Brain, Meaning, and Computation

Von der Philosophisch- Historischen Fakultaet der Universitaet Stuttgart
zur Erlangung der Wuerde eines Doktors der Philosophie (Dr. phil.)
genehmigte Abhandlung

Vorgelegt von

Michael Klein

aus Leonberg

Hauptberichter:	Prof Dr. h.c. Hans Kamp PhD
Mitberichter:	Prof Dr. Guenther Palm
Tag der muendlichen Pruefung:	1. Februar 2007

Institut fuer maschinelle Sprachverarbeitung
Universitaet Stuttgart
2007

Acknowledgment

This work was supported by the German Research Foundation (DAG), the German Academic Exchange Service (DAAD), and the National Institute of Information and Communications Technology of Japan (NICT).

I am very grateful to Hans Kamp for all his time, patience, good advice, many discussions and for teaching me the essentials of semantics, logic and philosophy of language. His criticism shaped my writing abilities and finally made me able to write scientific text.

I would like to thank Guenther Palm for his support and feedback, and for his advice in questions related to neural modelling. His outstanding knowledge about the modeling of unsupervised associative learning and the simulation of binding by synchronization gave this thesis its first main theoretical backbone.

It was Michael Arbib's Brain Simulation Laboratory where I leared how neural models can be designed on the basis of neuroscientific data. He also gave me encouraging feedback, as well as many critical comments on my ideas.

Kenji Doya invitated me to Japan, and there I learned a lot about internal models and reinforcement learning, the second theoretical backbone of this thesis. He also teached me how to write scientific articles.

With Peter Indefrey I had a very stimulating discussion about the modelling of individual concepts and how my theory could be tested experimentally.

I also had stimulating discussions about how the brain computes with Christof von der Malsburg.

Almut Schuez and Valentino Braitenberg shaped my way of thinking about the brain by discussing the biological evidence for specific learning mechanisms with me.

Aude Billard taught me the methods of synthetic brain imaging and helped me to program my very first neural network. She also introduced me to C programming. Laurent Itti introduced me to C++ programming. He always had helpful and stimulating answers to my questions concerning visual processing and attention in the brain.

Dirk Wildgruber taught me the methods of functional brain imaging. With him I shared many exciting discussions about philosophy and language processing in the brain.

Finally, I would like to thanks Cristina Rosazza and Majken Hulstijn for comments on the manuscript.

Contents

Chapter 1

Introduction

Everyone agrees that human beings can acquire a natural language only because they are biologically prepared to do so and only because they are exposed to other people in the culture speaking the language. The difficult part is in specifying the exact nature of this biological preparation, including the exact nature of the cognitive and learning skills that children use during ontogeny to acquire competence with the language into which they are born.
Michael Tomasello (2003)

1.1 Background Problem

Experimentalists (especially psychologists) have the tendency to regard the generation of data as more important than the development of theories. They often view theories merely as a means to explain the data and maybe to generate new experiments. Theoreticians (especially linguists) on the other hand often tend to regard theories as the primary goal of the scientific work and data only as a means to evaluate them. But while a theory is always a theory about something (i.e. empirical observations of some kind), data can stand on its own feet. To publish data without an underlying theory is common practice. Especially data which contradicts a widely accepted theory is interesting and an important and stimulating scientific finding. However, in the neuroscience of language, the emergence of functional imaging methods such as EEG, MEG, PET and fMRI has, within the last ten years, led to a very strong prevalence of empirical results over theoretical understanding. The enormous flood of (often inconsistent) empirical results currently resists coherent interpretation by the few theories that are available. These theories might be categorized into (i) *structural* theories and (ii) *computational* theories.

Structural theories usually come in the form of boxes and arrows, in which cognitive functions, described by a general label such as e.g. *lexical retrieval* or (which is worse) *semantics* (Deacon 1997), are attributed to a certain anatomical brain region. They are often generated by scientists working with patients or using functional imaging. Using these methods, correlations have been established between some aspects of language processing

and local changes in neural activity, or lesions have been described that cause the impairment of certain linguistic capabilities. While traditional structural models, e.g. Geschwind (1965), which are generally agreed to capture at least some part of the truth, are too coarse-grained to deal with the specific data of current imaging experiments, no consensus has been reached about more recent models of this type (Deacon 1997). The labels used to denote the cognitive functions in those box and arrow models allow little insight into the exact nature of information processing that is claimed to take place in the respective brain area and, therefore, does not allow precise predictions that can be experimentally evaluated. Also, the box-and-arrow models are still very distant from a neural explanation of language processes as the linguist understands and describes them.

Computational theories make stronger claims about the exact bits and pieces of information that are stored and retrieved during language processing, but they usually do not relate a specific kind of information processing to a neural substrate, such as a brain region (Levelt 1989, Plaut and Kello 1999) let alone physiological entities such as neurons, synapses or cortical columns. There are several kinds of theories that might be called computational. Here I should mention three of them: linguistic theories, psycholinguistic theories and connectionist models.

Linguistic theories often offer precise, even algorithmic accounts of language processes. Some of those theories, such as Chomsky (1965, 1994), claim to describe cognitively real processes, while others do not. But even those linguistic theories which claim to describe cognitive processes are often not in a form that allows testing by behavioral experiments. and so far the relation to cognitive processes could only been demonstrated experimentally for very few of them. And to use theories which do not describe cognitively plausible processes as a basis for experimental brain research is not very promising. Especially most semantic theories (Montague 1973) with their anti-psychologistic roots (Frege 1892) keep a safe distance to mental processes or mental representations while trying to describe meaning in terms of correspondence between utterances and states of the world. For theories which do not aim to describe the cognitive processes and cognitive representations involved in meaning processing, this is, of course, a justified approach. But despite the *non-cognitive* character of its approach, theoretical semantics (in accord with philosophy of language), offers a large and articulate body of theories about the computation of meaning, as well as detailed observations about all kinds of meaning phenomena which a cognitive or neural theory of meaning also has to deal with. For their part, theoretical semanticists, with their insights into the structure and use of language, could give guidance to neuroscientific research to a greater extent than has been the case hitherto. Semanticists have, in spite of their growing knowledge of how meaning is expressed in language, hardly played a part in efforts to account for meaning-related phenomena in neural terms. In experimental brain research, in psycholinguistic theories and experiments, and in connectionist models, all of what has been learned about the representation and processing of language by the methods of theoretical semantics has thus far been ignored. One of the main reasons for this lack of attention is the result of the *non-cognitive* tradition, mentioned above, in which semantic theories have thus far been developed. It makes them not only hard to test in behavioral and especially physiological experiments (since their basic concepts seem to be

so orthogonal to cognitive or physiological ones), they also do not seem a realistic point of departure for the enterprise of understanding the neural basis of meaning processing (e.g. the interpretation of neurophysiological data or the design of new neuroscientific experiments). This is particularly salient when we turn to very basic semantic relations and operations which the working semanticist usually takes for granted as he grapples with problems that count as the real challenges within his community. In the light of results that have been emerging within the neurosciences about the neural representation of information generally, theoretical semanticists will, if they are to make a useful contribution to a neural theory of meaning, have to rethink their theories, and establish a new conceptual foundation for them which is compatible with these results.

Most current psycholinguistic theories, such as Levelt (1989), offer statements about cognitive processes involved in language production or comprehension, as they take into account cognitive concepts such as retrieval from memory or spreading activation or lateral inhibition. This enables these theories to make good predictions about empirical data such as reaction times and speech errors. But when you compare them to linguistic theories, then, of course, they are pretty meager in terms of the variety of linguistic phenomena and the complexity of sentences they can explain. In particular, when it comes to the compositional aspects of meaning, they usually do not give a precise description of how linguistic form is used to determine meaning and vice versa. This is not surprising because, of course, these theories always have to carry the additional burden of having to agree with behavioral (and not only linguistic) data. The need to show that such data are consistent with the theory slows down the theoretical development considerably (and most psycholinguists are anyway fully engaged in obtaining this data). Especially when it comes to meaning, most experimental work in psycholinguistics focuses on single word processing, using picture naming and priming studies. As it was already pointed out by Levelt (1999), there is a need in the experimental sciences for a cognitive theory of semantics which can also serve as a basis for brain research. In general, psycholinguistic theories appear to describe cognitive processes involved in language processing for which a brain area which activates when those processes are executed can be found, as has been shown by Levelt et al. (1998) and Roelofs and Hagoort (2002). The problem with this mapping is that psycholinguistic theories do not take the computational organization of the brain into account. I will try to show in this thesis that the *modular* organization of the brain may well be very different, following different principles than the ones linguistic and psycholinguistic theories often appear to assume. Simply trying to map components of these theories onto the brain may therefore not be the right approach.

Computational neural models on the other hand use small neuron-like computing units, and connections between those units that resemble biological synapses. This allows those models to employ knowledge about the computational possibilities of various types of neurons and synapses, as well as insights about the distribution of computations over different brain areas. Models that connect behavioral data with data obtained in electrophysiological single cell recordings (in animal brains) are often of this type. But for higher cognitive functions, especially language processing functions, single cell recordings are hardly ever available. Since this major source of evidence about the localisation of processing is lacking

(although other sources are in fact available), almost all connectionist models of language processes do not make strong claims about the brain areas that carry out the computations executed by specific artificial neurons of the model. Moreover, most current connectionist models are even simpler than the current psycholinguistic theories.

1.2 Massive Integration

What we need is a computational (algorithmic) and precise (formal) description of the processes and representations involved in a) neural theory that describes how interesting linguistic (semantic/pragmatic) phenomena are processed on the neural level. Such a theory would bridge the gap between cognitive neuroscience and theoretical semantics and create a framework in which both communities can think. To do this, the theory would, on one hand, have to connect to linguistic theories and explain linguistic phenomena. On the other hand, it would have to be compatible with results and knowledge provided by neuroscience. In computational and experimental neuroscience a variety of methods has meanwhile led to theories and experimental results about the nature of neural representations and computations as well as about the regions of the brain where they are located. In this way they have created a platform on which it now seems possible to develop accounts of how the human brain understands and produces language.

A reformulation of the foundations of semantic theory in terms of the brain will have to take into account that *in the brain* representations and computations manifest themselves as states of physically real entities (i.e. neurons and synapses) and it is the properties of those entities and their interaction that determines what kind of computations or representations are possible. Our current understanding of neural information processing is based on experimental investigations of these entities (both on the microscopic and on the macroscopic level).

The knowledge about information processing in the brain, which I claim needs to be integrated in such a theory, can be divided into several kinds of data. These data can be classified into two broad categories: the *behavioral* level and the *neurophysiological* level. On the behavioral level there is linguistic data, i.e. data which is generally used by theoretical linguists to evaluate their theories. In large parts it consists of the utterances which normal speakers, who have fully acquired their native language, produce under normal circumstances. But it also includes the utterances (or pre-linguistic sounds) of children in their early stages of development, and the transitions between the utterances made in the early stages and those made in the later stages. Further, it includes the errors children and adults make under normal circumstances. It also includes reaction time (RT) data and learning rates.

Behavioral data can be obtained either in patients (e.g. people with injuries, diseases, or brain lesions which impair their linguistic performance) or in (so-called) normals (i.e. people without such impairments). Behavioral data of patients already give insights into the locations of the brain mechanisms involved and the neural architecture of the cognitive processes.

More information about the biological implementation of those processes can be obtained by measurements on the *neurophysiological* level (the second broad category). Here, it isn't the behavior of the tested person that plays the central role, but some measured physiological brain response.

One possible way to categorize data of the neurophysiological level is by considering the number of neurons they measure. Small scale measurements are the conduction or stimulation of single or small numbers of neurons by introducing an electrode directly into the brain. In some rare cases, when patients undergo some sort of epilepsy surgery, electrodes can be directly used to measure or to stimulate single neurons while the patient is performing a linguistic task. But in general this type of data is obtained with animals. Nevertheless, animal brains and the human brain appear to be very similar in terms of their general mechanisms of computation (this is discussed in a bit more detail below). Therefore, the insight that animal research provides into the general learning and processing mechanisms of the brain can be applied by imposing biologically plausible constraints on theoretical models of cognitive processing in the human brain. I will call the insights into information processing in the human brain provided by this kind of data *design constraints*. Design constraints include observed learning mechanisms of brain areas, as well as connections between brain areas, and conduction times.

Large scale measuring methods can detect the activation of large numbers of neurons or synapses. The ones most frequently used are fMRI and EEG (others are PET, SPECT, MEG, and TMS). Functional imaging data, such as fMRI, give us clues about the location of certain brain processes with a fairly good space resolution, but pretty bad time resolution. EEG on the other hand gives us insights into the time response of neural activation, but the space resolution is very low. EEG (and MEG data) can also give us information about activity in different frequency bands (such as the γ-band). While the results of small scale measurements are mainly used contrain the design of a neural model, large-scale measurements can be used to test predictions such a model makes. To obtain such predictions, functional imaging data or EEG data need to be simulated with the neural model. So far this has hardly every been done in relation to language.

By using computational models, all these different kinds of data can potentially be integrated. Connectionist models allow a mathematical and algorithmic formalization that helps to make the assumptions and consequences of different possible computational alternatives precise. The box-and-arrow models discussed above or theories that are not specified in a mathematical or computational way often resist falsification because their consequences are not clear. In contrast computational neural models are based on precise algorithmic theories and make quantitative predictions. Unlike the theories in which a box or a module is denoted merely by a verbal label (e.g. *concept selection*), a connectionist model needs to be specific in terms of input, output and exact information processing. The theoretician is faced with a number of choices or possible computational theories of how - *precisely* - a cognitive process is executed. The exact specification of the process allows one to predict and compare the consequences of those different theories. Therefore, such theories may have different potentials for explaining linguistic data (i.e. on the basis of biologically plausible computations they could give reason why certain natural language

expressions have a certain form). Depending on the capability of models to account for linguistic data in this way some models could already be eliminated. In addition, a computational neural model can, if it is designed in the right way, also simulate behavioral data, such as reaction times or learning rates. Comparison on the behavioral level is then likely to rule out several of the computational alternatives because their predictions are not in accordance with the behavioral data. If computational possibilities and algorithms of the real brain are taken into account, as I will do in this thesis, some of these possible alternative theories cannot be considered plausible from a neurobiological point of view. This further reduces the number of possible theories. Decisions between alternative theories can also be made by including the simulation of brain lesions so that the resulting behavior of the model can be compared with patient data. Finally, computational neural models can be used to simulate functional imaging (fMRI or PET) or electrophysiological (EEG or MEG). Based on the predictions made by these types of simulations, experiments can provide strong evidence for or against some possible theories. This is a rough description of the way in which the massive integration of data enabled by computational neural models allows for specification of theories and predictions that can lead to scientific progress by falsification (Popper 1934) and maybe lead to a true understanding of how meaning is really processed and represented in the human brain.

Another reason why a *computational* neural model is needed, instead of just a imathematical theory, is because of the complexity of neural dynamics. When we model a cognitive task in terms of neurons, connections, and learning rules, we specify *static* properties of the brain, but what we are interested in is dynamical behavior and changes of values, such as for example the activation of neurons, the changing strength of synapses or focal changes in neural activity within certain brain regions (as it is measurable in the real brain with functional imaging). The interaction and the number of variables in such a model is so complex that the outcome cannot be calculated by hand so that an implementation on a computer is needed.

The number of processing units in the real brain (10^{11} neurons and 10^{15} synapses are estimated for the human cerebral cortex alone) is so big that a simulation with a realistic number of neurons or synapses is still not possible (not even for a single brain area). Therefore, a certain degree of abstraction is necessary if an enterprise such as the one sketched above is to be practically possible. Large-scale neural modeling is one possible approach. In large-scale neural models the claim is not to model the behavior of single neurons, but rather of larger neural groups. Whichever the method of abstraction is, it is important to keep the essential features of the neural organization of computing, such as the distribution of computing units, or, as I will show in the next section, the learning algorithms used by the different parts of the brain.

1.3 Learning

The way a neural model processes information is determined by its neurons, the interconnection of those neurons (i.e. the structure of the model), and the strengths of those

connections. These strengths can either be fixed (by the programmer) from the very beginning or they can be trained with some learning algorithm. There are several reasons why a neural theory which accounts for semantic processing should include learning.

The first reason is a very pragmatic one: Learning algorithms help to fix the synaptic strength between the neurons in a way that allows the model to actually perform the task. Many neural models are so complex that to fix the strength of the connections between its neurons by hand is more costly than to train them with learning algorithms.

The second reason is that the brain (in most parts) is a learning system. It can learn new languages and forms of communication until death. Modeling it as such as a learning system is likely to help us to arrive at biologically plausible neural representations, and algorithms as well because the way the brain learns determines the way it can represent what it has learned. One of the major goals of the theory developed in this thesis is to take into account the mounting evidence that different major brain structures (cortex, basal ganglia, cerebellum) differ mainly with respect to the learning method they use (Doya 1999, 2000). To accomplish complex learning tasks, such as language acquisition, the brain uses these interacting subsystems to achieve the global goals of the system. In particular, unsupervised learning in the cortex (Hebb 1949, Bliss and Lomo 1973, Artola et al. 1990) is used for concept acquisition and word - meaning associations, supervised learning (cerebellum) can be used to train forward models (Jordan and Rumelhart 1992) for prediction, and reinforcement learning (basal ganglia) can be used to train a value function (Sutton and Barto 1998) necessary for goal-directed utterance selection. While physiological investigations have shown us how the brain learns, behavioral studies on language acquisition have taught us what kind of information the language acquiring child uses (and in particular what kind of clues parents and the environment provide it with) to learn the meaning of words and more complex utterances. We also have some knowledge about how the brain represents this information. Combining the theories about the learning processes *inside* the brain with the theories about language acquisition observed from the *outside* allows us to see the causal link between behavior and biological mechanisms and can bring us another step closer to an understanding of the real neural processes and representations involved in the production and understanding of meaningful language.

The third reason is that to include learning makes the theory subject to testing by empirical data on language acquisition. Such a theory provides descriptions of representational structures not only for the mature language use of the grown-ups (as most theories of theoretical semantics do), but also of the language of children in all stages of language acquisition; and, what is almost more important, it describes the transitions between them (computational accounts of these transitions are important and too little systematic research has been done to provide them). But developing a theory about language acquisition requires the inclusion of additional non-linguistic capabilities and cognitive functions, such as attention (to describe gaze-following and shared attention capabilites of the child) as well as role-reversal imitation. The ultimate goal of such a theoretical framework for the description of meaning acquisition on a neural level would be to describe the information contained in the represented semantic abilities at any point in time. As this information always is the basis for further learning, it is possible to account for how similar information

provided to the child by the environment in an observed usage-event enables it to learn very different aspects of language depending on its stage of development. The theory, therefore, could describe how the semantic abilities of the child are building up. And due to the integrative character of the theory, it could also simulate physiological data during acquisition that can be tested with brain imaging techniques.

The fourth reason for having an adaptive model is its use in the modeling of information exchange. Meaning, from a cognitive point of view, emerges through learning - from the first words acquired by children from the interaction with their parents to the change of (semantic) knowledge by verbal information exchange as it was extensively discussed already by Frege (1892) (in a very different context). Although the acquisition of basic syntactical structures of a person's native language might be over at an early stage of childhood, it is obvious that her communication capabilites have the potential of further extension and refinement through her entire lifetime. An adaptive framework of meaning, like the one I will introduce in this thesis, does not only allow the modeling of early meaning acquisition but also the more elaborate acts of information transmission we find in adults.

The last reason for modeling linguistic ability as a learning system is that this allows a clear distinction between innate structures and acquired capabilities - the topic of the next section.

1.4 The Question of Innateness

The question which linguistic capabilities are innate and which can be learned has been a hot topic of discussion for many years. The starting point of this dispute was Chomsky's (in many points justified) attack on the behavioristic research philosophy and especially on Skinner's book *Verbal Behavior*. Chomsky's subsequent work led to the research program of *Universal Grammar* and *Principles and Parameters* in which (mainly) linguists tried to describe the innate human grammatical system of the brain by specifying linguistic rules. Computational neural modelers, on the other hand, have been trying for many years to show that general learning principles are capable of acquiring the representations that a system needs if it is to be able to use natural language grammar. In recent years the two sides have shown some convergence and the dispute has somewhat cooled off. Since natural language syntax does not play a significant role in the simple artificial languages I study in this thesis, this particular issue does not arise. However, in principle the same issue is relevant in connection with semantics and in my theoretical approach to the processing of meaning in the brain with computational neural models, I will specify certain biological structures, learning algorithms, and parameters (which might been seen as genetic dispositions) and try to show how such a system can acquire linguistic capabilites. In view of this I would like to say some things about the question of innateness and also about the general approach of addressing this problem that is taken in this thesis.

As put nicely by Michael Tomasello[1], it has always been clear that the brain is not a complete *tabula rasa* and that language is only acquired because humans have certain

[1]the citation at the beginning of this chapter

genetical dispositions that allow them to learn language (in contrast to other animals). On the other hand everyone is aware that at least some aspects of language are learned. This leaves us with the question of *exactly* what is innate and what can be learned. More specifically the question is: *what is the necessary innate basis of learning natural language*. Computational neural models offer a way of investigating this question.

The genome specifies structures and their biochemical components, i.e. what the genome specifies is basically the design of the human brain. Computational neural models make it possible (and require!) to describe such a design. All the genetical dispositions for language must - in one way or the other - be in the brain by virtue of some architectural properties, types of neurons, their connections or the strength of those connections (for genetically specified computations), and the (learning) algorithms with which they interact. The insight that computational models serve very nicely to model this difference between dispositions and acquired knowledge was pointed out by Elman et al. (1996). They allow us to work out the dispositions to acquire language which must be genetically specified and to specify a brain that has what Arbib (2000) calls *language readiness*. The general approach of describing the mechanisms for language acquisition in terms of a computational neural system is not far removed from the description of a language acquisition device. However, to describe a language-ready brain in terms of neural architecture is, of course, very different from the attempt to describe the genetical determination for human language processing by grammatical rules. But the description of the architecture makes it possible to actually *investigate* what kind of dispositions are needed to allow natural language to be learned (while the mere specification of grammatical rules does not). For example, concepts are acquired, but they are based on innate basic features, and the unsupervised learning algorithm which the brain uses to acquire them is also innate.

For this approach - called synthetic psychology by Braitenberg (1984) - to work, it is important to start with a simple system that comes with a simple *initial* structure and to test what kind of behavior (i.e., in the case of language acquisition, what kind of communication abilities) it can develop. Then it will be obvious whether the amount and kind of initial information used was sufficient. It can be regarded as the ultimate goal of this approach to describe the essential structures of the human brain which are necessary for the acquisition of a human language. However, one feature of the synthetic approach to cognition (Braitenberg 1984) is to begin by assuming a simple mechanism and to discover its limitations. It is important to understand this principle as a way of doing research and not as a claim. I do *not* claim that the structural and parametric dispositions I will describe in this thesis are sufficient to learn human languages - I am just trying to see how far I can get with the particular assumptions I make - introducing each time I reach a limit another disposition into the model. Such a system, for which some initial information or structure can be specified, but which has the potential to fully develop its functional capabilites by interacting with its environment, is called a *dynamical system*. Mathematically, a dynamical system can be described by a set of possible *states*, a *state function* (mapping previous states and the input onto new states), an *output function* (mapping states onto outputs), and an *initial* state. The initial state, the state function

and the output function can be thought of as representing the genetical disposition[2]. The disposition given by the initial state and these functions needs to be specified in terms of neural properties (i.e. connections, initial connection strengths between neuron etc.). Although we do not know yet what exactly the genetical dispositions for language are, it is evident that genes only determine the way the human brain processes by creating the necessary neural structures; any form of processing algorithm built into the brain is built in terms of neurons and connections.

Although the approach sketched so far allows me to stay fairly neutral with respect to the question which dispositions are innate, I still would like to point out that three of Chomsky's main arguments for a universal grammar can no longer be considered tenable. The first one is his point that parents do not teach their children language explicitly and that children do not get a *reward* for successful language use. Recent data of Chouinard and Clark (2003) has shown that parents actually do teach language to their children in a subtle but nevertheless efficient way. For example, parents might not point out the mistakes children make, but they often present the correct form of an utterance to them as a reaction to a wrong usage. Children can also observe the consequences of their utterances to improve their linguistic skills (the expected effect can be contrasted to the real effect resulting in some kind of error signal), a capability that can be implemented in neural systems, as was demonstrated by Jordan and Rumelhart (1992). And although children do not get a reward from their parents for correct usage, their own awareness that their use was correct will produce an internal sense of success that can play the role of a positive reward in language acquisition. Moreover, they will no doubt find that by using language they can accomplish more than without it, in particular that language will help them to get others to do what they want (e.g. given them the thing that they want).

Chomsky's *poverty of stimulus* argument[3] has come to appear increasingly problematic as computer simulations of syntax acquisition, such as Chang (2002), and language acquisition data (Tomasello 2003) are showing more and more convincingly that the learning mechanisms of the human brain might well be able to acquire syntactic rules from the limited linguistic input they get. A weak point of the simulation studies remains the long training times, which seem to be far from being comparable to the actual learning times of children.

Another argument is about *Language Universals* [4]. Here it has to be said that the structure of language is given by three factors, the first is the limits and capabilites of the human brain, the second is the optimality of linguistic structures for information exchange, and the third one is the conventions within a language community. These three factor are not independent of each other. One one hand, the optimal structures for information exchange, depend on the possibilies of our brain to process linguistic structures. In that

[2]Note however that, for good reasons, most neural models do not start with a *new-born* brain, but usually the initial capabilities in a modeled cognitive system already corresponds to those at a certain point of development.

[3]The argument that children do not receive enough stimulation to acquire the complex structure of natural language syntax.

[4]It has been observed that many languages in the world use similar kind of linguistic structures.

sense our brain determines optimal structures of information interchange. On the other hands, it is possible that genetic changes have led to brains, that can process structures that are more efficent in the transmission of information. This would mean that the optimality of linguistic structures for efficient information exchange could have determined the processing capabilites of the brain. For example, it can be mathematically shown that a compositional language is more optimal for information exchange than a language that only uses non-compositional information. Therefore, brains that can use compositional language have a clear advantage over those that cannot. Only with brains that are capable of using compositional language can a language community come up with a compositional language and use it conventionally. This means that a community can only adapt to more optimal forms of language after the brain has evolved to process them. In summary, if language is similar across cultures, then that is because we have the same brains and we have to use them in the best possible way in order to communicate. It is however, not very likely that the brain's capacity to process structures that can be used for efficent information transmission has evolved solely for the purpose of communication, which brings us to the question of *domain specificity*.

The problem of domain specificity is closely related to the whole discussion about nature vs. nurture. For language, being processed with domain specific brain mechanisms can mean two things: (i) language is processed in brain structures or brain areas in which no other cognitive function is being processed or (ii) language is processed and acquired with neural algorithms that are not used to process any other cognitive function. Since it is clear that certain cognitive functions used in language are also used in other cognitive tasks (e.g. basic auditory processing) the more reasonable way of asking the question is - here as elsewhere - *which* components of language are domain specific. In this thesis, I will not discuss whether there are brain areas and structures used solely for language processing. I will, however, demonstrate that the simplified forms of meaning that I deal with in this thesis can be acquired with algorithms that can be (and are) used to carry out other cognitive tasks as well. But since I am not dealing with complex syntactic structures, my simulations and theoretical insights do not answer the question whether the processing of syntax requires special algorithms.

One of the cognitive tasks I investigate in this thesis is the learning of utterance effects. Here, a good analogy can be found in the motor system. To choose the right action, in a certain context, the motor system needs to predict the outcome of a certain action. Models that make this kind of prediction are called *internal models*. But internal models can be used not only to choose the appropriate action in a certain context, but also to understand the intentions of others. If you observe a person performing a certain action in a certain situation and you know which effect this action will have, you can use this model to understand the actor's *intention* (Wolpert et al. 2003). Since utterances can be regarded as goal directed in production, and because understanding the intention behind an utterance has to be regarded as the ultimate goal in comprehension (Grice 1975), internal models might give new insights into the representation of meaning in the brain. But although my theory will use general methods of neural computation that can be used for other cognitive tasks as well, the architectures of the system components are designed to map a certain

input structure to a certain output structure in such a way that they can aquire language meaning. Such aspects of the architecture, as well as way the components are connected, may well be where the genetical disposition lies.

1.5 Goal-Directedness

In order for a species to persist, genetic modifications (mutations) occurring within it should enable its members to better exploit their environment. Better communication skills give humans a major advantage since they lead to social interaction and cooperation in achieving a goal. Compositional language for example allows us to have precise communication with lower learning costs.

A major accomplishment of evolution is that it has equipped the brain of humans (and other animals) not only with specific processing capabilities, but also with the ability to learn and adapt to the environment. One of the major learning methods of the brain is *reinforcement learning*. Reinforcement learning uses genetical information about which actions and situations are positive for the persistence of the organism (or its genes, to be precise) and rewards those, causing the desire to bring the situation about again or repeat the action. It also punishes actions and situations with the effect of causing the desire to avoid them in the future. The organism learns which situations and actions (and also objects) are more desirable and which are less. This in turn leads to humans and animals having goals (or intentions) and thus to their functioning as goal-oriented systems. If we see the brain as a goal-directed system generally, the goal-directedness of language can be regarded as a special case of goal-directed behaviour, and language as one of the means which agents can use to achieve their goals. Thus, the learning of language becomes an optimization process, in the course of which the ability of the language learner to handle the environment increases with higher levels of linguistic sophistication: the more accurately we communicate, the more likely it is for our goals to be accomplished.

The awareness that utterances have extra-linguistic motives and extra-linguistic effects goes back a long time (Austin 1961, Wittgenstein 1953). Wittgenstein and Austin saw and stated clearly that the purpose of language is not - in the first place - to describe the world. Linguistic utterances - like actions - are used to accomplish certain goals. As subsequent work within speech act theory has made increasingly clear, utterance effects vary from the most direct and directly observable to the indirect and only indirectly observable (such as the sharing of information and other more subtle social effects). Speaking involves conceiving an intention and selecting the relevant information to be expressed for the realization of this purpose (Levelt 1989). It is, further, generally recognized that the utterance of a speaker achieves its effects by way of getting the addressee to recognize the speaker's intention to achieve it (Grice 1975). In other words, to really understand an utterance is to understand the speaker's communicative intention. To understand intention, as pointed out by Bloom (2000), Tomasello (2003) and many other researchers in language acquisition, appears to be an essential prerequisite for acquiring the meaning of words. Since the purpose for which utterances are used plays such a prominent part in language use and lan-

guage acquisition, it seems natural to model acquisition of meaning and linguistic structure in a setting where it goes hand in hand with the learning that and how utterances of given forms can be used to achieve certain extra-linguistic goals. It is reasonable to assume that if we really want to understand communication, we must understand the relation between the utterances and their goals. Language production should be modeled as involving the transition from goals to utterances and language comprehension as involving the transition from perceived utterances to inferred goals. Such an approach to language use needs to answer several questions. From a language production point of view such a theory needs to explain: (i) how we know in what situations it is useful, given certain desires, to speak instead of performing some other, non-verbal action, or simply doing/saying nothing, and when to say what to whom; and (ii) how we transform the goal we want to achieve in the given situation into an utterance (or a sequence of utterances). From the view point of language comprehension, such an approach offers the possibilty of modelling language understanding in the context of knowing what kind of goals another agent is likely to have in a certain situation. In other words, such a framework has to offer some sort of probability filter helping to process utterances based on the question: What can this person possibly want from me? The addressee can use this mechanism to estimate which communicative goals are more likely than others in this particular context, and based on this estimation, the probability of certain words and phrases could be computed in a top-down manner. In that way, such a mechanism would influence the processing of not only pragmatic and semantic, but also of syntactical, and even of phonetic and phonological information.

Although the problem of intention in communication has generally been recognized in philosophy of language, linguistic theories (with some exceptions), but especially psycholinguistic ones have not really worked out a theoretical framework which makes it possible to address this issue. The approach I take towards the problem of meaning in this thesis differs from those commonly adopted in theoretical linguistics, in psycholinguistics, or in neuroscientific approaches. It differs in paying attention to the extralinguistic purposes of language use from the very start.

1.6 Scope of this Thesis

The goal described above - to develop a neural theory of meaning, bridging the gap between theoretical semantics and experimental neuroscience - is, of course, a very difficult one. The work presented in this thesis cannot be more than a first step in this direction. I develop a computational neural theory that describes how interesting phenomena at the linguistic (semantics/pragmatics) level are processed on the neural level. The bridge between neuroscience and theoretical linguistics isn't quite finished. The main shortcoming is that the language my theory is able to process, while it captures what I consider the most essential features of human language, is still not natural human language (the topic of research for theoretical semantics). It involves a moderate degree of compositionality, but it is still comparatively simple. However, building the bridge has naturally led to the integration of concepts of cognitive psychology, such as working and long term memory, as

well as semantic and episodic memory, into the theoretical framework.

I have dealt with what I consider the most essential requirements of human communication: concept acquisition, goal-directed utterance selection, compositionality, topic-comment structure for information transmission, and a variety of different types of speech acts. I have not dealt with syntax, although syntax is of course a defining feature of human language. But until the semantic structures my theory can deal with are rich enough to capture the meaning of utterances in which the details of syntax become decisive, syntax will not be relevant.

I am going to present a theory of meaning representation developed in the approach I have sketched above. These ideas will be used to motivate a general formal framework for describing meaning representations as they are likely to exist in the human brain. I will start by explaining the emergence of conceptual representations in the brain and from there go on to language use, compositionality and different kinds of speech acts. Since in this thesis I am developing a neural theory of meaning, the notion of *concepts* familiar from the theory of linguistics has to be recast in terms of physiologically real entities (such as neurons and synapses); it is the properties of these entities and their interactions that determine the range of possible representations and computations.

Since I want to model agents who want to communicate with other agents in order to achieve certain goals, I have equipped my agents with the capacity of having desires. To this end, I have placed them in a environment in which they develop these desires by receiving rewards or punishments in specific states of the environment. Thereby they acquire an internal representation of which states are desirable and which states should be avoided. They also learn from which other states these desirable states can be brought about by their actions (linguistic and otherwise). In that way they learn that these other states are also desirable.

In the theory presented in this thesis (and the computer simulations that complement it), internal models are used to predict context dependent utterance effects. A child that is learning to speak can be regarded as a system that interacts with the environment by observing the utterances of other people and their effects, as well as the effects of its own utterances. In other words, children learn under which conditions which utterances have which results. In line with this, observing the consequences of utterances is the major principle of learning in the approach I employ. However, learning from this kind of experience makes sense only if the agent has the capacity of developing, on the basis of his observations, means of representing what effect his actions will produce in given circumstances. In the theory presented in this thesis these means are given in the form of an internal model which predicts the effects produced by different utterances. The internal model of an agent is trained by his observing the changes in the environment that his utterances and those of other agents bring about.

However, to know the effects of an utterance cannot be sufficient by itself to make the appropriate selections of utterances in a context. A selection mechanism is needed that finds the best utterances (or actions) for a specific goal. There is good evidence that reinforcement learning (Sutton and Barto 1998) is a good approximation of this selection mechanism in the human brain. By means of reinforcement learning a so-called *value*

function can be trained, which computes a value for every possible sensory state of the world. Such a function can be thought of as specifying the desires of an agent. A state with a high value is strongly desired and a state with a low value is one the agent tries to avoid.

I use a multi-agent game simulation to show that the combination of reinforcement learning and internal modeling is most effective to teach an agent language by imitation and to produce optimal communicative behavior. Previous work in multi-agent modeling of communication (Steels 2001, Cangelosi et al. 2002, Smith et al. 2003) generally investigated what might be called the descriptive aspects of language. These approaches deal with problems such as how object categories emerge, how these categories are given linguistic labels, or how such labels are shared within a language community. These aspects are essential for understanding how a successful repertoire of reference acts can evolve within a language community. However, to go beyond the descriptive use of language, which is the aim of the studies described in this thesis, it is necessary to investigate the relation between utterances and goals. Of course - in a certain sense - the agents in some of those previous studies can also be said to have goals - the goal of a successful act of reference or the goal of successfully transmitting a correct description of some state of affairs. But to get a better picture of communication it is important to move on to the general phenomena of goal-directed communication.

In my first study the language that is to be acquired consists of a small set of one-word sentences which, when used in the right situation, have the effect of a certain kind of object being handed to the speaker by the addressee. In this game, a learner (*child*) is placed in an environment with two other agents (*parents*) who can speak already (i.e. they are pre-programmed agents who make pragmatically correct use of the utterances of the language). The agents have to perform a certain task (feeding), for which they are rewarded (by reduced hunger). The language learner learns to predict the outcomes of utterances in varying contexts (e.g. he learns that to ask another agent for a certain object is only useful when the addressee has the desired object). In this way the agent learns to choose that utterance in a given situation for which the predicted effect will be more beneficiary to him than that of any other alternative utterance he could have used. At the same time the agent learns to *understand* the utterances with which others address him - that is, he learns to infer the speakers' desires from the utterances they make - and to react accordingly.

In the computer simulations described in chapter 5, the design of the agents is demonstrated to work for one type of speech acts: *requests*. In chapter 6 the theoretical basis for these simulations is extended to questions and assertions. For questions, I change the input to the value function and the internal model, which up to that point has just been a model of the current state of the world. This is necessary because the goal of questions is to increase the knowledge of the agent. Hence, agents need to be equipped with the possibility of acquiring knowledge. In the earlier models, the complete state of the world was assumed to be accessible to them at every moment in time. In the revised model, I introduce properties of the world that are not accessible for the agent at every moment and, add as a new component a knowledge base in which the (usually incomplete) knowledge

of the agents about the world is represented. Based on these extensions it is possible to show mathematically why an agent will be motivated to ask questions in order to obtain knowledge.

To model assertions as expressions of desires is only possible if the use of an assertion leads to a state with a higher value (i.e. a desired state). The basic effect of an assertion is to change the mental state (the knowledge or the focus of attention) of the addressee. To regard this effect without its social context is not sufficient if the goal is to explain why agents should choose to make assertions in the first place. In chapter 6, I discuss how the change of the mental state of the addressee can lead to higher rewards in the long run (and, thus, to a state with a higher value). One of the reasons involves the idea of social rewards through mutual cooperation.

In chapter 7, I investigate the acquisition of compositional language. Here I employ simple attention mechanisms in both theory and simulations. The internal models used in these later simulations do not predict the next sensory state on the basis of utterances and the previous sensory states, but focus on the objects which the utterances describe as changing their location. In this model agents develop the power to predict location changes of objects from the form of the utterances in which these objects are mentioned. In this way they can aquire the passive and active command of some basic form of compositionality. In particular, the agent can understand, i.e. predict the effect of, utterances he has never heard before. It is also in chapter 7 that I unite the theoretical insights about goal-directed communication gained hitherto with the ideas about concept representation described in chapter 4. In this integrated formalism, current states and desired states are represented as conceptualized states of the environment represented in terms of long term and working memory binding of feature neurons through synaptic changes and neural oscillations. This framework is used to model the speaker's representation of the addressee's knowledge about the world. Based on those representations, I model goal-directed information transmission in terms of (among other things) the topic - comment structure of the utterance.

The chapters 6 and 7 are mainly theoretical extensions of the computational framework developed in chapter 5 and the theoretical approaches of chapters 2 and 4. I have refrained from doing too many computer simulations, because at this point it was more important to get a larger theoretical understanding of the possibilites of the framework than to simulate some more details. The thesis ends with a chapter on how the theoretical insights to which it has led can be used to predict results which can be tested in behavioral and brain imaging experiments with humans.

Chapter 2

Learning, Representation, and Processing in the Brain

2.1 Rethinking Modularity

2.1.1 A Short Historical Overview

The dispute about the organization of processing and representation in the brain goes back a long way in time. The nineteenth century scientist Franz Joseph Gall (1757-1828) claimed in 1810 that the brain is a collection of 35 centers, each corresponding to a specific mental function (among them: mother love, generosity, secretiveness). This idea was called *phrenology*. To see whether the idea is true, Pierre Flourens (1794-1867) removed portions of the cortex in animals to see whether he could disturb specific mental functions, but he did not find evidence for the phrenological system. In 1825 he put forward the idea that all regions of the brain participate in all functions. This idea was called the *aggregate field view*.

The aggregate field view was held by many influential scientists, such as Ivan Pavlov (1849-1936), Henry Head (1861-1940), or Karl Lashley (1890-1958). The empirical basis of Karl Lashley's argument against the view that brain functions can be localized was his failure to find a learning center. In the light of the view which I will advocate in this thesis, this was an interesting observation. Although now it is generally agreed that the hippocampus serves a more central role in (at least certain types of) learning than other structures, it is also true hat most parts of the brain are actually involved in learning.

However, the aggregate field view had strong opponents, such as Paul Broca (1824-1880), John Hughlings Jackson (1835-1911), Gustav Theodor Fritsch (1838-1927), Julius Eduard Hitzig (1838-1907), David Ferrier (1843-1928), Carl Wernicke (1848-1904), Ramon y Cajal (1852-1934), Korbinian Brodmann (1868-1918), Wilder Penfield (1891-1976), and Norman Geschwind (1926-1984). Jackson was the first to postulate on a solid empirical basis that mental functions can be localized in areas of the cerebral cortex and, therefore, special abnormal mental conditions may result from structural damage of certain parts of the brain.

Especially electrophysiological research was able to create a stable basis for the idea of brain functions being localized in certain areas. This started with the work of Fritsch and Hitzig, who showed experimentally that electrical stimulation of the cortex (of dogs) produced contralateral limb movements. Ferrier was able to confirm this finding with monkeys. Penfield and Roberts (1959) were the first to work with the brains of humans (epileptic patients) and to localize motor and speech functions with electrical stimulation.

Due to this evidence, today there is no one who would doubt anymore that at least some cognitive functions are localized to a certain degree. However, the dispute about the structural organization of cognitive functions in the brain is far from over.

2.1.2 Current Issues

Through the popularity of *Modularity of Mind* (Fodor 1983) and the invention of modern functional imaging methods like PET (Peterson et al. 1988) and fMRI, a large community of neuroscientists has emerged who are trying to localize cognitive modules in the brain. The research approach they often follow can be characterized in the following way: Scientists try to localize a certain cognitive function, say, *syntactic processing*, by e.g. contrasting the brain activity measured during the execution of a task which appears to involve syntactical processing with the activity measured during a task which does not. While this approach might generate valuable data, there are nevertheless several assumptions underlying this method which are questionable. In particular, assumptions about the nature of localization need to be thought through in much more detail than has been done hitherto. For a start, we don't know why specific functions might be represented in particular areas. We also don't know exactly what kind of functions we are looking for. Further, we don't know to what degree the exact localization is predetermined (i.e. specified in the genes) and what the nature is of the genetical specification that causes a cognitive function to be localized in a certain brain area. For example, the part of the brain in which visual processing is located might be genetically specified in the following way: Genes specify certain neural structures, i.e. nerve fibers that project information from the eye via the thalamus to the primary visual cortex. As a consequence, the visual cortex starts to process visual information. For that reason the position of a cortical area with its motor and sensory pathways and its connection to other brain areas are likely to determine its function. Experiments have shown that connecting another cortical area in such a way that it receives visual input results in this other area processing visual information in a very similar way as the primary visual cortex.

It has been observed, in particular by Brodmann (1909), that different parts of the cortex have slight differences in their layer structure (see figure 2.1), e.g. the primary visual cortex probably has properties that allow it to have a slight advantage for visual processing in comparison to other cortical areas. But this advantage might consist only of small differences in the anatomy (e.g. a certain layer with a certain type of neuron is slightly thicker than it is in other cortical areas). Besides those minimal differences (that might make a significant difference in the processing of particular types of information), the cortex as such seems to have a very similar structure everywhere and doesn't seem

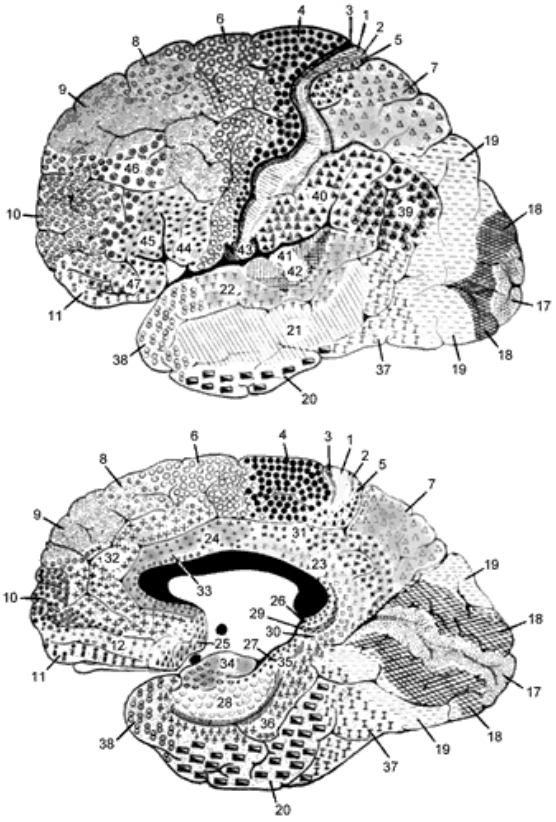

Figure 2.1: Brodmann Areas of the Cerebral Cortex

to be very domain specific (Braitenberg and Schuez 1998). Data about interindividual differences (i.e. differences between subjects in the localization of some cognitive function in the brain) and local reorganization of cognitive functions (i.e. the relocalization of a cognitive function to another brain area after lesion caused by e.g. strokes) are supporting this assumption. For example it can be seen that the *lower* the cognitive function is, the smaller the interindividual differences are. Therefore, lower sensory and motor functions can be studied with functional imaging in a much better way.

Furthermore it is still very unclear what the *elementary* cognitive functions are which are localized in certain brain areas. It is not known how a mental function is decomposed into subfunctions, what those subfunctions are and which of these subfunctions can be located in a given brain area.

There is, for example, data showing that (i) the syntactical capabilities of a patient are impaired if his Brodmann Areas 44 and 45 are lesioned. There is also evidence from functional imaging that these areas are activated during cognitive tasks which involve some syntactical processing. However, this still does not mean that there is something like a syntax module in areas 44 and 45, because (i) the areas which are involved in syntactical processing are very likely to be involved in other functions which do not deal

with syntactical processing; (ii) not all of the area is necessarily involved in syntactical processing and (iii) syntactical processing might involve other areas as well.

These considerations are especially true for the processing of meaning in the brain. Some researchers, such as Deacon (1997), have postulated a *semantics module*. However, the processing of meaning involves a large variety of complex processes and representations of different nature and it is not very likely that all this is done in one area of the brain. First of all, the very first relations between meanings and words are learned by connecting perceptual categories with words. The localization of the representation of the perceptual categories is likely to vary with the sensory modality of the information they categorize. In other words, the basic conceptual representations used in semantic processing are probably represented in almost all cortical areas (Hebb 1949, Braitenberg and Pulvermueller 1992). More abstract concepts might not depend so much on the sensory modalities but quite possibly they are represented in the vicinity of the sensory features they abstract from. To compose a complex meaning from the primitive representations of the meaning of content words and function words involves processes (such as the choice of speech act, the generation of the topic-comment structure of the utterance, the selection of quantifiers, definite or indefinite determiners, the selection of anaphoric expressions) of a very different nature, which are related to other kinds of processes in the brain that are used in different cognitive functions as well (such as memory, motor control, sequencing of movement, perceptual processing etc.). Looking at functional imaging data of semantic processing, different studies find activation in areas in different lobes of the cortex and it is likely that this activation is very much dependent on the exact task, and that these different tasks cannot easily be subsumed under the term *semantics*, and that they are confined to one (domain specific) brain area.

A semantic module characterized as an abstract function would be something like equation 2.1.

$$\vec{y} = Sem(\vec{x}) \tag{2.1}$$

where $\vec{x}$ is a vector consisting of all input that is required for all semantic processes and $Sem()$ is the function that computes every semantic task necessary in language processing. Of course, a semantics module is needed in both production and comprehension. But in the case of production the function maps meaning into forms, while form needs to be mapped into meaning in comprehension. Supposedly the form and the meaning representation are not part of the semantics module, then (about) half of the output of the module needs to go to a site representing meaning and the other half to a site representing form; and the same holds for the input. So let us assume for a moment that the module is only involved in production, i.e. the function $Sem()$ maps meanings on forms. Now if we also assume that the sole purpose of this module is to map simple lexicalized concepts on content words, then a semantics module of this type might be capable of doing the job. But as soon as you consider what kind of complex interplay of information for a mapping from meaning to form is required in utterance production, you will soon see that meaning processing is a complex interaction of many functions. For instance, to perform the simple task of

referring to a certain target person (not present) the speaker needs to do several things. As a first choice, the name of the person needs to be checked as a possibility to refer to the absent person. However, a proper name can only be used if the addressee is aware of that person's name. The speaker therefore needs to check whether the addressee has used the name in the past to refer to the target person. This involves probably subconscious retrieval from episodic memory. If the name is not a possibility then a definite description has to be found that allows the addressee to identify the target. This also involves retrieval of knowledge of the addressee about the target person. This is just one simple example of what it takes to process meaning in the brain.

Postulating a *semantics module* or searching for a *semantics area* in the brain is for these reasons not very likely to lead to interesting insights or progress in our understanding of meaning processing in the brain. What we need to do is to specify the different functions that are necessary for the processing of meaning, to spell them out mathematically and in terms of neural computations, and to describe how they can interact with each other and with functions that are not merely involved in meaning processing (such as retrieval from episodic memory). Such a set of functions should be able to learn. This means that the functions each involve a dynamic state s_i that can change during learning. Such a state consists of synaptic strengths for long term learning.

$$\vec{y_1} = F_1(\vec{s_1}, \vec{x_1}) \tag{2.2}$$

$$... \tag{2.3}$$

$$\vec{y_n} = F_n(\vec{s_n}, \vec{x_n}) \tag{2.4}$$

One or several of these functions $F_i()$ might actually have *meanings* as an input and another one might have *forms* as an output, but it is the exact processes, the retrieval of the required information, and especially the interaction between the different functions that we should try to understand. Postulating such possible sets of subfunctions, testing their interaction in computational simulations and using them as a basis for experimental research is likely to result eventually in the understanding of meaning processing in the brain.

In this thesis I will specify some potential candidates of those functions, including their possible neural realization. Each of those functions might be localized in a different area of the brain; some might not even be represented in the cerebral cortex, but in subcortical areas such as the cerebellum and the basal ganglia. And each of these functions might contribute to cognitive processing that does not fall under the broad category of *semantic processing*.

2.2 Brain Structures and Learning Algorithms

Traditionally higher cognitive functions are thought to be represented in the cerebral cortex. Subcortical structures, such as the cerebellum and the basal ganglia were always regarded

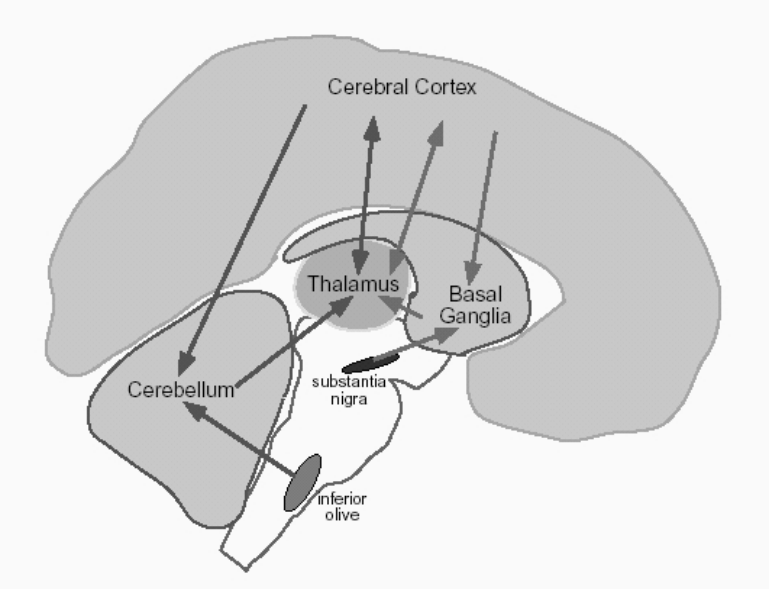

Figure 2.2: Interactions between Cortex, Basal Ganglia and Cerebellum

as being concerned with motor control. However, there is more and more evidence pointing towards the involvement of the cerebellum and the basal ganglia in higher cognitive and non-motor functions (Desmond and Fiez 1998).

Above I sketched the theoretical idea that cortical tissue is basically interchangeable and the function of a region is determined by its connection to motor and sensory information as well as to other cortical areas. This would mean that the way the cortex processes information follows the same principles in every cortical region and that visual processing, auditory processing, as well as higher level cortical processing employ the same basic mechanisms. This actually appears to be case. The cortex appears to be a very big associative memory device in which associations are learned by an unsupervised learning algorithm characterized for the first time by Hebb (1949). The involvement of the cortex in the processing of meaning might therefore be accounted for in terms of a computational description of these general mechanisms, the information flow between the brain and the outside world (in terms of sensory input and motor output) and the information flow between the different cortical areas. In this section I will describe a widely accepted theory about the general mechanism of cortical processing, as well as the data supporting it.

Just as for the cerebral cortex, basic computational principles have also been found for the basal ganglia and cerebellum (see figure 2.2). For the cerebellum it is environmental dynamics and supervised learning and for the basal ganglia it is state evaluation and reinforcement learning. So it is not unlikely that these three brain structures can be described best not by reference to specific cognitive domains, but by their computational and learning principles. This position was stated by Doya (1999).

The three learning types that are attributed to the brain structures are supervised,

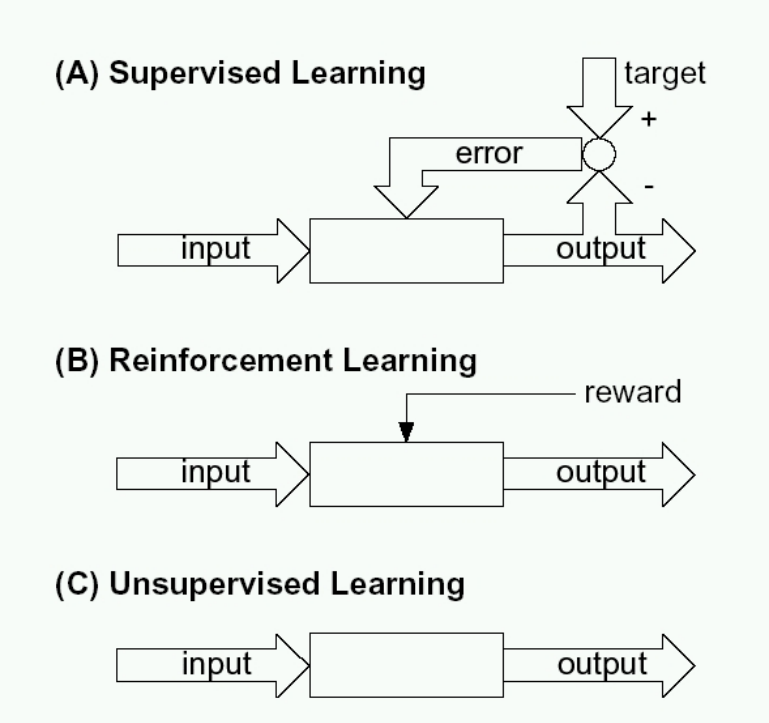

Figure 2.3: Difference between different learning algorithms in terms of error signal.

reinforcement and unsupervised learning. The difference between them is in terms of the *error signal* they use (figure 2.3). Supervised learning requires *directional error vectors*, i.e. information about the specific error of every output neuron. Reinforcement learning uses only *scalar rewards*, i.e. information to what degree the overall performance is good or bad. Unsupervised learning works completely without any error signal. Note that an error signal does not mean that there needs to be a teacher who provides this signal but the signal can be computed from information about the environment.

Different learning algorithms are suited to solving different kinds of learning problems. Supervised learning is very effective in acquiring *internal models* of the environment. The reason for this is that an internal model has to learn a mapping between a current state and a future state. While interacting with the environment this future state is provided by the environment. Therefore, an error vector can be computed from the actual future state and the predicted future state. This error vector is needed for supervised learning.

Reinforcement learning is well suited to learning how to *evaluate* environmental states. The reinforcement learning algorithm is able to do this, because, again, the necessary error signals (rewards and punishments) are provided by the environment (or more precisely - an internal evaluation of the environment). Unsupervised learning is capable of associative learning, acquiring categories and, therefore, suited for the representation of environmental and internal states. Category representations emerge because certain features are generally observed within a certain type of object. To increase the associative links between these es-

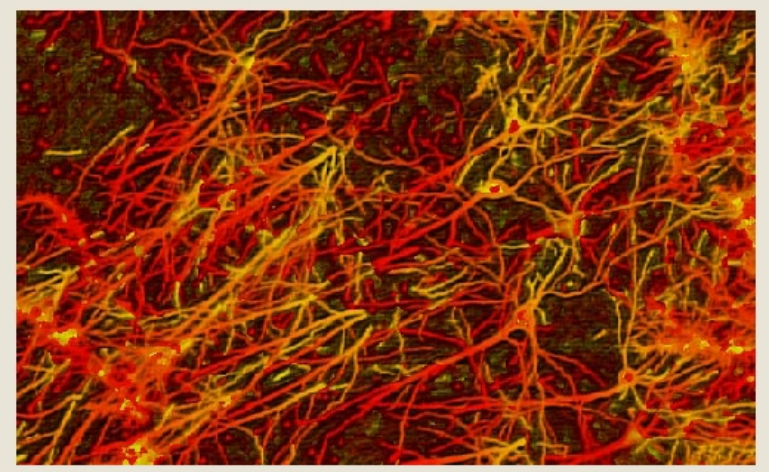

Figure 2.4: A section of brain tissue showing real neurons under the microscope.

sential properties of the category does not require an error signal. Therefore, unsupervised learning is the best algorithm for this kind of learning.

In the remainder of this chapter I will discuss the evidence for the view that the processing in the three brain structures can be characterized in terms of their learning algorithms. As a consequence it seems reasonable to assume that the functions those structures execute are those which their learning mechanisms can acquire. If this is indeed the case, then it is very likely that those structures contribute the same kinds of functions for all cognitive domains, rather then being involved in distinct domains. In other words, a complex function of a specific cognitive domain (e.g. control of arm movement) might be realized by a global network combining the specific learning capabilities of the different brain structures. This view is also supported by the strong connection between cortex and basal ganglia, and between cortex and cerebellum, as well as by imaging studies showing correlated activation between the cortex and the cerebellum (Tamada et al. 1999a).

After discussing the computational principles of the different brain structures in this chapter, the rest of the thesis will be an attempt to describe how the specific abilities of these structures can be assembled into the acquisition and the use of a goal directed communication system. The emphasis will, of course, be put on the meaning aspect of these systems. But before I discuss the computational principles of different brain structures, I need to briefly describe the relation between theoretical accounts of neural computation (that are the basis of simulations using neural models) and neurobiological reality.

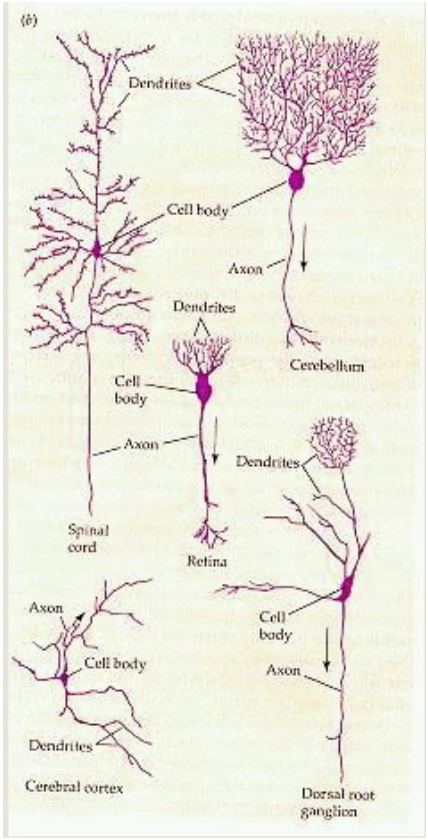

Figure 2.5: The structural similarities of different kinds of real neurons.

2.3 Real Neurons and Artificial Neurons

2.3.1 Neurons

Artificial neurons are models of idealized biological neurons. The degree of similarity between the neuron models of computational neuroscience and real neurons usually depends on the purpose of the simulation in which it is used. While some neuron models pay attention to details, such as ion channels and the speed of electric transmission in the axon, the neuron models used in simulations of higher level functions such as language cannot, usually due to a lack of computational power, go into such details. A good approximation used in most of the simulation in this thesis is the neuron model described in figure 2.6. The output of the neuron in this model is computed by an output function that maps the sum of the weighted inputs (see section 2.3.2) to a so-called *spike rate*. This neuron model, however, does not take into account the timing of single spikes and can, therefore, not be used for the simulation of neural oscillations required in some parts of this thesis. Another model, the so-called *leaky integrator neurons* (equation 2.5) forms the underlying theoretical account of these parts. The leaky intergrator neuron models the temporal

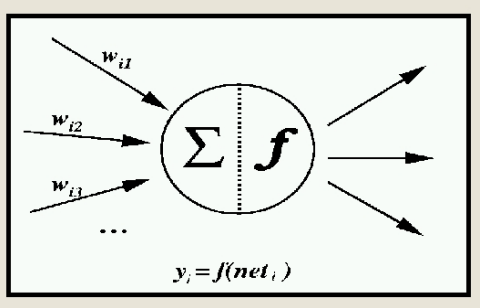

Figure 2.6: A standard neuron model.

dynamics of a neural cell membrane m_i, which receives input from other neurons y_j and leaks a constant amount of charge l while above its resting potential.

$$m_i(t+1) = m_i(t) + \Sigma_{j=0}^{n} y_j(t) - l \tag{2.5}$$

2.3.2 Synapses and Learning

Learning takes place in at the junction between neurons, the so-called *synapses*. Learning is a change in synaptic strength. In most neural models, synapses are modelled by so-called *weights*, usually a factor with which the input coming from another neuron via the modelled synapse is multiplied. The different learning algorithms that I will discuss in sections 2.4, 2.5, and 2.6 use different principles to change these weights. Weights seem to be a pretty good approximation of real biological synapses. But while an artifcial neuron usual gets input from about 10 to 100 synapses, a real neuron (e.g. in the cortex) gets input from up to 10000 synapses (Braitenberg and Schuez 1998). For a more detailed account of the computations happening in real synapses see Koch (1999).

2.4 Cerebellum

2.4.1 Supervised Learning

Supervised learning amounts to the acquisition of an input-output mapping $\vec{y} = F(\vec{x})$ from a set of training pairs $< \vec{x}, \vec{y}^* >$ that specifies for every input $\vec{x}$ in the training set the desired output $\vec{y}^*$. Like in many neural networks, the input to a neuron y_i is computed as the sum of the activations of all neurons x_j that are connected to y_i times the strength (or weight) w_{ji} of the connection from x_j to y_i (equation 2.6). In this way, an output vector $\vec{y}$ is computed from the input vector $\vec{x}$. From $\vec{y}$ and the desired output $\vec{y}^*$ a directed error vector $\vec{e}$ can be computed (equation 2.7) that specifies the error for every output neuron y_i (y_i^* is the desired output of this neuron).

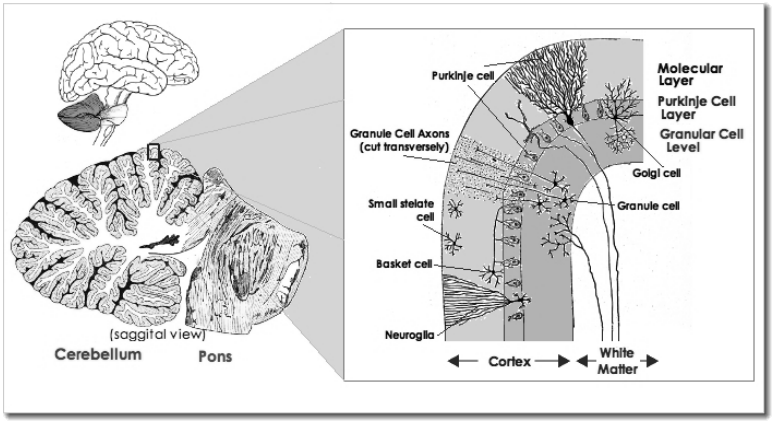

Figure 2.7: Layer structure of the cerebellum

$$y_i = \sum_{j=1}^{n} w_{ji} x_j \tag{2.6}$$

$$e_i = y_i^* - y_i \tag{2.7}$$

The learning algorithm tries to minimize the expected output error by estimating the change for every weight Δw_{ji} in dependence of whether it contributed to the error (i.e. whether the input neuron connected by this weight was active) and also as dependent on a learning rate α that determines how fast the weight is changed (equation 2.8). Then the weights are modified by the computed difference Δw_{ik} (equation 2.9).

$$\Delta w_{ji} = \alpha e_i x_j \tag{2.8}$$

$$w_{ji} \leftarrow w_{ji} + \Delta w_{ji} \tag{2.9}$$

2.4.2 Anatomy

The cerebellum is attached to the brain stem by three major input and output peduncles (i.e. fiber bundles). These are the superior (brachium conjunctivum), middle (brachium pontis), and inferior (restiform body) cerebellar peduncles. There are three sources of input to the cerebellum, in two categories consisting of mossy and climbing fibers, respectively. Mossy fibers can originate from the pontine nuclei, which are clusters of neurons located in the pons that carry information from the contralateral cerebral cortex. They may also arise within the spinocerebellar tract whose origin is located in the ipsilateral spinal cord. Most of the output from the cerebellum initially connects to the deep cerebellar nuclei before exiting via the three peduncles. The most notable exception is the direct inhibition of the vestibular nuclei by Purkinje cells. The cerebellum is covered by an outer cortex (or

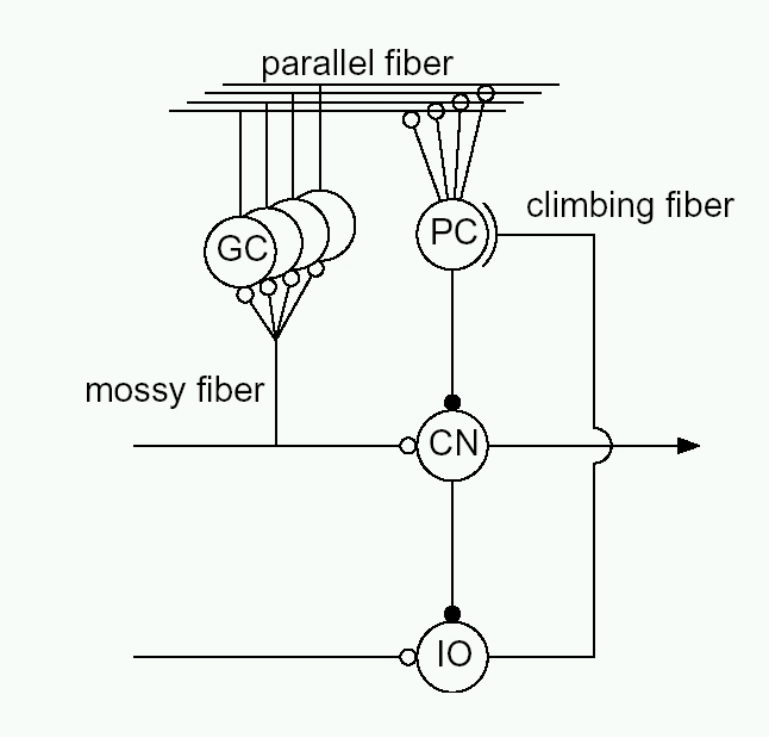

Figure 2.8: Supervised learning in the cerebellum

gray matter) containing neural cell bodies overlying a region that contains predominantly myelinated axons (or white matter). Two shallow grooves running from rostral to caudal divide the cerebellar cortex into the vermis and two hemispheres. The cerebellum has a uniform anatomical organization with characteristic multiple inhibitory pathways, a nearly feed-forward structure, granule and Purkinje cells, mossy fiber input, and climbing fiber input from the inferior olive (Ghez 1991).

2.4.3 Physiology

Physiological experiments have shown simple spikes as response to parallel fiber input and complex spikes as response to climbing fiber input. In motor tasks the climbing fibers seem to transmit registrations of errors in movement. These error signals produce learning in terms of long term depression (Doya 2000)

2.4.4 Theoretical Model

These anatomical and physiological findings suggest that supervised learning in the cerebellum is implemented in the following way: The output of the granule cells is linearly combined by a Purkinje cell (equation 2.6). The error signal e is carried by the climbing fibers, while the output vector x is carried by parallel fibers (Doya 2000)

Supervised learning is often used in multi-layer feedforward networks (trained with backpropagation). Such networks have proven to be able to learn e.g. the internal models which I will discuss and use below. However, the cerebellum clearly is not a multi-layer network. Nevertheless, the strong connections between the cerebral cortex and the cerebellum might allow the two structures in interaction to compute and represent similar to multi-layer (recurrent) networks.

2.5 Basal Ganglia

2.5.1 Reinforcement Learning

The purpose of reinforcement learning is to use experience to decide on a course of action. It is usually based on simplifications such as discrete time, a finite set of discrete states of the environment, and a discrete set of possible actions. A learning system (agent) selects its actions by considering possible future states. The goal of the agent is to maximize the sum of rewards.

To this end the agent trains a mapping of states into actions, which is called a policy π (equation 2.10).

$$a(t) = \pi(x(t)) \tag{2.10}$$

2.5.2 Anatomy

The striatum is the input side of the basal ganglia, receiving afferent projections from the cerebral cortex. Three nuclei comprise the striatum: caudate nucleus, putamen, and nucleus accumbens. There are three nuclei on the output side of the basal ganglia: the internal segment of the globus palidus, ventral pallidum, and substantia nigra pars reticulata. The output nuclei send their axons primarily to thalamic nuclei which, in turn, project to different areas of the frontal lobe. The striatum also consists of two compartments: the *striosome* (which sends output to the dopamine neurons in the substantia nigra pars compacta) and the *matrix* (which sends output to the internal segment of the globus palidus and the substantia nigra pars reticulata (Côté and Crutcher 1991).

2.5.3 Physiology

It was found that cells in the striatum increase firing if a reward is given (Doya 2000)

2.5.4 Theoretical Models

The striosome compartment works as value prediction and the matrix compartment works as an action selection mechanism (equation 2.10) (Doya 2000).

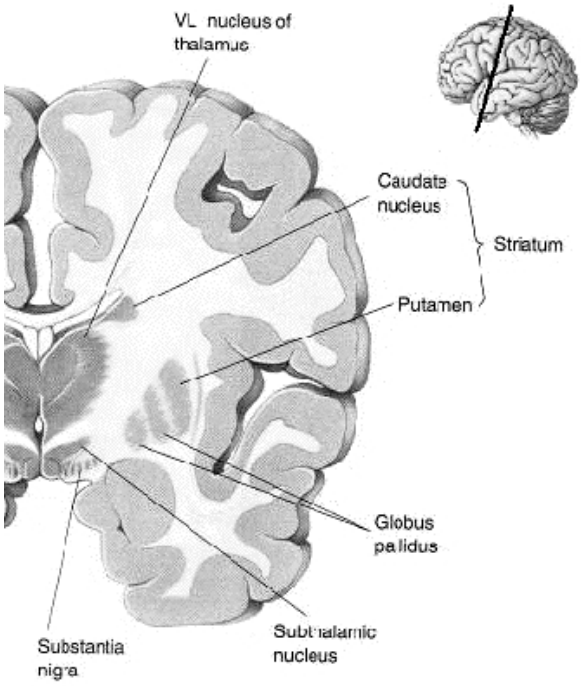

Figure 2.9: Anatomy of the Basal Ganglia

2.6 Cerebral Cortex

2.6.1 Unsupervised Learning

Unsupervised learning learns mutual information that some elements of the vector give about other elements of the vector. This can take many forms and serve several functions. In this thesis the main function of unsupervised learning is the formation of conceptual representations.

2.6.2 Anatomy

The neocortex consists everywhere of a similar six layered structure. It has massive recurrent connections. This homogeneous anatomical structure suggests that all the cortical areas perform homogeneous computations.

2.6.3 Physiology

If from the perspective of neural implementation the conceptual representations which are used to mould our sensation into the perception of entities of different sorts are identical to the conceptual representations which are associated to lexical items as lexical meanings (as the above reflections suggest) then what is known about the neural dimension of perception

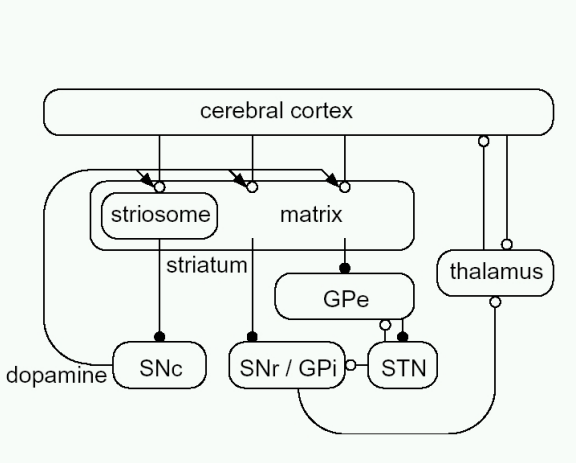

Figure 2.10: Reinforcement learning in the Basal Ganglia

may help us in developing a reasonable theory of how the brain represents concepts used in the processing of meaning. This is important because there is a large quantity of experimental results and theoretical insights about the neural realization of perception processes and their results - very much more than we currently know about the neural aspects of semantic processing. As things stand at present, so much is known about neural aspects of perception (even about vision alone!) that anything resembling a survey of the literature would be out of the question. All I can do is mention a few facts and hypotheses which are pertinent to the ideas which I will develop in this thesis.

First of all, perception is a hierarchical process. There is convincing evidence that at the lowest level certain sensible features - colors like green and yellow, qualities of taste such as sweet and sour, etc - are *represented* by single cells or small cell clusters (Hubel and Wiesel 1962). The activation of such a cell or cluster is the neural correlate of the feature perception, so that the cell (cluster) can be regarded as a *feature identifier* at this level. In particular, the visual field is represented by a field of identifiers for any one of the basic colors (as well as a number of other visually perceptible features) and the activation of any one of these identifiers signals the instantiation of that feature in the part of the visual field to which the cortical position of this identifier corresponds. In these cases it is the direct neural connection between the identifiers and the corresponding detectors in the sensory organs which, it might be said, gives the identifiers their *meaning*. On the other hand the feature identifiers are linked to more complex cell clusters *higher up* in the processing hierarchy. These represent more complex concepts and derive their meaning from their links to the identifiers *below*. Beyond this second tier of cell clusters there are other tiers; and the farther removed a cluster is from the primary sensory areas, the more abstract (from the low-level perceptual features) the concept it represents.

Furthermore, the geometrical lay-out of the brain is such that identifiers are spatially grouped by perceptual channel. Roughly speaking, visual features are represented in ar-

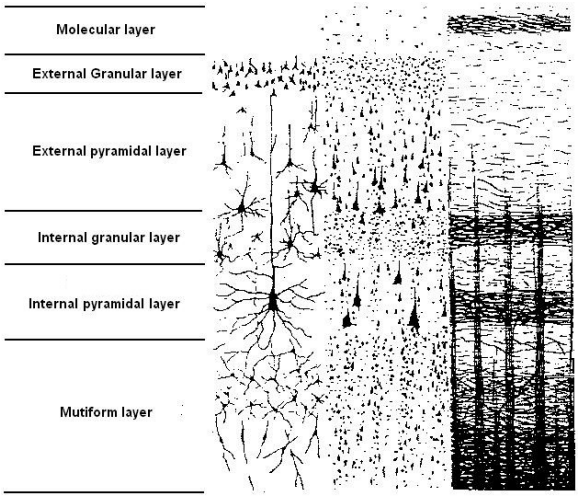

Figure 2.11: Layer structure of the cerebral cortex

eas of the visual cortex, auditory features in the auditory cortices, taste features in the paralimbic cortex, and so on (Fuster 1995). These areas are further subdivided according to feature type. Thus, visual features pertaining to the type of object are processed and represented in the so-called ventral stream while other visually detectable characteristics pertaining to the position and possible actions upon the object are processed in the dorsal stream (Ungerleider and Miskin 1982, Milner and Goodale 1995). These streams are further subdivided into areas representing specific properties of the object. E.g. in the dorsal stream the area MT (mid-temporal cortex) is known to be sensitive to stimulus movement (Hildreth and Koch 1987), while in the ventral stream area V4 is important in color processing (Zeki 1980).

Dynamic Binding

Given this organization of the representation of perceptual features, a perceived individual object will activate neurons corresponding to the perceptual features that are part of the individual object which is perceived. Corresponding to the sensory modalities which the object stimulates (think for instance of its shape, its color, its smell, the sound it produces), these neurons are distributed over different parts of the cerebral cortex. Given such a representation of different features of an object, how can these features be perceived as features of the same thing? This problem has been called the *binding problem*.

The currently most plausible theory about how the brain accomplishes binding was first put forward by von der Malsburg (1981). The theory postulates that features can by dynamically bound by synchronizing the firing of the neurons representing the features of the object. The synchronization of neural activity can be explained by the ability of neurons to detect temporal coincidence (von der Malsburg 1985). This ability can in turn be explained by the properties of the cell membrane (Koch 1999).

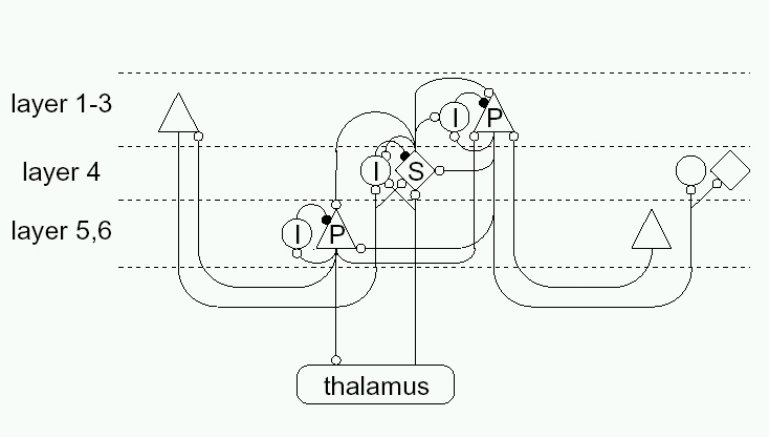

Figure 2.12: Unsupervised learning in the cerebral cortex

Synchronization of the firing of cortical neurons has been experimentally confirmed by multi-electrode recording from the visual cortex (Eckhorn et al. 1988, Gray et al. 1989, Kreiter and Singer 1996). Also EEG studies (for a review see Muller and Gruber (2001)) could show increased γ-band activity for processes related to binding.

Some people argue that for some difficult cases of synchronization, i.e. cases in which the synaptic stengths between the neurons which are to be synchronized are not strong enough, a short term increase of synaptic strength between those neurons is necessary[1]. However, computer simulations have shown that synchronized activity can usually be achieved by the general inhibitory mechanisms of the cerebral cortex (Wennekers and Palm 1997, Knoblauch and Palm 2002a,b).

Another theory about how features of one entity are bound in the brain is binding by spatial attention (Treisman and Gelade 1980, Treisman 1986). This theory is based on the assumption that an attention map is used to represent the location of attention (i.e. by using a position coding, active neurons in the attention map represent the location in space to which attention is directed). Active feature neurons outside of the attention maps are then interpreted as representing properties of the object located at that particular point in space. Due to various sources of evidence, such as patient data[2] and functional imaging studies, this attention map is now considered to be in the intra-parietal sulcus (Freidman-Hill et al. 1995, Robertson et al. 1997, Itti and Koch 2001).

A unification of these two theories was proposed by Robertson (2003): An external source that synchronizes neural responses between cortical feature maps seems to be necessary, and this source seemed to be spatial attention, mediated by the parietal lobe. But there is currently no evidence supporting this claim, so the interaction between mechanisms of dynamic binding by temporal synchronization of neural activation and by spatial

[1]There is also some evidence for such a mechanism in the cortex.

[2]Patients with bilateral lesions of the parietal lobe suffer from Balint's syndrome - one of its symptoms being *simultanagnosia*, the inability to see more than one object at a time

attention remains an exciting topic of research. Since I do not deal with issues of attention in this chapter, I will focus on binding by temporal synchronization.

In contrast to binding in long term memory, which I will discuss below, binding by syncronized activation is (i) dynamic and (ii) temporary (although, as we will see, it can lead to long term binding). Therefore I will call this type of binding *working memory binding*. Once the binding of those entities is accomplished by the mechanisms of synchronization, its activity can be sustained by the prefrontal cortex, as was demonstrated by the computational model of Raffone and Wolters (2001).

Long Term Binding

Some of the concepts we represent are memorized, i.e. the features characterizing them are bound in long term memory from which they can be retrieved again at some later point. Long term binding is assumed to be realized by long term changes of the synaptic strengths (physiologically this is done by lowering the synaptic thresholds). It gets established by a learning mechanism first postulated by Hebb (1949), who introduced the term *cell assemblies* for the groups of cells (or of cell clusters) that get bound together by synaptic strength modification as the result of such learning. Hebb's learning rule was confirmed a quarter of a century later by observations on the hippocampus (Bliss and Lomo 1973). This form of learning is called *long term potentiation* (LTP). It and its counterpart, *long term depression*, LTD) are now considered the general neural mechanisms underlying long term learning in the cerebral cortex (Artola et al. 1990). Note that the dynamically bound representation of concepts by synchronized activation basically consists of neurons which fire repeatedly almost exactly at the same time, and thus facilitate long term binding by mechanisms such as LTP. In other words, synchronized activation as mechanism of dynamic binding and LTP as a mechanism of long term binding fit very well together.

Another mechanism of long-term binding is the formation of so-called *grandmother cells*. A grand-mother cell is a neuron or neural cluster that represents a certain feature combination present on a perceptually lower level. Evidence for this kind of representation can be found in many single cell recording studies. It can also be inferred from the hierarchical organization of the perceptual system.

Hebbian learning and grandmother cells do not exclude each other. They might both exist, but serve different and complementary functions in the brain.

2.6.4 Theoretical Models

There are two main types of computational theories explaining the formation of grandmother cells in the brain, both using unsupervised learning. The first one is the theory of *self-organizing maps* (SOM) by Kohonen (1982), the other is *adaptive resonance theory* (ART) by Grossberg (1980).

There are also computational theories about conceptual representations by means of synchronized activity. In physiological terms, synchronized activity means that neurons *spike* at approximately the same time. To simulate spikes the neuron model needs to

takes into account important properties of the real biological neuron, in particular of the cell membrane. Synapses which connect a neuron with other neurons are generally of two sorts, (i) excitatory and (ii) inhibitory. The cell membrane of a neuron integrates excitatory and inhibitory synaptic activity over time, and *leaks* (i.e. looses) electrical potentials when there is no input. If the membrane reaches a certain *threshold* the neuron spikes. A good mathematical model of this behavior is the well-known *leaky-integrator* neuron. The leaky-integrater neuron has been succesfully used to simulate synchronization. In these simulations the inhibitory synapses function as an inhibition mechanism. This mechanism has been shown to be very important (Wennekers and Palm 1997).

Mathematically, the membrane potential of neurons in a brain area can be given by a vector $\vec{m}(t)$ - a vector of states of membranes of cells located in this area, in which every element m_x of the vector represents the electrical potential of the membrane of one of the cells. The synaptic strength within and between areas can be approximated by a so called *weight matrix* $\mathbf{W}(t)$. The state of a brain area r at time t will then be modelled as determined by the membrane potential vector $\vec{m}_r(t)$ and the area's weight matrix $\mathbf{W}_r(t)$ thus as the pair $< \vec{m}_r(t), \mathbf{W}_r(t) >$. The state of the entire brain at t, $A(t)$, is composed of the states of its regions at t together with further *distal* weight matrices $\mathbf{W}_{r,r'}(t)$ which represent the strengths of the connections between the cells in area r and those in area r'.

$$\begin{aligned} A(t) =< \{< \vec{m}_{r1}(t), \mathbf{W}_{r1}(t) >, ..., < \vec{m}_{rn}(t), \mathbf{W}_{rn}(t) >\}, \\ \{\mathbf{W}_{r,r'}(t) : r = r_1, ..., r_n; r' = r_1, ., r_n\} > \end{aligned} \tag{2.11}$$

As in standard dynamical systems, the next state of the brain, $A(t+1)$ (the state of the brain at time $t+1$), depends in such a model only on the current state $A(t)$, and on the sensory input $i(t)$ which reaches the brain at t. The transition can be schematically expressed as in (2.12)

$$A(t+1) = S(A(t), i(t)) \tag{2.12}$$

The function $S()$ can be made explicit in the following way: The state of the membrane potentials of the neurons in a region r can be computed from the input to r and the neural activity within r. The input will consist wholly or partly of sensory input for those regions which lie directly at the sensory periphery, while for regions that are not directly connected with sensory organs it will consist entirely of input from other regions. The neural activity within a region (i.e. the *spike* activity $\mathbf{F}(t)$ at t in each of the different regions) is a function of the membrane potential in r at t.

Within this framework the acquisition of concepts can be modelled as a reaction of the brain to certain patterns of sensory input. The model assumes that the synaptically determined strengths of links between clusters is changed according to Hebb's rule, i.e. approximately as in equation 2.13.

$$\Delta w_{ij} = c u_i u_j \tag{2.13}$$

Here, u_i is the pre-synaptic neuron (the one that fires), u_j the post-synaptic neuron (the one the receives the input), and c is a variable determining the learning rate.

Such a model might display the following behaviour:

(i) Frequent sensory perception of certain (types of) objects produces very strong connections between the neurons representing the different features of the object.

(ii) At the same time the model will also be able to form concepts involving comparatively weak connections between the cell groups involved. Such concepts will be hard to activate and their activation will require joint activation either of a large number of the features that are part of them or of one or more features which are distinctive of the concept in that they have even weaker (if any) links to other concepts.

The reader interested in such simulations of binding by synchronizations might want to look at the work of Wennekers and Palm (1997), Knoblauch and Palm (2002a), and Knoblauch and Palm (2002b).

Chapter 3

A Goal-Directed Communication System

3.1 The Importance of Considering Goals

Language is used to achieve a variety of goals - from the most simple and straight-forward (getting someone's attention) to the more complex or subtle (touching someone's heart with a piece of poetry). We can use utterances to accomplish goals because utterances have an effect on the world (a direct effect on the mental states of the addressees and an indirect one on states of the world that the addressees can bring about by actions motivated by the change of their mental state)[1]. The importance of considering the effects of utterances was already pointed out by Wittgenstein (1953)[2]. However, these aspects have not yet been taken into account sufficiently by researchers investigating the neural basis of language. To really understand how meaning processing in the brain enables meaningful communcation it is essential to relate the goals and the utterances people make to achieve them in terms of a neural algorithm. The goal of this chapter is to outline how the different learning methods employed by the different brain structures (as described in chapter 2) contribute to the processing of meaning in the brain and how these contributions can be integrated into a theoretical framework explaining the neural processes involved in goal-directed communication.

[1]The possibility of using language to pursue our goals not only gives us a reason to speak, but also gives us reason to learn. For the more precisely we relate our utterances to the relevant properties of the world and our goals, the more *effective* our utterances will become in achieving them. Unfortunatly, this relation between motivation and learning is not discussed in this thesis.

[2]According to von Savigny (1998) the theory of meaning developed in Wittgenstein (1953) can be summarized in the following way: the meaning of an utterance is the (context) conditions under which it can be used and the result this utterance leads to.

3.2 The Overall Architecture

3.2.1 Essential Cognitive Functions

As explained in chapter 2, the cerebral cortex, the basal ganglia and the cerebellum provide us with three basic learning mechanisms: unsupervised learning in the cortex, reinforcement learning in the basal ganglia and supervised learning in the cerebellum.

I will relate these learning algorithms (and brain structures) to three cognitive functions that are essential to the creation of a goal-directed communcation system: *conceptualized state representation*, *state evaluation*, and *internal model*. The best way to learn these three functions is by unsupervised leaning, reinforcement learning, and supervised learning, respectively. Conceptualized state representations are needed for an efficient coding of relevant information about actual and possible states of the world (including mental states). States are evaluated in order to determine how desirable current and possible states are - in particular states which can be brought about by the speaker (either by some non-verbal action or by using a communicative act to get another person to bring about this state). Finally, it is the function of the internal model to compute which states are brought about by which verbal or non-verbal actions in which contexts.

3.2.2 A Formal Description of the Architecture

The exact nature of the (conceptualized) state representation will vary considerably throughout this thesis and will become more and more complex from chapter to chapter. For now, it is sufficent to say that a state representation s is a variable containing the information represented in the brain about a current or potential state of the world.

Such states of the world are evaluated by a so-called *value function* $V()$. The value function maps states s onto real numbers (values). The higher the value, the more desirable is the state.

The third component, the internal model, computes which verbal actions (utterances) u (or non-verbal actions a) bring about which states s of the world. Internal models can be either *forward* or *inverse* models. A forward model is a function $F()$ which computes the next state s_{t+1} from the current state s_t and the utterance (equation 3.1). An inverse model is a function $I()$ which computes the utterance which will bring about the change from the current and the state s_t to the desired state s_{t+1} (equation 3.2).

$$s_{t+1} = F(s_t, u) \tag{3.1}$$

$$u = I(s_t, s_{t+1}) \tag{3.2}$$

Both functions can be used in utterance selection. A major question that will be addressed in this thesis is whether the brain uses forward or inverse models, or both. It might also need additional mechanisms, as we will see by the end of this thesis.

The proposed architecture can select utterances by an interplay between value function and internal model. Depending on whether the system uses a forward model or an internal model there are two possible ways of selecting utterances.

With a forward model the system would use an algorithm described in equation 3.3. Here, the forward model $F()$ predicts the outcome of every possible utterance in the current context and each outcome is mapped onto a value with the value function $V()$. The utterance yielding the highest value is chosen.

$$u = argmax_u V(F(s_t, u)) \tag{3.3}$$

The inverse model $I()$ needs the desired state as an input and therefore starts by computing a value for every possible state with the value function $V()$ (equation 3.4). From this state s and the current context s_t, the utterance that will bring about the transition from s_t to s is computed with the inverse model $I()$.

$$u = I(s_t, argmax_s V(s)) \tag{3.4}$$

3.3 Cognitive Components in More Details

3.3.1 The (Conceptualized) State Representations

The state representation s is a representation consisting of a set of concepts. To be more precise, the state representation includes categorized information about the agent himself, the outside world, other agents, etc. I will briefly describe it here. The exact structure of the state representation will be developed during this thesis.

First of all, the state representation $s(t)$ includes information about the sensory accessible world, i.e. everything the agent is able to perceive at a point t. This is already a tricky business. Suppose there is an apple to the left of the agent and a pear to the right and the agent is not able to look at both objects at the same time. A realistic theory about the representation of sensory input should include some attention component, and also some short term or working memory device which allows the brain of the child to represent the items and events of the suroundings. To keep things simple, I will in my first computer simulation (described in chapter 5) assume that all objects of the world are visually available simultaneously at all points in time. However, it will become necessary at some point to include in the state representation also objects which are not part of the immediate surroundings, viz. objects the agent encountered at some time in the past and still remembers. The same holds for the properties of these objects. It must be possible for some of those properties to be accessible at the current moment while others are remembered.

Besides objects, the state representation also needs to include events, such as actions. Furthermore, there are also facts concerning the internal world or internal states of the

language user/language learner that the state representation needs to take into account, for example, whether the agent is hungry, thirsty, or sleepy, to name only the most basic ones. More complex types of representations include our knowledge about the mental states of other agents and our knowledge about our own mental states (including our own knowledge about our own knowledge).

Humans are not born with the concepts of objects, actions or events preestablished in the brain. Children must learn them. The cortex, due to the learning algorithm it uses (as described in chapter 2), is the optimal device for acquiring concepts. The sensory information a child is provided with comes in the form of innate basic sensory features represented along with innate basic motor features in the primary sensory and motor areas of the cortex. When children learn to categorize objects, actions and states of their environment, more complex motor and sensory concepts emerge in higher motor and sensory areas, as well as in multimodal areas of the cortex (integrating areas). Through processes of association and abstraction, the cortex learns that certain features always appear together and thus a concept is formed. This learning leads to a more efficient coding of the representation of the world. Many concepts are formed long before the children learn what these objects and states are called.

Exactly how these concepts are formed, represented and processed, and what neural mechanisms are involved, is the main topic of chapter 4.

3.3.2 Evaluation

The value function V maps states to values. The input state can be the current state or any other possible state. The values asigned to the states are estimations of how good it is for an agent to be in that state. The value is a positive or negative real number expressing how much an agent desires a particular state. The value function is defined in terms of future rewards that can be expected, or, to be precise, in terms of the *expected return* (Sutton and Barto 1998). The expected return is the sum of discounted rewards, which can be expected in and after the given state s_t (equation 5.1).

$$V^{\pi}(s_t) = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\} \tag{3.5}$$

$V^{\pi}(s_t)$ is the estimation of the value of state s_t at (discrete) time step t under a *policy* π. A policy is a function mapping every state onto an action. The choice of action has a strong influence on the next state s_{t+1} of the world.

The γ-parameter is the discount factor, which determines the value of future rewards. The higher the value of γ is, the more importance is given to future rewards.

3.3.3 Internal Models

Internal models are called *internal models* because they are models *in* the brain of the dynamics of the *outside* environment. Usually, internal models are used in motor control,

but in this thesis I will apply the concept to language use. In this context, internal models are models of the effects of certain verbal actions (i.e. utterances) on the state of the world (i.e. the agent's internal representation of it). Children can train their internal models by being exposed to utterances used by themselves and others and observing these utterances to have specific (indirect) effects on the world as they perceive it. This includes experienced (and, at a later learning stage, inferred[3]) mental states of the addressees. A related approach has been proposed by Langacker (1987).

It must also be assumed that the child can experience the ways in which the effects of a given action vary depending on certain configurations of the world - in other words, that these effects are context-dependent. Again, the context includes experienced (and, later, inferred) mental states of the addressees.

As I have explained in detail in chapter 2, the principal capability of the cerebral cortex is to detect and store correlations of events and states in the environment. It is, therefore, not unlikely that the correlations between expressions, context configurations and effects are stored in the cortex. Those stored correlations would form a probabilistic internal model which can be used to predict the probability of the effect of a certain utterance in a particular context. However, since supervised learning is the best learning algorithm to acquire internal models (because actual and desired output are given), I am assuming that the cerebellum plays a central role in the acquisition of the internal model. But, as I will discuss later, the cortex might nevertheless store those correlations and use it to reduce the number of necessary computations significantly.

3.3.4 Other Essential Functions

Of course, there are other (and maybe more basic) cognitive functions needed to support the acquisition and use of these three functions. One important function is *intention reading*. This is now regarded as a necessary prerequisite for word learning. Therefore, any theory of language acquisition must take intention reading into account. Intention reading develops in several steps, from simple (and possibly instinctive) gaze-following to the child's ability to use a number of complex social and pragmatic cues to determine the referent of a word.

Another important cognitive function which the theory presented here presupposes is *imitation*. To acquire symbolic communication by imitating an adult, role reversal imitation has to take place (Tomasello 2003). In comparison to normal imitation behavior, the child must not only substitute herself for the adult as an actor, but also substitute the adult for herself as the target of the intentional act. In this thesis, I will assume the capacity for imitating behaviour as given.

Another essential capability is the ability to acquire and represent word forms (including the ability to segment speech and recognize words). I will discuss this capability in some more detail in section 4.2.

[3]As explained above, the *theory of mind*- capabilites are not developed before the age of four and until that time the child is not able to draw such inferences.

3.4 Remarks on Utterance Comprehension

The focus of this thesis is on the production and not on the comprehension of utterances, and I will not deal with mechanisms of comprehesion in detail. Nevertheless, I would like to propose a simple form of comprehension which does not require any extra mechanism or algorithms in addition to the ones described so far. The mechanism I will describe is simple, but I think it nevertheless captures an important aspect of language comprehension.

The framework introduced so far describes a system which can store the relation between utterances, contexts and effects in an internal model. A speaker A_1 can use it to select an utterance u when its estimated effect d is desired.

If speaker A_1 and addressee A_2 speak the same language (which I assume throughout this thesis), they should have represented the same relations between expressions, contexts and effects. In other words, A_2 has the very same internal model as A_1 (or at least a very similar one) and can use it (a forward model is required for this process) to map u directly onto d. In that way A_2 can compute the desire d (which has triggered the usage of u) of A_1 from A_1's utterance u.

There is also empirical evidence of a neural mechanism that maps first-person experience onto third person experience in monkeys (Rizzolatti and Arbib 1998), which makes it pretty likely that such a mechanism also exists in humans.

Chapter 4

Basic Units of Meaning

4.1 Form and Meaning

The enterprise of creating a neural theory of meaning representation and meaning processing that is undertaken in this thesis is based on the assumption that in the brain there is a level of representation of language form and a level of representation of language meaning. The language forms that play a role in this thesis are word forms, simple sequences of words, as well as topic, comment, and speech act markers. I will dedicate the next section to a brief discussion of how word forms could be represented in the brain, because word forms are the most important form unit in this thesis. However, since the purpose of this thesis is not to develop any theoretical account of language form, I will use simple representations of form that do not take into account questions of biological plausibility.

To talk about the representations of meanings is far more controversial. First of all, it is not even clear what a representation of a meaning is, since the notion of meaning is already far from clear. In this thesis, two major types of meaning representations are used. The first one are representations of individual objects and object categories functioning as the meanings of content words. The second type are representations of current states and desired states, functioning as descriptions of circumstances of utterance use.

For us to communicate, the brain has to be able to transform the representation of meaning to a representation of form in language production, and to transform the representation of form into a representation of meaning in language comprehension. These two transformation processes cannot simply be understood as a two way channel, but they differ at least in some of the brain areas involved. For auditory speech comprehension, we definitely use the primary auditory cortex and higher auditory areas, while for speech production, we use the primary motor cortex and higher motor areas. Of course, it would be uneconomical for the brain to represent e.g. semantical or phonetical information twice in the brain - once for production and once for perception. This argument from economy has lead to theories such as the motor theory of speech perception, which claims that motor representations are also used in speech perception. Also, the more we get away from the primary sensory and motor areas, and the higher the language related cognitive processes

are, the more likely it is that the same representations are used. Still, the transformation of any type of meaning to any type of form might be a process of a different nature than the inverse transformation, and involvement of similar processing units cannot by any means be taken for granted.

If you regard the complete transformation processes - starting from the perception of sound to the complete understanding of the utterance, including implicatures, allusions etc. , the representations of form involve different representational levels with different types of information and probably different brain areas; and the same holds for the representations of meaning. What those representation are, where they are, exactly what information they code are questions which to answer would involve a summary of the research of almost all fields concerned with language. Since this thesis is about the creation of a computational theory of meaning in the brain, I will focus on the meaning side of the representation (and even here I will only be able to deal with the levels and aspects that I consider to be the most central).

In this chapter, I will introduce what I consider to be the most fundamental neural representation of what might be called the basic unit of meaning: the *concept*. Before I discuss the representation of concepts in the brain, I will give a brief outline of the representation of the most important unit of form in the brain: the *word form*. In this outline of the representation of the word forms, I do not even get close to a survey of the literature. Nevertheless, my description should suffice to give the reader an idea about the complexity of word representations in the brain. In particular it should clarify what I mean by *word form representations*, a notion that will be used throughout the rest of this thesis.

4.2 Representation of Word Forms

4.2.1 Functional Neuranatomy

In the following discussion of the experimental literature on the representation of word forms in the brain, I will place special emphasize in the relation between production and perception and the question in how far representations used in production appear also to be used in perception and vice versa. As a starting point I would like to contrast two classical and widely known models of how words are represented in the cortex. The first one is Lichtheim's model of single word processing. Lichtheim (1884) postulates an auditory speech center projecting to a center of concepts during language understanding. The concept center projects to the motor center of speech during production. The model postulates one concept center used in both production and perception. Word forms, however, are proposed to be represented independently for production and perception and the model postulates independent form-meaning interactions for production and perception. The auditory speech center further projects directly to the motor center (for repetition). The second model is the Wernicke-Geschwind-Model of language (Geschwind 1965), which is widely accepted and still used in contemporary neurology. Broca's Area is postulated

to be the center of motor control of speech (i.e. it transforms phonemes into articulatory programs). Geschwind postulated that phonemes and word forms are represented in Wernicke's Area (BA 22) and that these representations are used in both speech perception and speech production. The angular gyrus (GA), a major multi-modal association area (BA 39), was not only postulated as an essential structure in reading (transformation of letters in phonemes), but also claimed to play the fundamental role in all kinds of transformation of visual information to language (e.g. in picture naming). This model was based on the work of Carl Wernicke (1874) who, through post mortem analyses of patients with brain lesions, was able to show that comprehension of speech was impaired when the lesions were located in the posterior BA 22 (superior temporal gyrus). Since Wernicke's early studies much information has been gathered about the role of the superior temporal gyrus (STG) in speech perception.

But speech perception does not start in STG. The first cortical area of relevance for speech is the primary auditory cortex (A1). Its tonotopic organization has been demonstrated in monkeys with single cell recordings (Rauschecker et al. 1995). Neurons in these areas are responsive to pure tones, and narrow frequency bands. The tonotopic organization which has been demonstrated in monkeys could be confirmed in humans with fMRI (Wessinger et al. 1997). A1, in turn, projects to the lateral belt areas (Hackett et al. 1998), i.e. the secondary auditory cortex (A2). The responsiveness of the lateral belt areas to more complex sounds has also been demonstrated in monkeys (Rauschecker et al. 1995). The superior temporal gyrus, which includes the lateral belt areas, projects to the supplementary motor area (SMA), which, as we will see, is important in speech production (Deacon 1992). Projections from the lateral belt to the superior temporal sulcus (STS) were found by Hackett et al. (1998). The responsiveness of the lateral belt areas to more complex sounds has also been demonstrated in humans with functional imaging (Wessinger et al. 2001). The anterior superior temporal sulcus (STS) projects widely to a-modal association cortices (Jones and Powell 1970), as well as to lower order visual areas (Maioli et al. 1998). Deacon's major finding is that the most extensive projection to prearcuate (frontal) regions comes from STG and even denser projections come from STS. No connection was found from STS to A1.

Evidence for the involvement of auditory areas in speech production comes from clinical data. Important production disorders that can be caused by temporal lesions are paraphasias (Damasio 1992), a problem attributed to errors in the process of phonologically encoding words. Severe Wernicke's Aphasia can lead to severe anomia (Goodglass 1997), the inability to find high information words. Those two phenomena might both be caused by problems of phonological encoding, i.e. a disorder or a disability. The respective words are often preserved in comprehension (Goodglass 1997). This, however, does not necessarily indicate different representations in perception and production, since it can be explained by independent access routes. The counterpart of anomia in perception is *word deafness* (WD) with intact word and speech production and intact perception of complex sounds. Bilateral lesions in STG result in pure WD (Tanaka et al. 1987). WD can be caused by several cognitive deficits, such as the inability to recognize word forms or to connect the detected forms with a meaning. Evidence for the involvement of the

superior and middle posterior temporal gyrus in prelexical phonological processing, e.g. in perception of an unfamiliar language, was found by Mazoyer et al. (1993). It could also be shown that speech sounds are processed more ventrally than non-speech sounds (Zatorre et al. 1992). Two PET investigations observed activation of the superior temporal gyrus in speech perception. One study showed activation of the left anterior superior temporal sulcus only for intelligible speech (Scott et al. 2000) while the second found activation in the superior temporal sulcus during performance of a verbal fluency task. The authors, therefore, postulate this region to be involved in the acquisition of phonetic sequences (Wise et al. 2001). This later PET study further observed that supra-temporal cortical plane responding to both speech and non-speech stimuli and that posterior media at the junction to the inferior parietal cortex responds to speech production.

In the frontal lobe it is especially SMA and the primary motor cortex (M1) that are of interest for speech. SMA is involved in the representation of complex motor programs in general, which would imply a major role in speech motor control. SMA projects to the primary motor cortex (M1 or BA 4). M1 is organized somatotypically (Greyer et al. 2000). M1 and the premotor cortex (BA 6) are reciprocally connected, (Pandya and Kuypers 1969, Jones and Powell 1970). Price et al. (1996) could show with a PET study that the activation of what the authors called *posterior Broca*, i.e. BA 44/6, correlated with the phonetic encoding of phonological word form into an articulatory plan. Their assumption is that words are stored in the form of auditory information. The brain area was found by subtracting the brain activity obtained during a condition in which participants had to listen to words from the brain activity obtained during a condition in which the same participants had to repeat words. The authors therefore claim this area to be required for motor plans including speech. An articulation task in the same study activated the sylvian sensorimotor cortex bilateral in tongue, mouth, and respiratory areas, proving further evidence for the involvement of auditory areas in speech production. It is further notable that stimulating M1 (with an electric wire) produces rudimentary vocalizations, while no other area could trigger these responses (Penfield and Roberts 1959). Hence it seems that the primary motor cortex includes representational structures that directly trigger motor output (Ojeman 1983).

The area that has been known to be related to language for longer than any other area is Broca's area (Broca 1865). But to this day it is also the most controversial brain area. Its location was redefined several times and is now relatively well agreed to be Brodmann's Areas 44 *and* 45 (Damasio 1992). The role of Broca's area in the motor control of speech is disputed, as *apraxia*, the inability to coordinate speech movements with unimpaired perception of speech, seems to be caused by lesions in BA 4/6 and not by lesions in BA 44/45 (Levine and Sweet 1982). The integration of linguistic knowledge in brain research on language gave rise to the alternative view that the areas 44/45 are housing syntax (Zurif 1982). Is is now widely agreed that the main symptom of damage to areas 44 and 45 is agrammatism, but there is still evidence supporting an important role of Broca's area in speech motor control as well.

4.2.2 Learning of Words

For the learning of speech sounds it is essential that the brain learns that speech motor commands produce specific sounds. To imitate the sound of the parent's language, children need to predict the sound of their speech motor commands (or find a motor command to produce a specific sound). This is done by so-called internal models (of which I already talked in chapter 3 and will talk a lot more in the chapters to come). The internal models of speech motor control are trained during the babbling phase, in which sequences of motor commands are produced and, by means of auditory feedback, are associated with sequences of auditory events. The motor representations can get a strong bidirectional link with the auditory pattern that occurs when the produced sound is perceived. The knowledge introduced by this link can be seen as the ability of the child to intentionally produce a certain sound. These internal models of speech sounds can be seen as representing an established connection between the motor and the auditory aspects of speech sounds. A word form can then be seen as a sequence of such speech sounds. Of course, to give a full account of word forms, other aspects than the production and perception of sounds have to considered, such as the word's morphological properties, its syntactic behavior and, in case the agent is literate, its orthography. But in this thesis they will not be dealt with.

4.2.3 Computational Aspects

One might imagine a simplified computational model of word form representations in the following way: During the babbling phase motor events are triggered in M1. The produced sequences are detected, encoded and represented in SMA. The role of SMA is also to reactivate these sequences of motor events in M1. The sequences of articulatory movements lead to sound, which is registered and analysed in A1. A2 detects auditory features. Sequences of auditory features are detected in STS. With the motor sequences in SMA are then associated auditory sequences in STS. The resulting structure is the auditory-motor representation of sounds. These sound representations can, if required, reactivate the motor sequences which produce the respective speech sounds.

The learning of word forms starts with the auditory processing of the sound of the word. Sequences of sounds are detected and represented in STS. They can, then, be associated with some form of semantic representation. In this theoretical sketch, it is assumed that a word form is perceived when a familiar speech sound sequence is redetected. This takes the form of reactivation of a representation of a sequence of auditory features. If this sequence representation is meaningful, i.e. if it was associated with some form of meaning in the past, then it will activate this meaning representation. During naming, the (visual) meaning representation can reactivate the word form representation is STS, which in turn activates the auditory part of the representation of the sound sequence. The auditory representation of the sounds sequences can in turn activate the motor commands that produce the sounds of the word form, i.e. it can trigger the execution of the respective motor events by M1[1].

[1]For a mathematical formalization of this model, a description of its implementation, and simulation

4.2.4 Abstract Word Form Units

This theoretical sketch of the representation of word forms in the cerebral cortex is just one of several theories. It is in contradiction with several of those other theories, such as the motor theory of speech perception (Liberman and Mattingly 1985) which states that the object of speech perception are the intended phonetic gestures of the speaker represented in the brain as invariant motor commands. However, no case of word deafness caused by frontal lesions is known to me. Also, there exist hardly any reports that lesions in frontal areas result in disturbance of speech perception.

My theoretical sketch also ignores the undoubted role of the cerebellum in speech motor control. However, my only purpose here in describing how word forms might be represented in the brain, is to show that, despite the fact that a word is sometimes produced (involving motor modalities of the brain) and sometimes perceived (involving auditory modalities), it is not unlikely that there is some abstract representation of word form in the brain which is used in both production and perception. It is this abstract representation that I will refer to in the rest of this thesis when I will talk about word form representations.

4.3 Representations of Concepts

4.3.1 The Basic Unit of Meaning

The focus of this chapter is a theoretical investigation of *concepts*. Concepts are a suitable anchoring point in an interdisciplinary investigation of meaning in the brain such as the one untertaken here because linguistic semanticists, psycholinguists, psychologists, and experimental and computational neuroscientists all use this term. However, there is some disagreement upon the exact function of a concept and the information it carries. In general, the term *concept* refers to a cluster of information that can be used to categorize objects, actions, or situations. This means that the concept is, first of all, a representation used to categorize perceptual information. Of course, there are concepts which cannot be directly grounded in perceptual information. However, it is reasonable to assume that even those are linked to concepts which *are* grounded in sensory and motor experience via various kinds of abstractions. If such a concept has a certain word associated with it, then this concept carries the information as to which objects, actions, or situations the word associated with it can be applied to. Such a concept I will call a *lexicalized* concept.

The theory that it is one and the same type of conceptual representation which (i) carries the information necessary to categorize non-linguistic perceptual information and (ii) carries the information which constitutes the meaning of a word cannot be taken for granted and is not accepted by everybody. Also, there is some evidence, summarized by Levelt (1989), that between the word form representation and the conceptual representation there is another level of representation called the *lemma*, which is a combination of semantical and syntactical information and has an important role in the production of

results, consult Klein and Billard (2001)

complex utterances. The existence of the *lemma level* is disputed.

It is, however, beyond doubt that the representation of perceptual categories and the representation of words which can be used to refer to objects, actions or situations of this category must have a very close relationship. That is the case for several reasons:

The first one is that people can verify linguistically presented information by perceptually processing the world and they can easily describe perceived scenarios.

Second, from an ontogenetic point of view, children learn to organize what they perceive of the world into entities of different categories before they learn their first words. It is, therefore, justified to assume that they acquire their first set of representations of categories (concepts) based on non-linguistic perception alone. Then they learn that objects belonging to such a category can be referred to by a certain word, i.e. they lexicalize some of their acquired concepts. This in turn makes it very plausible that the child associates the word representation with the category representation it has already acquired.

Since there is no evidence supporting the view that the conceptual representations used in perception are different from those used in language processing, I will - at least until contratictory evidence is found - subscribe, in keeping with Occam's Razor, to the theory that the concepts used for categorizing perceptual information and those used to represent the meaning of standard content words are identical.

How concepts are implemented in the brain in terms of neurons and synapses is not known and the theoretical assumptions about this issue diverge considerably. Below, I will discuss some of those approaches, adopt the most plausible ones, and extend them to a neural theory of concepts that can meet some of the most important requirements for the processing of meaning in the brain.

But before doing that I will try to explain the important function that concepts have in the representation and processing of meaning in the brain in more detail. I will use the theoretical notion of *predication* - arguably the most important building block of the entire edifice of linguistic semantics - and explain how concepts are used to implement its function.

4.3.2 A Neural Theory of Predication

As most linguists and philosophers see it, predication is an operation in which one entity, the *predicate*, is applied to another, the *predicable*; the result of the application is either positive, when the predicate is true of the predicable, or negative, when it is not. In more psychological terms, predication could be regarded as the process of attributing a category to an *individual*. Here, *individual* is a broad term for individual *instances* of categories of objects (including e.g. humans), abstract entities, events, states, and actions.

An important aspect of this conception of predication is the view usually associated with Frege (1879) that predicate and predicable are entities of fundamentally different sorts. The predicable has a reality all of its own (it can, so to speak, stand on its own feet), but the predicate has only a kind of virtual or potential reality - it yields something of genuine reality only through predication, i.e. when it is combined with a predicable and the result is either a *truth* or its opposite.

When we try to relate this apsychologistic (or, as some would see it, antipsychologistic) conception of predication to the mental act of predication - to what it is people do when they attribute a property to a thing and to what goes on in their heads when they do so - we run into all sorts of problems. In fact, as soon as one starts to think about the relation, one can't help noticing that there isn't just one type of mental act that corresponds to the linguists' and philosophers' notion of predication, but a number of them, and it is crucial not to confuse these. Mental acts of predication involve particular ways of representing, and sometimes recognizing, both the predicate and the individual involved, and the recognition and representation of these two components of the predication may vary significantly from one instance to the next. Consider sentence (1).

(1) The keynote lecture was boring.

Igoring the questions of how the brain deals with the definite description, for example the problem of how a particular instance of *keynote lecture* is selected in the context, this sentence involves two predications, involving the categories *keynote leture* and *boring*. Further, it is important to consider the processing difference between the production of a sentence and its comprehension. In the production, the speaker who wants to give the addressee some information about an individual entity has to select an appropriate category for an individual entity (represented in her brain) to refer to it (predication 1). This category has to be selected in a way that allows the addressee to identify the particular individual. If the speaker would have chosen another category, as in sentence (2) then the addressee might identify a different individual (e.g. considering that this conversation takes place at a conference with several talks each day).

(2) The talk was boring.

To transmit information about the selected individual, the speaker, then, has to select another category - a lexicalized concept that represents the property which the speaker wants to communicate (predication 2). This selection does not have to deal with the problems of referential uniqueness, since this predicate is not used to single out a referent, but to give information about it. However, it has other problems which I will discuss later in detail. In contrast to these processes involved in production, in sentence comprehension the addressee has to identify a specific individual based on the categorial information given to him by the speaker (predication 3), and then attribute a certain property to it (predication 4).

In theoretical linguistics (where one is not dealing with the cognitive processes involved in production and comprehension), these four types of category attributions are all treated in the same way. In the light of such reflections what little current semantic theory has to say about predication may not seem all that helpful; at the very least it is incomplete, and needs to be complemented by an account of the various ways in which the mental representations of predicate and predicable can be combined so as to yield, at the level of consciousness, the sense of a predicational judgment. Below I will formulate some first

hypotheses about what the representations involved in these mental predication operations may be like and about some of the ways in which they can interact to produce an act of predication.

An assumption already made above is that lexicalized concepts (i.e. concepts which have a word associated with it) carry the information as to which objects, actions, or situations this word can be applied to. For example, there is a word *banana* and a conceptual representation of a banana which carries information on the basis of which a speaker is able to determine whether what he sees is or is not a banana. The concepts which will represent this kind of categorical information will be called *categorical concepts*. But not only the categories are represented in the brain - individuals need to be represented as well. Humans are able to remember individual entities, such as, for example, the particular banana which I put in the refrigerator yesterday, or the smaller one of the two bananas on the table. The representation of such an individual entity in the brain (whatever that may be) I will call an *individual concept*. The application of the term *concept* to individuals might seem to be contractictory to cognitive psychologists who usually distinguish concepts (i.e. mental representations of categories) from *instances* of the concepts. I hope that the rest of this chapter makes clear why I have chosen similar terms for both types of representations; the reason is that one aspect of my story will be that between individual concepts and categorial concepts there is not the fundamental difference that linguists, philosophers and psychologists assume exists between predicates and predicables or categories and their instances.

A theory of mental predication should, at the very least, have something to say about the following mental acts and processes: (i) the acquisition of a categorical concept through perception, (ii) the acquisition of an individual concept through perception, (iii) categorization of a perceived object as belonging to a certain known category, (iv) identification of a perceived object as a known individual, (v) retrieval of an individual concept based on the understanding of a denoting predicate or predicate complex, and (vi) modification of an individual concept through communication.

In contrast to other aspects involved in the processing and acquisition of meaning, I assume that conceptual representations are purely a cortical matter and, therefore, are based on the computations and representations attributed to the cortex, as outlined in chapter 2. The two major points made there were that (i) certain neurons in the sensory areas respond to certain features in the environment and (ii) synapses between neurons in many areas increase their synaptic strength when the connected neurons fire at approximately the same time. Based on these and other findings, there is a growing consensus that basic concepts are neurally represented by comparatively strongly connected neurons, which thereby form so called *cell assemblies* (Hebb 1949); coordinated firing activity in such an assembly signifies activation of the represented concept. A longer discussion of the physiological data and the theoretical modelling of this data can be found in section 2.6.

Based on the insight discussed there, I will now outline a mathematical description of the behaviour of neurons and synapses involved in long term and short term binding. This description is on a very high level of abstraction. It is still compatible with the physiological data discussed in chapter 2.6 and can be implemented using leaky-integrator

neurons. However, it allows for a much more concise description of the states of neurons and synapses as far as they are involved in conceptual processing.

The starting point of this framework is, as discussed above, the fact that different features are localized in different brain areas. Therefore the *conceptual* state of the brain C of an agent A_1 at time t is represented as a feature-value matrix of different brain areas (equation 4.1). Here VFA is for visual form areas, $V4$ is for visual color areas, and OA is for olfactory areas. Note that these areas are functional areas and not anatomical labels, i.e. they are the areas in which these functions are represented. Since there might be still a dispute over which anatomical areas these functions are located in, I do not want to make any strong antomical claims here. However, my choice of areas is motivated by data on the localization of brain functions.

$$C_{A_1}(t) = \begin{cases} VFA : ... \\ V4 : ... \\ OA : ... \end{cases} \tag{4.1}$$

In each of these areas there are representations (neurons and neuron cluster) of certain features, again represented as a feature value matrix (equation 4.2).

$$C_{A_1}(t) = \begin{cases} VFA : \begin{bmatrix} feature\_1 : & <\{\},\{\}> \\ ... & ... \\ feature\_n : & <\{\},\{\}> \end{bmatrix} \\ V4 : ... \\ OA : ... \end{cases} \tag{4.2}$$

The value of the feature is a pair of sets of labels. The first set of labels represents the activation and binding of features, i.e. working memory. If features are bound, then they have the same label. Having the same label means that the activation of the neurons representing the features is synchronized with the activation of the neurons representing the other features. In equation 4.3, feature 1 and 2 are active but are not synchronized, i.e. they are features of different entities. However, feature 2 is sychronized with feature 3, i.e. they are features of the same entity.

$$C_{A_1}(t) = \begin{cases} VFA : \begin{bmatrix} feature\_1 : & <\{w_1\},\{\}> \\ feature\_2 : & <\{w_2\},\{\}> \end{bmatrix} \\ V4 : ... \\ OA : \begin{bmatrix} feature\_3 : & <\{w_2\},\{\}> \\ feature\_4 : & <\{w_3\},\{\}> \end{bmatrix} \\ ... \end{cases} \tag{4.3}$$

Binding of entities in working memory for a sufficient amount of time will lead to synaptic changes, i.e. binding in long term memory. This is represented in the second set of labels. In equation 4.4 the activation in working memory has ceased, but has resulted in a long term binding of features 2 and 3.

$$C_{A_1}(t+5) = \begin{cases} VFA: \begin{bmatrix} feature\_1: & <\{\},\{\}> \\ feature\_2: & <\{\},\{l_1\}> \end{bmatrix} \\ V4: ... \\ OA: \begin{bmatrix} feature\_3: & <\{\},\{l_1\}> \\ feature\_4: & <\{\},\{\}> \end{bmatrix} \\ ... \end{cases} \tag{4.4}$$

But not only can working memory binding lead to long term binding, also long term bindings facilitate joint activation whenever one part of the assembly gets triggered.

$$C_{A_1}(t_x) = \begin{cases} VFA: \begin{bmatrix} feature\_1: & <\{\},\{\}> \\ feature\_2: & <\{w_1\},\{l_1\}> \end{bmatrix} \\ V4: ... \\ OA: \begin{bmatrix} feature\_3: & <\{\},\{l_1\}> \\ feature\_4: & <\{\},\{\}> \end{bmatrix} \\ ... \end{cases} \tag{4.5}$$

Because of the strong synaptic connection between feature 2 and feature 3, feature 2 activates feature 3 and their activation is synchronized.

$$C_{A_1}(t_x+1) = \begin{cases} VFA: \begin{bmatrix} feature\_1: & <\{\},\{\}> \\ feature\_2: & <\{w_1\},\{l_1\}> \end{bmatrix} \\ V4: ... \\ OA: \begin{bmatrix} feature\_3: & <\{w_1\},\{l_1\}> \\ feature\_4: & <\{\},\{\}> \end{bmatrix} \\ ... \end{cases} \tag{4.6}$$

In general it is not necessarily the case that activation of one cell cluster automatically triggers activation of the other clusters that are long term bound to it (e.g. if the connected cluster are not excited by other sources). On the other hand, clusters that are weakly connected can synchronize when they do receive strong input from other sources. However, in my simplified framework I do not deal with those cases.

The picture that emerges from these neuroscientific facts and assumptions based on them is one of fairly directly implemented basic concepts (or features) and of complex concepts that are built in some way from simpler concepts as components. This conception is consonant with a well-established perspective within cognitive psychology, according to which many, most or perhaps even all concepts have a *prototypicality structure*, with the concept's prototype consisting of a weighted set of features. The assumption of prototypicality structure has been used in explaining a range of observations about the way in which concepts are used and about their acquisition (Rosch 1973, Clark 1973, Lakoff 1987).

This discriptive formalism introduced in this section is, of course, a broad simplification of matters. However, it captures some essential properties of how concepts are represented in the brain, such as the distinction and the relation between long term and working memory binding. This distinction is - as far as I know - not made in any theories within

theoretical linguistics or psycholinguistics. But especially in psycholinguistics it is important to distinguish these levels of representation, since they are necessary to characterize the processes involved in the production and comprehension of informative sentences, as we will see in chapter 7.

4.4 Individual and Categorical Concepts

4.4.1 Categorization

In psychological research about picture naming and categorization, the term *concept* usually means the representation of categorical information in the brain. An individual instance (e.g. a picture of a particular car) is shown to a subject whose task is to come up with the correct category label (*car* in this case).

The implicit assumption seems to be that the category is represented as a concept in the brain, i.e. it is memorized and then retrieved, whereas an individual is something that is perceived at a certain point in time and is not represented afterwards. This is not the way things are. It is true, that perception is always perception of one or several individuals - the object in the world is never a category, but it is certainly the case that individuals can also be memorized and retrieved in the brain. It is, therefore, justified to speak about the representation of both, individuals and categories, as concepts.

A more helpful distinction that is made in standard psychological theories is that between episodic and semantic memory. One could argue that the individual concept is a matter of episodic memory, while the categorical concept is a matter of semantic memory. This is because the encounter of individual entities happens in episodic events - we meet particular individuals on particular occasions. This is important because it implies that the individual concept carries information about time and place of the encounter.

The encounter with several individual entities that have a certain property can trigger the formation of a categorical concept. The formation of such a category can generally be characterized as an abstraction of particular features that distinguish the individual entities of this category (including the abstraction of space and time, i.e. episodic information about the individuals which, in many cases, might be the only recognizable difference between individuals of one category).

It might be justified to say that it is possible to distinguish two major types of theories about the neural mechanisms of categorization in the cerebral cortex: (i) generation of a higher level representation (possibly in a different brain area) in the way this was postulated in self-organizing maps (Kohonen 1982) or in adaptive resonance theory (Grossberg 1980). (ii) feature *reduction* on the same level based on a Hebbian associcative framework (Hebb 1949). The first method associates a category representing neuron to a set of features (the category neuron might at first fire accidentally when the features of the category are active). Categorization is the activation of the category neuron. This neuron can also reactivate the category (but does not have to). According to the second account, repeated encounters with individual objects with the same property lead to a stronger association between the

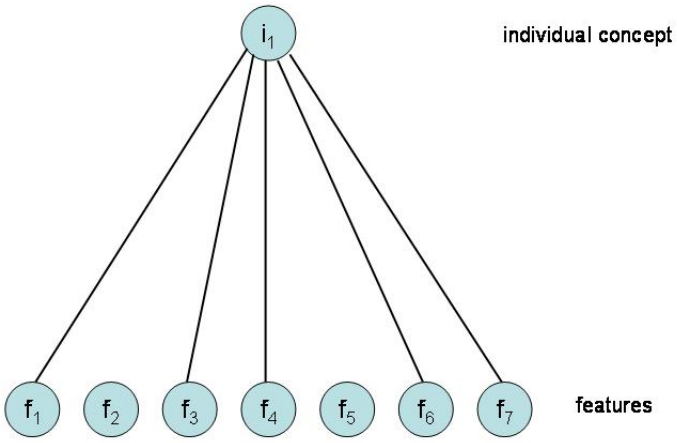

Figure 4.1: The representation of an individual concept with hierarchical feature binding.

shared features. Categorization is then simply the activation of the shared features of the category. Features of the category can reactivate each other, since they have strong associative links. Note that the two types of theories are not mutually exclusive. The brain could and probably does use both mechanisms to categorize and represent categories (Palm 1982).

Note also, that, as discussed in section 2.6.3, these two mechanisms are also the general mechanisms of feature binding in long term memory. This means that the features of an individual concept can also be bound by a abtract neural representation on a higher level. The essential difference between individual and categorical concepts in the brain therefore boils down to the pruning of feature information (see figures 4.1 - 4.4).

In this thesis I will not address the question whether the features are bound on the same level or on a higher level[2]. Far more interesting for the theoretical framework which will be developed here is that the facts and assumptions of current neuroscience suggest that at the neural level there is no fundamental distinction between representations of individuals and representations of categories - for either representation takes the form of a bound set of features, realized by a network of linked feature-representing cells or cell groups.

Suppose that this is so. Then we must face the question: How could such a uniform mode of representation for both individuals and categories be compatible with the view of philosophical and linguistic semantics according to which they are entities of fundamentally different kinds? Or - asking the question from a slightly different angle - how can the brain, assuming that it does represent individuals and properties in this uniform manner, execute

[2]I hope to answer this question in forthcoming empirical work which is done in collaboration with Ton Dijkstra and Herbert Schriefers.

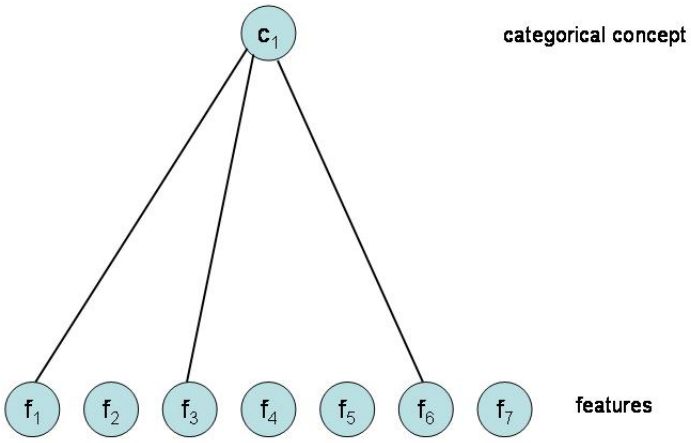

Figure 4.2: The representation of a category concept with hierarchical feature binding.

acts of predication if it is of the essence to the predication operation that the two entities involved, the predicate and the predicable, are as different as semantics claims they are? It is to this question in particular that this section explores some first, tentative answers. I will develop these answers with the help of the theoretical framework skteched above. I begin by reflecting on the acquisition of both individual and categorical concepts in the context of perception.

First consider the case where a person A_1 observes a certain object - a banana, say - and the features of this individual object are allowed to enter long term memory. Equation 4.7 might describe the representation of the individual banana that A_1 encounters. Note that this indidual representation comes with, besides form and color features, also information about the time and place of encounter (I simplify by assuming that the representation involves a 2×2 grid world and 4 possible points in time).

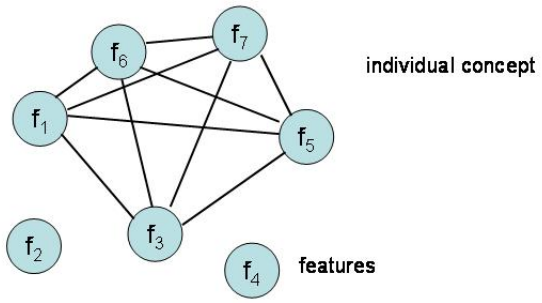

Figure 4.3: The representation of an individual concept with associative feature binding on the same level.

$$C_{A_1}(t) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature: & <\{\},\{l_1\}> \\ other\_form\_feature: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature: & <\{\},\{l_1\}> \\ other\_color\_feature: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature: & <\{\},\{\}> \\ other\_taste\_feature: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \end{array} \right. \tag{4.7}$$

Now A_1 encounters a second individual banana with similar general features but different time and location, leading, in combination with the representation of the previous banana, to a state which is described in equation 4.8,

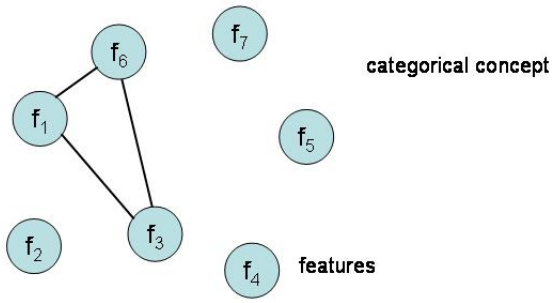

Figure 4.4: The representation of a categorical concept with associative feature binding on the same level.

$$C_{A_1}(t) = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} banana\_form\_feature\_1 : & < \{\}, \{l_1, l_2\} > \\ other\_form\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ V4 : \left[\begin{array}{ll} banana\_color\_feature\_1 : & < \{\}, \{l_1, l_2\} > \\ other\_color\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ TA : \left[\begin{array}{ll} sweet\_feature\_3 : & < \{\}, \{\} > \\ other\_taste\_feature\_4 : & < \{\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[1,1] : & < \{\}, \{l_1\} > \\ position[1,2] : & < \{\}, \{\} > \\ position[2,1] : & < \{\}, \{\} > \\ position[2,2] : & < \{\}, \{l_2\} > \end{array} \right] \\ Time : \left[\begin{array}{ll} t[1] : & < \{\}, \{l_1\} > \\ t[2] : & < \{\}, \{l_2\} > \\ t[3] : & < \{\}, \{\} > \\ t[4] : & < \{\}, \{\} > \end{array} \right] \end{array} \right. \tag{4.8}$$

We now have a situation where two individual objects share two features. This means that these two features are active together more often than the other features, and this allows the formation of a categorical concept consisting of these two features. The new categorical concept is represented as label l_3 in equation 4.9.

$$C_{A_1}(t) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{\},\{l_1,l_2,l_3\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{\},\{l_1,l_2,l_3\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{\},\{\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{l_2\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{l_2\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \end{array} \right. \tag{4.9}$$

Note that such a mechanism for the generation of categorical concepts automatically abstracts from time and location information, as well as from specific properties of particular individuals.

4.4.2 Distinguishing Individuals and Categories

The Bottom of the Lattice

Given that individual concept and categorical concept use the same mechanisms of binding, what is it about a concept that makes it into either a categorical or an individual concept? The answer that comes to mind when looking at the individual concepts l_1 and l_2 and the categorical concept l_3 is that an individual concept is one which is richer in content. The main thing here is that an individual concept, no matter how little information it may carry, always comes with the time and place (or times and places) of encounter. This means that there is a history (at least a little one) for every individual we remember. This episodic information serves to give every individual concept sufficiently specific information to qualify as uniquely identified by the represented features. This uniqueness of features appears to be the essential property of individual concepts.

Look at the list of bound features given in equations 4.10. Since I use this list only to illustrate a theoretical point about individual concepts, the features are not assigned to any brain areas.

$$C_X(t) = \begin{cases} f_1: & <\{\},\{l_1,l_2,l_3,l_4,l_5,l_6,l_7\}> \\ f_2: & <\{\},\{l_1,l_2,l_3,l_4,l_5,l_6,l_7\}> \\ f_3: & <\{\},\{l_2,l_4\}> \\ f_4: & <\{\},\{l_3,l_5,l_6,l_7\}> \\ f_5: & <\{\},\{l_4,l_5,l_6,l_7\}> \\ f_6: & <\{\},\{l_5,l_6,l_7\}> \\ f_7: & <\{\},\{l_6\}> \\ f_8: & <\{\},\{l_7\}> \end{cases} \tag{4.10}$$

To make my point more visible to the reader, I will, for a moment use a different notation which expresses the same relation between features and labels (equation 4.11-4.18).

$$l_1 = \{f_1, f_2\} \tag{4.11}$$
$$l_2 = \{f_1, f_2, f_3\} \tag{4.12}$$
$$l_3 = \{f_1, f_2, f_4\} \tag{4.13}$$
$$l_4 = \{f_1, f_2, f_3, f_5\} \tag{4.14}$$
$$l_5 = \{f_1, f_2, f_4, f_5, f_6\} \tag{4.15}$$
$$l_6 = \{f_1, f_2, f_4, f_5, f_6, f_7\} \tag{4.16}$$
$$l_7 = \{f_1, f_2, f_4, f_5, f_6, f_8\} \tag{4.17}$$

Here, l_1 clearly is a categorical concept, since it shares the features with all other concepts. l_2 and l_3 are also categorical concepts and they are subcategories of l_1. The first individual concept is l_4, since it does not share this feature combination with any other concept. Although it has more features than l_4, l_5 is still a categorical concept, since it shares its features with l_6 and l_7. To summarize, it can be said that in the subsumption lattice of all of a person's concepts at any one time the individual concepts always occupy bottom positions (see also figure 4.5).

Recognition and Categorization of an Individual

Given this structural relation between indviduals and categories, perception of an individual object can lead to either (i) recognition of the individual (which corresponds to the activation of an individual concept) or (ii) categorization of the individual (which corresponds to the activation of a suitable categorical concept).

The categorization of an object is taking place when its perception activates the specific features of the categorical concept. The recognition of an individual can take place as soon as perception has features which are only present in that specific combination in one particular individual concept. The activation of other features that follows from the identification of the individual concept I will call *expansion*.

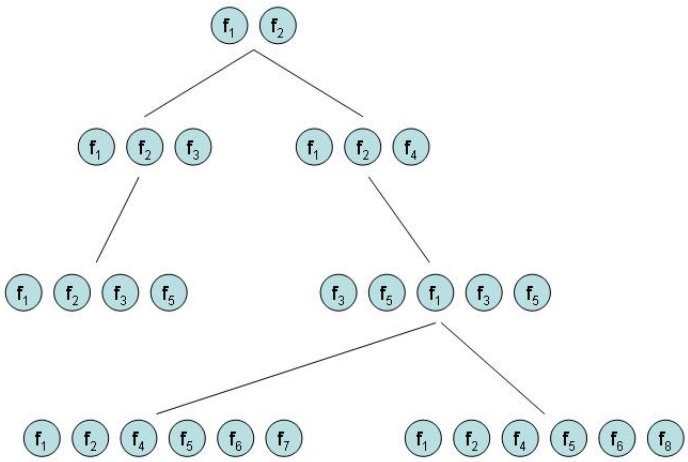

Figure 4.5: Lattice of all of a person's concepts

4.4.3 Types of Predication

Production

Under the assumption that both individual concepts and categorical concepts are representated in the form of a bound set of features and the assumption that the only difference is the uniqueness of bound features (usually achieved in form of information about time and place of encounter), how can the brain execute acts of predication?

I will now try to answer this question. Here it is important to recall the point made in section 4.3.2 about the distinction between four different kinds of predication. Two of those main types of predication are important for language production and the other two for language comprehension. Since there is a topic-predication and a comment-predication in both production and perception, the four types might be called:

- topic predication in production
- comment predication in production
- topic predication in perception
- comment predication in perception

As I have mentioned before, topic and comment predication have to deal with additional complications, such as representations of the mental states of the addressee, and I will deal with those in section 7.5. In this section, I will just sketch the basic mechansims involved and how they can be described in the theoretical framework I have just sketched.

To illustrate the processes consider a situation in which a person A_1 observes a certain object - a banana, say - and then becomes acquainted with some further feature of it. For instance, we may assume that after having first become aware of it, A_1 takes a bite of the banana and finds it to be (deliciously) sweet. At this point A_1 has an individual representation of this particular banana (bound by l_1 in equation 4.18). A_1 also has two categorical concepts, one for *banana* (l_2) and one for *sweet* (l_3).

$$C_{A_1}(t) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{\},\{l_1,l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \end{array} \right. \tag{4.18}$$

The process of topic-predication in production involves the finding of a lexicalized concept which can be used by the addressee to identify the individual about which the speaker wants to give information. This process of *denoting* by the speaker as well as the process of comment-predication in production involves the representation of the addressee's state of knowledge. Both topic and comment predicate are found by contrasting the desired mental state of the addressee with the current mental state. In chapter 7 the theoretical framework will be extended to allow represention of the mental state of the addressee. Therefore, topic and commment-predication will be discussed in detail there. However, for the comprehension side the description is already possible.

Comprehension

Topic-predication in comprehension involves the identification of an individual, i.e. the activation by an addressee A_2 of an individual concept described by the topic predicate. Before understanding the utterance A_2 has to have a representation of the categorical concepts bound by l_2 and l_3 and an individual concept bound by l_1 (as in equation 4.19). Note that in contrast to A_1's representation of the indivudual banana, A_2's representation does not include the feature *sweet*. This is because A_2 does not know (yet) that this particular banana is sweet.

$$C_{A_2}(t_1) = \begin{cases} VFA: \begin{bmatrix} banana\_form\_feature\_1: & <\{\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{bmatrix} \\ V4: \begin{bmatrix} banana\_color\_feature\_1: & <\{\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{bmatrix} \\ TA: \begin{bmatrix} sweet\_feature\_3: & <\{\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{bmatrix} \\ Location: \begin{bmatrix} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{bmatrix} \\ Time: \begin{bmatrix} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{bmatrix} \end{cases} \tag{4.19}$$

To understand and process a sentence such as (1), A_2 has (i) to understand the topic-predicate as identifying the indvidual that A_1 is talking about, and (ii) to understand the comment-predicate as attributing a property to that individual.

(1) The banana is sweet.

By understanding the topic predicate *banana*, A_2 activates the representation of his categorical concept l_2 (equation 4.20).

$$C_{A_2}(t_2) = \begin{cases} VFA: \begin{bmatrix} banana\_form\_feature\_1: & <\{w_2\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{bmatrix} \\ V4: \begin{bmatrix} banana\_color\_feature\_1: & <\{w_2\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{bmatrix} \\ TA: \begin{bmatrix} sweet\_feature\_3: & <\{\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{bmatrix} \\ Location: \begin{bmatrix} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{bmatrix} \\ Time: \begin{bmatrix} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{bmatrix} \end{cases} \tag{4.20}$$

This triggers the activation of the individual concept l_1 (equation 4.21).

$$C_{A_2}(t_3) = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} banana\_form\_feature\_1 : & < \{w_1\}, \{l_1, l_2\} > \\ other\_form\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ V4 : \left[\begin{array}{ll} banana\_color\_feature\_1 : & < \{w_1\}, \{l_1, l_2\} > \\ other\_color\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ TA : \left[\begin{array}{ll} sweet\_feature\_3 : & < \{\}, \{l_3\} > \\ other\_taste\_feature\_4 : & < \{\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[1,1] : & < \{w_1\}, \{l_1\} > \\ position[1,2] : & < \{\}, \{\} > \\ position[2,1] : & < \{\}, \{\} > \\ position[2,2] : & < \{\}, \{\} > \end{array} \right] \\ Time : \left[\begin{array}{ll} t[1] : & < \{w_1\}, \{l_1\} > \\ t[2] : & < \{\}, \{\} > \\ t[3] : & < \{\}, \{\} > \\ t[4] : & < \{\}, \{\} > \end{array} \right] \end{array} \right. \tag{4.21}$$

The preception of the word *sweet* (the comment-predicate) triggers the activation of the categorical concept of *sweet* (equation 4.22).

$$C_{A_2}(t_4) = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} banana\_form\_feature\_1 : & < \{w_1\}, \{l_1, l_2\} > \\ other\_form\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ V4 : \left[\begin{array}{ll} banana\_color\_feature\_1 : & < \{w_1\}, \{l_1, l_2\} > \\ other\_color\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ TA : \left[\begin{array}{ll} sweet\_feature\_3 : & < \{w_2\}, \{l_3\} > \\ other\_taste\_feature\_4 : & < \{\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[1,1] : & < \{w_1\}, \{l_1\} > \\ position[1,2] : & < \{\}, \{\} > \\ position[2,1] : & < \{\}, \{\} > \\ position[2,2] : & < \{\}, \{\} > \end{array} \right] \\ Time : \left[\begin{array}{ll} t[1] : & < \{w_1\}, \{l_1\} > \\ t[2] : & < \{\}, \{\} > \\ t[3] : & < \{\}, \{\} > \\ t[4] : & < \{\}, \{\} > \end{array} \right] \end{array} \right. \tag{4.22}$$

The attribution of the property *sweet* to the individual banana takes the form of synchronizing the activity of the individual concept with the categorical concept, leading to a uniform representation of the sweet banana (equation 4.23).

$$C_{A_2}(t_5) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{w_1\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{w_1\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{w_1\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \end{array} \right. \tag{4.23}$$

This can lead to a binding in long term memory of the feature *sweet* to the individual concept in long term memory (equation 4.29).

$$C_{A_2}(t_6) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{w_1\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{w_1\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{w_1\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \end{array} \right. \tag{4.24}$$

4.4.4 Meager Individual Concepts

There seems to be an immediate objection to the suggestion that in the subsumption lattice of all of a person's concepts at any one time the individual concepts always occupy bottom positions. It appears that many of the individual concepts we actually entertain are quite poor in content. If you tell me: "I saw a man in the street yesterday." and you get called away to the phone before you can say anything else, the effect your utterance will presumably have on me is that I have set up an individual concept for the man you

mentioned, but there is very little I know about the individual which this individual concept is supposed to represent. It is only natural for me to assume that there are many men you saw in the street yesterday. So the property of being a man seen by you in the street yesterday almost certainly will not pick out the individual uniquely. In fact, suppose I happen to know that yesterday you were in the street at least twice, first in the morning on your way to work and then on your way from work in the evening, and that I consider it a practical certainty that you saw men in the street at both times. Then the concepts "man seen by you in the street yesterday on the way to work" and "man seen by you yesterday on the way from work", while each qualifying intuitively as a (non-empty) categorical concept, are both subsumed by the concept I have set up in response to your utterance.

Or so it might seem. But on a closer look the semblance disappears. My individual concept for the man you mentioned before you were called away isn't a concept for just any man you saw in the street yesterday, it is my concept for the person that *you were referring to when you said those words to me*. This is information that won't tell me much about the intrinsic properties of the man in question, but it may nevertheless uniquely identify him, even if it does so in some indirect way which won't help me to find out who he is (unless I rely on you as a source of further information).

I believe that it is this kind of *anchoring* information that in such cases distinguishes individual concepts from categorical ones and that places them necessarily at the bottom of any subsumption hierarchy irrespective of how much (or little) intrinsic information the concept may contain about the thing it represents. Usually such anchoring information takes the form of an explicit descriptive condition on the represented object (i.e. of a combination of categorical concepts, in the terminology used hitherto) of which it is (practically) certain that only one thing can satisfy it; an example are conditions to the effect of the thing having been located in a given place at a given time. But there are also cases where the anchoring condition is not explicit in the manner of this example. Thus, in the case just discussed, where my individual concept is based on the information you have given me in this one sentence before you got called away, the condition presumes there to exist *some* causal relation between me and a particular man, a relation which has been established via (i) your earlier perception of that man and (ii) your referring to that man when speaking to me. In this case the anchoring condition confers uniqueness upon the individual concept by other than straightforwardly descriptive means. There are even cases where this anchoring information cannot really be construed as pertaining to the represented individual in any straightforward sense at all. This happens for instance when an individual is introduced at a particular point in the course of a piece of fiction. The reader will at that point introduce, as part of his representation of the content of the story, an individual concept. And the concept will qualify as an individual concept insofar as one of the concepts it includes will be that the individual was introduced at such and such a point of the text, and/or with such and such words. Of course this information cannot be construed as a property of the (fictional) character described: it is not like his being characterized as a handsome prince, or a toad, or as having lived for most of his life in Baker Street. Nevertheless, as a unique identification of the fictional character the information serves as well as, say, the concept of having been mentioned by someone on a

given occasion serves to uniquely identify an individual in the real world.

Individials and Prototyes

In the way in which this theory about individuals and categories in the brain has been stated so far, is not consistent with prototype theory (Rosch 1973). My theory so far states that an individual concept l_1 can be categorized as l_2 because the features of l_2 are *all* present in l_1. Prototype theory states that an individual can still belong to a certain category (represented by a prototype concept l_4) even if the individual (e.g. l_3) does not include all features of the prototype (e.g. l_4).

$$l_1 = \{f_1, f_2, f_3, f_4, f_5, f_6, f_7\} \tag{4.25}$$
$$l_2 = \{f_1, f_2, f_3\} \tag{4.26}$$
$$l_3 = \{f_1, f_2, f_3, f_5, f_6, f_7\} \tag{4.27}$$
$$l_4 = \{f_1, f_2, f_3, f_4\} \tag{4.28}$$

Another related problem for my theoretical framework is possible change of individuals. How can an individual be recognized when it changed location or say its color (e.g. in the case of bananas).

One way that comes to mind of dealing with this problem is to use a soft mapping algorithm, which allows to categorize l_3 as belonging to l_4 (or to recognize an individual even if not all the features match those represented by the corresponding concept). However, such a soft mapping cannot be the solution as it stands, given the way in which individual concepts have been defined so far. If l_1 - l_4 would be represented in the brain of an agent, l_4 would be considered to be an individual like l_1 and l_3 and the distinguishing criteria of what an individual is would be lost.

To solve this problem, I will take the importance of episodic information one step further. In the last section episodic information was used to distiguish individual from categorical concepts by adding enough features to ensure that indvidual concepts will always be at the bottom of the conceptual lattice. As I have shown above, being at the bottom of the subsumption hierarchy is neither a sufficient nor a necessary condition for being an individual concept. However, it is possible to distinguish individual from categorical concepts simply by testing the presence of certain kinds of episodic information. In particular, if the concept carries information about *time and place of encounter* then the concept is the representation of an individual. This still allows me to hold on to my original claim that the learning mechanisms and principles of representations are the same for individuals and categories.

4.5 Coding of Events

4.5.1 State of the Art

The approach sketched so far describes how the brain can code information about all types of individual objects (e.g. fruits, tools, animals, humans) and object categories. The framework is sufficient to model knowledge about certain objects, e.g. the knowledge that a certain individual banana is sweet, or that John is married. However, the brain must also be able to represent states or events corresponding to the sentences like (1).

(1) John gives Paul the banana.

All the concepts I introduced so far were a combination of sensory features. In contrast, the concept of *give* also involves motor features as well (Pulvermueller 2001). I will not make any detailed assumptions on how this concept is represented in the brain and just assume that it is a combination of sensory and motor features of a certain degree of abstraction.

The problem I must deal with at the moment has to do with the interaction of the different concepts in the representation of an event.

Suppose that *give* is represented by some motor feature, *John*, *Paul*, and the (individual) *Banana* by a combination of visual features with time and place of encounter (equation 4.29).

$$C_A(t) = \left\{ \begin{array}{l} V : \left[\begin{array}{ll} John\_feature: & <\{\},\{l_1\}> \\ Paul\_feature: & <\{\},\{l_2\}> \\ Banana\_feature: & <\{\},\{l_3\}> \end{array} \right] \\ M : \left[\begin{array}{ll} give\_feature: & <\{\},\{l_4\}> \end{array} \right] \\ Location : \left[\begin{array}{ll} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{l_2\}> \\ position[2,1]: & <\{\},\{l_3\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time : \left[\begin{array}{ll} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{l_2\}> \\ t[3]: & <\{\},\{l_3\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \end{array} \right. \tag{4.29}$$

Given this equation there are two possible ways of activating the required participants in the event: (i) they are all activated in asynchrony or (ii) they are all activated in synchrony. Both possibilities create a problem. In (i) we lose the knowledge that they all code the same event, since there would be no way of distinguishing the involved individuals from other indivuals represented in the scene, such as an additional apple. Another problem is that the event described by sentence (1) would be coded exactly in the same way as the event coded in sentence (2), i.e. this representation cannot distinguish which of the individuals functions as agent, object, or recepient.

(2) Paul gives John the banana.

The same is true for possibility (ii). But for this second possibility there arises an additional problem which is much worse: It appears that all the information coded in the concepts of *give*, *John*, *Paul*, *banana* are conflated into one concept. This means that it can no longer be distinguished which of the active features are part of Paul, which are part of John or the banana, and which features belong to the concept of the action representation. If we use this kind of representation (3) might be conceptually represented exactly in the same way as (4).

(3) John and Paul.

(4) John is Paul.

In the literature, several strategies can be found to overcome this problem. I will briefly sketch the most important ones and their problems. Then I will describe an alternative approach derived from neurophysiological data.

The first approach to avoid the conflation of concepts is to use a hierarchical architecture. As discussed before, the brain might recruit neurons in another (higher level) area to represent a concept whose features are represented in the original area (equation 4.30).

$$C_{A_1}(t) = \left\{ \begin{array}{l} VisualAreas : \left[\begin{array}{ll} JohnVisualFeature^{(01)} : & < \{\}, \{l_1\} > \\ ... & \\ JohnVisualFeature^{(n)} : & < \{\}, \{l_1\} > \end{array} \right] \\ HigherConceptArea : \left[\begin{array}{ll} JohnConceptualNode : & < \{\}, \{l_1\} > \end{array} \right] \end{array} \right. \tag{4.30}$$

This higher level representation can then be used to code the complex event. If only the higher level representation is activated, there is no danger of a conflation of features (equation 4.31).

$$C_{A_1}t = \left[\begin{array}{ll} JohnConceptualNode : & 1 \\ PaulConceptualNode : & 1 \\ GiveConceptualNode & 1 \\ BananaConceptualNode & 1 \end{array} \right] \tag{4.31}$$

Although such a representation prevents the conflation of the features of the individual concepts, it still does not allow a distinction between the content of the sentences (1) and (2) and since this distinction is essential I will therefore not use this strategy.

Another attempt to solve this problem is described, and shown to be problematic by (Chang 2002). It is based on an assumption which has been around in theoretical linguistics for many years - the idea is that actions are coded in some form of predicate - argument structures. For example in equation 4.32, the first column vector codes the action type, the second codes the agent, and further vectors can code all sorts of argument types (similar to the approach of)

$$C_{A_1}t = \begin{bmatrix} action & agent & object & receiver & \\ 1 & 0 & 0 & 0 & give \\ 0 & 1 & 0 & 0 & John \\ 0 & 0 & 0 & 1 & Paul \\ 0 & 0 & 1 & 0 & banana \\ 0 & 0 & 0 & 0 & ... \\ 0 & 0 & 0 & 0 & ... \\ 0 & 0 & 0 & 0 & ... \\ 0 & 0 & 0 & 0 & ... \end{bmatrix} \tag{4.32}$$

In this approach events can be distinguished that correspond to (1) and (2), repectively. Also, there is no danger of feature conflation since every column vector codes different objects, so that a simultanous activation of features is not to be interpreted as a single object with these active features. However, the fact that individuals can be in several of the argument positions poses another problem: They have to be coded at every position at which they occur. For example it has to be possible to represent John, who can be either agent or receiver in the vector coding the agent and, again, in the vecor coding the receiver. This is not only an immense waste of coding space (and surely not the way the brain organizes its conceptual representations), but also it poses the problem that during concept aquisition, every concept has either to be learned several times or dublicated in some way (see Chang (2002) for a more extensive discussion and demonstration of these problems). This is sufficient reason not to adopt this approach.

Another way to solve this problem based on temporal coding was proposed by Shastri and Ajjanagadde (1993). The representation they used is based in temporal synchrony. Corresponding to each verb, there are neurons for the thematic roles of the verb (e.g. for the verb *give*, there would a neuron for the agent, another one for the receiver, and one for the object that changes itss carrier. To represent the event, these role-coding neurons synchronized their firing with the conceptual representation of the individuals that fill the roles. Adopting this ideas to my theoretical framework, sentence (1) would be represented as in equation 4.33.

$$C_{A_1}(t) = \begin{cases} give\_verb\_features: & <\{w_1\},\{\}> \\ give\_agent\_representation: & <\{w_2\},\{\}> \\ give\_patient\_representation: & <\{w_3\},\{\}> \\ give\_object\_representation: & <\{w_4\},\{\}> \\ John\_features: & <\{w_2\},\{\}> \\ Paul\_features: & <\{w_3\},\{\}> \\ banana\_features: & <\{w_4\},\{\}> \\ ... & \end{cases} \tag{4.33}$$

This approach prevents conflation of the concepts, allows a distiction between the coding of (1) and (2) and there is no need to duplicate any conceptual representation. While Shastri and Ajjanagadde (1993) postulate the 4 events of synchronized activity to happen

within 40ms (because of synchronization in the γ-band), there is no empirical evidence supporting the idea that the brain is capable of such rapid changes in the activation and synchronization patterns. Simulation studies suggest that such changes need more time (Knoblauch and Palm 2002a,b). However, the general idea of representing an event as a number of concepts that are *not* synchronized with each other is not implausible and the representational structures I will use below follow a similar idea.

4.5.2 Extending the Formalism

The solution for the problems sketched above is inspired by electrophysiological studies, but the theory is still in a very speculative stage and direct evidence still needs to be obtained. Nevertheless, I would like to briefly sketch the general ideas.

The motor system is currently one of the most exhaustively studied brain systems. Area F5 codes specific actions (e.g. *grasping*, *holding*, *tearing*), rather than the single movements that are need to perform them (Rizzolatti et al. 1988). Since the performance of an action can be regarded as a property of an individual, there appears to be no good reason not to bind the action to the individual performing the action. Within my formal framework a sentence such as (5) could therefore be represented as in equation 4.34.

(5) John walks.

$$C_{A_1}(t) = \left\{ \begin{array}{l} VisualAreas : \left[\begin{array}{ll} JohnVisualFeature\_1 & < \{w_1\}, \{l_1\} > \\ ... & \\ JohnVisualFeature\_n & < \{w_1\}, \{l_1\} > \end{array} \right] \\ MotorArea(F5?) : \left[\begin{array}{ll} walking\_feature & < \{w_1\}, \{l_2\} > \\ other\_actions & < \{\}, \{l_3\} > \\ ... & \end{array} \right] \end{array} \right. \tag{4.34}$$

For sentences (1) and (2) this approach has the following implications. When John gives Paul the banana, Paul has to receive it, i.e. he has to open his hand and take the banana. This can be coded in the same way as John's action of giving the banana. Equation 4.35 gives the representation of John giving something to Paul.

$$C_{A_1}(t) = \left\{ \begin{array}{l} VisualAreas : \left[\begin{array}{ll} JohnVisualFeature\_1 & < \{w_1\}, \{l_1\} > \\ ... & \\ JohnVisualFeature\_n & < \{w_1\}, \{l_1\} > \\ Paul\_Visual\_feature\_1 & < \{w_2\}, \{l_4\} > \\ ... & \\ Paul\_visual\_feature\_n & < \{w_2\}, \{l_4\} > \end{array} \right] \\ MotorArea(F5?) : \left[\begin{array}{ll} give\_feature & < \{w_1\}, \{l_2\} > \\ take\_feature & < \{w_2\}, \{l_5\} > \\ other\_actions & < \{\}, \{l_3\} > \\ ... & \end{array} \right] \end{array} \right. \tag{4.35}$$

The remaining problem is the banana. Just to have a banana in the scene (i.e. an active representation of a banana) is not sufficient, since this banana can then not be distinguished from other bananas or other objects that are also present, perceived and represented. Also to add a *being given*-feature to the banana would not help, since utterances such as (6) can involve several objects on which the same action is performed.

(6) John gives the banana to Paul and George gives the apple to Ringo.

The solution is based on the coding of object positions.

$$C_{A_1}(t) = \left\{ \begin{array}{l} VisualAreas: \left[\begin{array}{ll} John\_visualFeature\_1 & <\{w_1\},\{l_1\}> \\ ... & \\ John\_visualFeature\_n & <\{w_1\},\{l_1\}> \\ Paul\_visual\_feature\_1 & <\{w_2\},\{l_4\}> \\ ... & \\ Paul\_visual\_feature\_n & <\{w_2\},\{l_4\}> \\ banana\_visual\_feature\_1 & <\{w_3\},\{l_6\}> \\ ... & \\ banana\_visual\_feature\_n & <\{w_3\},\{l_6\}> \end{array} \right] \\ MotorArea(F5?): \left[\begin{array}{ll} give\_feature & <\{w_1\},\{l_2\}> \\ take\_feature & <\{w_2\},\{l_5\}> \\ other\_actions & <\{\},\{l_3\}> \\ ... & \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{\},\{w_3,w_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{w_2\}> \end{array} \right] \end{array} \right. \tag{4.36}$$

Of course it is a broad simplification just to postulate a spatial overlap of *John* and the *banana*. A first improvement would be to represent an overlap between coding of John's hand position and the banana, but to do this one would need to take into account how the brain codes the position of limbs. Then, as a next step more abstract representations of spatial relations would be necessary, but to develop such a detailed theory about the representation of spatial relations is beyond the scope of this thesis. Instead, I would like to close this section by briefly discussing the representation of events corresponding to sentences such as (7) or (8).

(7) John throws the banana in the direction of Paul.

(8) John throws the banana over the fence.

These sentences describe events which do not involve an action of *Paul* or *the fence*, so they cannot be represented analogously to equation 4.36. However, it is known that

there are neurons in the brain coding a desired trajectory of a movement (Johnson and Ferreina 1996). Therefore, an activation of a neuron coding an appropriate trajectory and a position coding similar to the one used for *John* in the examples above can be used to represent the events described in sentences (7) and (8).

Of course these are just a few examples of event representations and I am aware of the broad simplications I have made and the aspects I have left out. But for the time being I need to restrict myself to the representation of these first ideas.

4.6 Lexicalization

So far, I have not said anything about how the word form representations explained in section 4.2 connect to the concept representations discussed in sections 4.3, 4.4, and 4.5. As indicated above, we assume that both word form representation and concepts are represented in the cortex. Therefore, following the theory outlined in chapter 2, word forms and concepts are connected with unsupervised associative learning. If a concept is connected to a word form representation I will call it a *lexicalized* concept. If a word form representation has a lexicalization link to a conceptual representation, the word form is rendered *meaningful* in that it enables the agent to associate the word form with certain features.

Lexicalization is a difficult process, since it is not easy for the language acquiring child to learn which concepts should be connected to which word forms (see Bloom (2000) for a discussion of the relevant problems and an informal account of a solution). I will say a great deal more about how this learning can be explained within the formal theory developed in this thesis. However, in the following chapters 3, 5, and 6 I will introduce a very different line of thought, which emphasizes the nature of language as a device to accomplish certain goals and where use produces certain effects on the world. The two theoretical strands will be pulled together in chapter 7. In that chapter I will also explain how concepts can be lexicalized through interaction of the language acquiring child with the adult.

Chapter 5

Expressing Desires

5.1 Simulating Goal-Directed Utterance Selection

In this chapter I will describe a possible neural architecture motivated by the biological considerations I have presented so far and investigate its possibilities with computer simulations. The architecture I will propose is capable of learning to use language by selecting utterances suitable for pursuing a specific goal. I have adapted the concept of the internal model (from control theory) to the problem of language use, and use it, in combination with reinforcement learning, to model utterance selection. I also use internal models to approach the problem of intention understanding in language. With simulation experiments in a multi-agent environment I show that our architecture is capable of deciding when to speak (instead of performing non-verbal actions), whom to address and what to say, as well as to understand the intention of another speaker. Further, I discuss data from functional imaging and electrophysiological experiments that point towards a possible relation of the components of our architecture to real brain structures. According to this data, the cerebellum and the basal ganglia appear to be among the major structures involved. This could shed new light on the role these structures play in higher linguistic tasks. Finally, I propose functional imaging experiments which could confirm our hypotheses about the localization of the major components of our model in the brain.

In recent years, a large number of studies have investigated language production in the brain. These studies have dealt with many types of linguistic skills, such as the transformation of meanings into words (e.g. in picture naming), the construction of correct syntactic or morphological structures and the production of phonetically well-formed speech. In any everyday act of communication these skills are employed with a purpose. They are used to realize a goal which goes beyond the production of an intelligible utterance as such and usually beyond the truthful description of a state of affairs. Hence, at the beginning of every act of communication stands the speaker's goal and - by means of her linguistic skills - this goal is transformed into an utterance. Such a transformation must involve a process which can compute what kind of linguistic construction is the most useful in order to pursue this goal. So far, however, linguistic skills have been investigated in abstraction from

the non-linguistic goals of language production. Therefore little is known about the neural basis of this transformation process. Apart from the selection of the appropriate linguistic construction, such a transformation also needs to include the decisions which goals can be pursued by utterances in the first place (instead of by non-verbal actions or not at all) and, in case the selected action is verbal, who it should be addressed to. Furthermore, since the work of Grice (1957), it is a widely accepted view that a speaker's utterance achieves its effects by getting the addressee to recognize the so called *communicative intentions* of the speaker. The communicative intentions are defined as those (and only those) intentions which the speaker wants the recipient to recognize as the goals of the utterance (Levelt 1989). In fact, if these goals are not recognized by the addressee the communicative act cannot be regarded as successful. So, not only does the speaker's goal stand at the beginning of the communication process, its recognition by the addressee is also the final result of the comprehension process. The neural basis of understanding the communicative goal of another speaker is also still very poorly understood. It is not unlikely that the processes of selecting utterances in accordance with goals and the process of understanding the goals of another speaker share some basic neural representations, brain mechanisms and even some anatomical brain structures.

In this chapter, I present a neural architecture using *value functions* and *internal models* for goal-directed and context-dependent selection of utterances. I also consider how an agent uses this architecture (internal models in particular) to understand the communicative intention of another speaker. A value-function is a function which assigns values to states of the world. The value reflects how desirable a specific state of the world is. The internal model, on the other hand, predicts the outcome of actions in certain contexts. An architecture which combines the two components can predict the outcome of an action and is able to assign a value to that outcome. So far, value functions and internal models have been investigated as the basis of goal-directed selection of motor actions (including speech motor control) in the brain (Doya 1999, 2000). I adapt this usage of forward models and value functions to language, because (despite the many differences verbal and non-verbal actions might have in terms of their complexity and in terms of their processing in the brain) both types of actions need to be selected in dependence of a context and with respect to a goal (Austin 1961) and, therefore, might also have similarities in terms of mechanisms and brain structures involved. Such similarities have also been proposed to exist between the processes of understanding people's intentions in using verbal action and in using non-verbal actions (Wolpert et al. 2003). This proposal is supported by the finding that the understanding of the intention of non-verbal actions of other persons precedes the understanding of the verbal actions of other persons ontogenetically (Tomasello 2003). There is also evidence supporting the view that understanding another's intentions (e.g. in gestures) also precedes understanding another's utterances phylogenetically (Rizzolatti and Arbib 1998, Arbib 2000).

There are a number of recent neurophysiological and imaging studies on the representation of value functions and internal models in the brain (Breiter et al. 2001, Seymour et al. 2004, Haruno et al. 2004, Tanaka et al. 2004, Watanabe et al. 2003, Kawagoe et al. 2004, Imamizu et al. 2000). I will discuss these studies in relation to our model and suggest

where the value function and the internal model are located in the brain. The evidence points towards an involvement of especially the basal ganglia (Breiter et al. 2001, Seymour et al. 2004, Haruno et al. 2004, Tanaka et al. 2004, Watanabe et al. 2003, Kawagoe et al. 2004), the cerebellum (Ito 1970, Kawato 1999, Doya 1999, Imamizu et al. 2000), the hippocampus and some cortical regions. The model therefore can give a first account of the involvement of the basal ganglia and the cerebellum in higher level linguistic processing and can be used to explain the effects on those processes of lesion in these areas. The model can also be used to interpret related functional imaging studies. It can also be used to design studies to test the proposed relation of the main components of our architecture to real brain structures.

5.2 Theoretical Framework

5.2.1 Value Function and Forward Model

Instead of the term *goal* I use the terms *desire* and *intention*, because I need to distinguish states of the world, which the agents know to be beneficial for themselves (desired states) from states of the world they are actually trying to reach by some action or utterance (intended states). An agent in our theoretical framework has many desires. However, only some of these desires actually become intentions[1]. This happens when the agent tries to fulfill a desire by some suitable action. In my model I represent desires as *valued* states of the world and intentions simply as states of the world which are linked to some verbal or non-verbal action that is made to bring this state about. The values of the desired states are estimations of how good it is for an agent to be in such a state. The value is a positive or negative real number expressing how much the agents desires a particular state. A function mapping every state onto such a value is called a *value - function* (equation 5.1).

$$V^{\pi}(s_t) = E_{\pi}\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}\} \tag{5.1}$$

The value function is defined in terms of future rewards that can be expected, or, to be precise in, terms of the *expected return* (Sutton and Barto 1998). The expected return is the sum of discounted rewards, which can be expected in and after the given state s_t. $V^{\pi}(s_t)$ is the estimation of the value of state s_t at (discrete) time step t under a *policy* π. A policy is a function mapping every state onto an action. The choice of action has a strong influence on the next state s_{t+1} of the world. However, in multi-agent systems (which I use in our experiments), the actions of the other agents also influence the future states of the world. As long as the actions of the other agents are not (yet) deterministic, as is usually the case as long as agents are still learning, the values of the states are also not stable; and

[1] I use desire and intention as technical terms here and do not claim these to be *real* desires and intentions. However, I think, that a certain analogy can be seen.

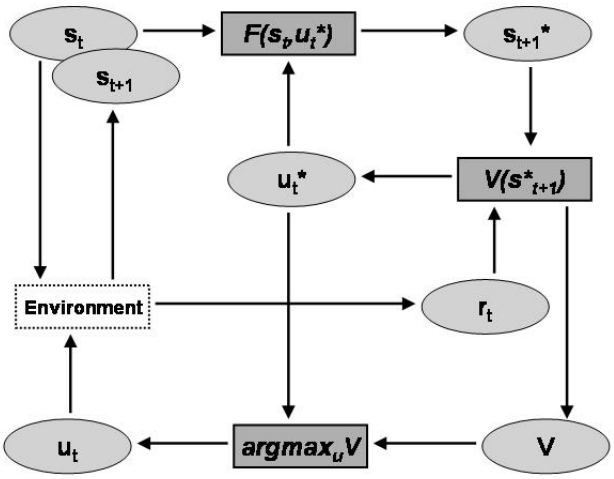

Figure 5.1: A simplified graph of how utterances (and other actions) are selected in our model: The forward model predicts the context-depended utterance effects and the value function determines whether the utterance is selected on the basis of how desirable the effect is.

so is the policy π which depends on the values of states. With the converging of the other agent's actions, the policy π can also converge, and so can the expected return $E_\pi\{...\}$.

The γ-parameter is the discount factor, which determines the importance of the values of future rewards for the evaluation of the current state. The higher the value of γ is, the more importance is given to future rewards. The value function allows to determine the *desired state* of every agent in every state: It is the state with the highest value. However, not every state can be reached from every other state. In fact, only the context state s_t and those few states which can be produced from s_t through a single action in a single times step are accessible. Therefore, the value function only needs to compute the value of those states which can be reached from the current state.

Given the value function, verbal and non-verbal action can be selected with respect to how much the *consequence* of the action is desired, i.e. a speaker uses a certain action because he desires the effects he expects the action to produce in the present context. The type of internal model which predicts the consequences of actions in a certain context is called *forward* model (Jordan and Rumelhart 1992). Within motor control, forward models are used to predict sensory consequences from efference copies of issued motor commands (Kawato 1999). In the model described in this paper, I use forward models for the selection of verbal actions in the following way: the outcome of all possible actions in the present context is predicted with the forward model and then the action which produces the most desired effect is chosen.

In general, the effect of actions and especially of utterances is not deterministic, but

probabilistic. This is an important problem, as it also affects action selection (do we rather select an unlikely effect which is very much desired or a very likely effect which is less desired?). However, in this chapter I limit myself to dealing with the general principles of the acquisition and utilization of internal models in language use and do not deal with probabilistic effects. Also, I do not deal with the fact that the effect of verbal actions on the observable world are indirect. They are achieved by a more direct effect on the (only indirectly observable) mental state of the addressee. To deal with this phenomena in a realistic manner would involve a method of computing the mental state of the addressee from his observable behavior and to relate this state to the effect of the utterance.

The forward model F used in this study does not deal with these problems. F simply predicts a subsequent state s^*_{t+1} based on a current state s_t (context) and a non-verbal or verbal action (utterance) u_t.

$$s^*_{t+1} = F(s_t, u^*_t) \tag{5.2}$$

What subsequent state an utterance can produce is one of the things that has to be learned by a language learning child. This learning involves, among other things, the observation of the utterances of other speakers (including the contexts and consequences of these utterances) and the capability to generalize from their utterances to the effects that the learner may expect from his own utterances. Further, the learner can observe the effects of his own utterance and also make use of that information for the fine tuning of his internal model of language use, e.g. he can experience whether or not a certain utterance in a particular context produces the predicted effect. Irrespective of whether the learner observes or speaks himself, in both cases the experienced context-dependent effects of the utterances can be used to train the internal model of language use.

Given the forward model F, I model language production (utterance selection) with a function that selects the verbal action which produces the most desirable state (equation 5.3).

$$u_t = argmax_u V(F(s_t, u^*_t)) \tag{5.3}$$

This function returns the u from all possible u which, given the context s_t, is mapped by the forward model F into a state s for which the value function V returns the highest value. This equation does not use an action-dependent component of the reward (action cost) since I do not impose any costs for actions themselves).

To understand the communicative intention of a speaker, an addressee can also use the forward model. If the forward model is applied to the utterance with which he was addressed and the context in which the utterance was made, the result is the effect which the utterance has been learned to produce in that context. Since human speakers share an understanding concerning which effects are produced by which utterances in which contexts, the result is likely to be the effect which the speaker intends to accomplish, i.e. his communicative intention.

Learning Algorithms

The value function is implemented as a neural network. To train this network, I used *TD(0) reinforcement learning* (Sutton 1988). In TD-learning, the so-called TD-error gives the distance and direction to the correct prediction and, thus, can be used to change the weights of a neural network. The TD-error δ is computed by subtracting the current state value of state s_t $V(s_t)$ from the sum of the reward r_{t+1} and the value of the next state $V(s_{t+1})$ times the discount factor (equation 5.4). Given δ, the value of the state $V(s_t)$ is changed to $V(s_t) + \alpha\delta$, where α is the rate of change (equation 5.5).

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{5.4}$$

$$V(s_t) \leftarrow V(s_t) + \alpha\delta \tag{5.5}$$

The forward-model, mapping utterances and game-states (contexts) onto game-states (consequences) is implemented by a single-layer perceptron. I use supervised learning to train this forward model. In supervised learning, the output of the network y_k is subtracted from the desired output y_k^* to compute the error e_k (5.6). The weight change Δw_{ik} is then calculated by multiplying e_k with the value of the input neuron x_i and the rate of change α (equation 5.7). Then Δw_{ik} is used to update the weight (equation 5.8).

$$e_k = y_k^* - y_k \tag{5.6}$$

$$\Delta w_{ik} = \alpha e_k x_i \tag{5.7}$$

$$w_{ik} \leftarrow w_{ik} + \Delta w_{ik} \tag{5.8}$$

From this theoretical framework I derive the following two hypotheses: (i) In an environment where only certain accomplishments are rewarded agents equipped with a value function (trained with reinforcement learning), a (rule-based) forward model, and a set of verbal and non-verbal actions can learn to behave in an optimal way, employing language and other actions whenever appropriate. (ii) In such an environment an agent equipped with an optimal value function and a rule-based model for non-verbal actions can learn to use language to achieve his goals by learning to express his desires and to understand the desires of other agents.

5.3 The Acquisition Environment

I test our hypotheses about language acquisition and communication in a simulation of a multi-agent-game. The goal of this game is to obtain food through verbal and non-verbal action. *Food* grows in certain intervals on *trees*. The present simulation uses three trees $T_1 ... T_3$ growing three types of food. Every tree T_i can hold maximally 5 pieces of food. Therefore, a tree can be represented by a 5-dimensional binary vector $\mathbf{t}$. The six possible states of such a tree vector, coding the number of food pieces (0 ... 5), is shown in equation 5.9.

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{5.9}$$

Further, there are three agents $A_1 ... A_3$. Every agent A_i can store 5 pieces of each of the three food types and can, therefore, be represented as a 3×5 matrix $\mathbf{A}$. The coding of the number of food piece of every food type in the agents store is analog to the coding of the trees.

Thus, a state s of the game is represented as a 6-tuple consisting of three 3×5 binary matrices and three 5-dimensional binary vectors (equations 5.10,5.11 and 5.12).

$$s =< A_1, A_2, A_3, T_1, T_2, T_3 > \tag{5.10}$$

$$A_i = \mathbf{A} \tag{5.11}$$

$$T_i = \mathbf{t} \tag{5.12}$$

The agents interact with the world by their perceptions and non-verbal actions and with each other by perception and non-verbal as well as verbal actions. At every point in time t, every agent can perceive the complete state s_t. Time is supposed to advance in discrete jumps, from t_1 to t_2, t_2 to t_3 etc. Each two successive times t_i and t_{i+1} are separated by an action a_{t_i} of one of the agents, so that the state $s_{t_{i+1}}$ at t_{i+1} is the result of the action a_{t_i} producing this effect in the state s_{t_i} at t_i. To learn the effects agents need to store the complete observable game state (including all utterances). To do this, every agent has his own short term memory device capable of storing game states for a constant number m of time steps.

Always after a certain time interval one piece of food gets *digested*, i.e. it disappears. Once the amount of food in the game is below a certain threshold, 3 pieces of food grow simultaneously on one tree. Because of this design, the agents need to act and cannot rest, after they have gained a sufficient amount of food items. Agents never starve to death. But for every time step in which they do not have any food, they get a very negative reward.

Agents can perform one of the following actions:

- harvest a tree (take down all the food)
- give one piece of food to another agent
- ask another agent for a piece of food of a certain type
- *no action*

Agents are allowed to perform *no action*, because, in cases where their current situation cannot be improved by any action, I do not want to force them to deteriorate because they have to do something. Also it appears to be a realistic approach that agents can do nothing unless an action improves their situation.

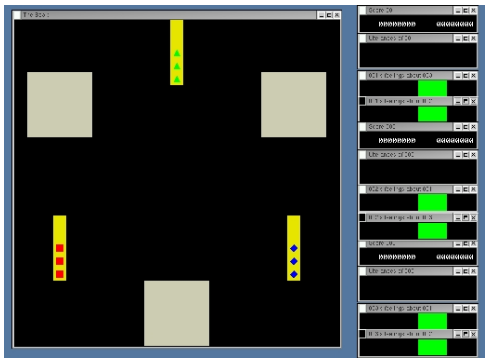

Figure 5.2: This shows the initial game state. The long yellow standing rectangles are the trees, each holding 3 pieces of food. The grey squares are the agents. They have the capacity of storing 5 pieces of each food type. The bar on the right displays scores and utterances. The green bars show, which agents cooperated with which other agents in their last move.

At each transition between two successive times, only one agent can perform an action. This agent can either perform one non-verbal or one verbal action. Generally, the agents take turns. However, when an agent asks another agent for a type of food, the normal order of play is suspended for one time step and it is the turn of the addressee to give the desired object to the speaker.

The goal of the agents in the games is to have one piece of each food type at all times. Therefore, the reward function was designed in the following way: Each agent gets a reward at every time step. If an agent has at least one item of every food type, he gets a reward of $+3$, otherwise he gets -1 for every food type which is missing in his store at that time.

An utterance of an agent is defined by its content (i.e. the word which is used), its speaker, and its addressee. The context of each utterance is the complete state of the game as described above. An agent can only address one of the other agents, never both of them, but the third party can observe every utterance that is made. This is important, because in our setting language is also learned by observing the context-dependent effects of the utterances of other agents. The word used in an utterance has to be one item of the *vocabulary* of the language of the game. In the present study, it consists of the three words *triangle*, *square*, and *diamond*.

To choose their actions or utterances, the agents predict the outcome of all possible actions in the present context with a forward-model. The forward model is rule-based for *no action*, harvesting trees, and donating objects, i.e. the agents do not have to learn the context dependent effects of these actions. With respect to verbal actions, our simulation uses two types of agents. The first type of agent employs a rule-based dialogue system to understand and produce utterances. The second, and more interesting type uses the neural network-based architecture described above.

This architecture uses two types of data for learning. On one hand it learns by observing

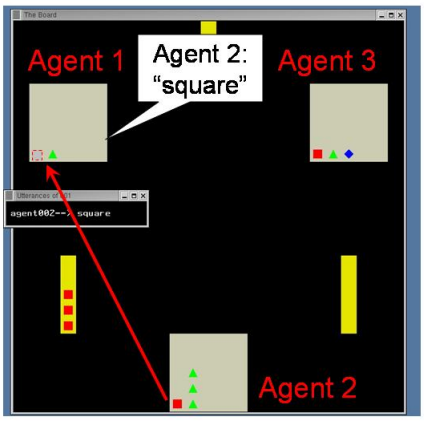

Figure 5.3: This shows an arbitrary state during the early stages of training. The last action of agent 1 was to ask agent 2 for the square. Obviously, this is not the best move. A better move would be to harvest the *square tree*, as with this action, the agent would get 3 squares instead of one.

the context-dependent utterance effects of other speakers. On the other hand, it learns by trial-and-error, i.e. by observing the context-dependent effects of its own utterances.

The neural network-based forward model trained with this data is gradually improving its ability to map context and utterance onto effects. The better the forward model is trained, the better it can be used for language production and language comprehension.

For language production, the predicted effects of the verbal actions are evaluated with the value function (this is also the case for dialogue system-based agents). The verbal or non-verbal action which will bring about the state with the highest value is chosen.

In language comprehension, the language learner, when addressed with an utterance, applies his neural network-based forward model to the utterance and the game state. The network then estimates what effect the utterance usually has in the given context. Using this method the learner computes the intention from the utterances and the context. In other words, the language learner understands the utterance, because he *wonders* what effect, according to his own experience, such an utterance has in the present context. Using the present state of the game and the estimation of the desired state of the speaker, the addressee then uses a rule-based algorithm to compute which action would bring this desired state about.

In the simulations reported in this chapter, I have decided to make the addressee fulfill the comprehended desire of a speaker, i.e. they give the desired object to the speaker even if they need it themselves. Alternatively, I could have made the agents cooperate only if it would be beneficial for themselves and make them develop a more strategic way of playing the game. This, however, would mean that the focus of this study would shift towards an investigation of cooperation in multi-agent systems, similar to the work of Grim and Kokalis (2004) or Mirolli and Parisi (2004). But cooperation is not the issue of the study

described in this chapter. Here I try to model the situation of a child that learns language through interaction with its parents. In such a situation, cooperation can be assumed to be the general attitude that agents have towards each other. The simulations are not intended to investigate the effects of non-verbal cooperation or non-cooperation, but to investigate a general mechanism of utterance selection, as well as the advantage a child gains by learning to communicate verbally with adults, given that these adults are willing to cooperate with the child and with each other.

In chapter 6, I will raise the issue of cooperation and strategic game play to motivate the selection of assertive speech acts, but then this issue is raised to explain the effect of a type of utterance and is, therefore, more relevant to a neural theory of meaning and use.

5.4 Simulations

5.4.1 Value Function

During the training of the value function agents learn which states are desirable. In that training phase they use a rule-based forward model, which computes the context-dependent consequences of all four kinds of actions (including verbal actions). Because of this, the value function enables the agents to decide in which situations it is better to harvest, donate, speak, or simply do nothing.

At the very early stages of the training, agents select *no action* or senseless actions very often (such as donating objects to other players without being asked). Sensible, but suboptimal actions, as described in figure 5.3, occur in the intermediate stage of training. After training, no more suboptimal action can be detected. Agents use language if appropriate, harvest trees whenever this was the best possible action, and almost completely stopped choosing no action (as one would have expected of a successful simulation since in most cases some action or request does in fact improve the state of agent). Figures 5.4 and 5.5 demonstrate the effect of the γ-parameter. The value function of each agent will compute positive values for objects in the store of this agent and negative values for objects in the store of other agents. With $\gamma = 0.9$, the value function computes positive values for more than one object of the agent, although the reward is given only for the first object. This is not the case for $\gamma = 0.1$. When the γ parameter is set this low, then agents only take immidiate rewards into account instead of collecting food items that they can use in the long run.

The TD-error decreased very fast in the beginning (figure 5.6) and optimal performance could already be observed after about a million time steps. In the end, agents performed approximately at the level of a human player or even better.

5.4.2 Language Learning

In the second type of simulation, a neural network based forward model was trained using the previously described learning algorithm. The task for the network was to learn

the context-dependent consequences of utterances. This network represents the linguistic knowledge of the agent. In our simulation, these consequences are assumed to be conventional within a language community. An agent is placed within a community of other agents and has to learn the conventions from them. For this task, the language learner is placed in an environment with two agents who use the rule-based dialogue system to produce and understand utterances. These rule-based agents have no interest in the learner-agent's learning of language. They do not perform any non-verbal actions or utterances with the goal to teach language to the learner. Instead, the learner-agent can learn language by observing the language use of the rule-based agents and by trial and error. The rule-based agents' knowledge about the conventions of their language is hard-coded. They act according to a set of rules translating their desires into utterances and computing the state desired by another speaker from the utterance they are addressed with. Their verbal actions and their reactions to verbal actions of other agents make it possible for the language learner to acquire the verbal repertoire they use. At the beginning of the second type of simulation, the language learner's value function (the neural network which computes values of the states) is already fully trained. With this working value function, this agent knows which states should be achieved. His language capabilities, on the other hand, could best be described as involving the production of *random utterances* with no understanding. Although the learner knows which state should be achieved, he has no idea which of the utterances are likely to bring this state about. This leads to situations where e.g. he needs the square of one agent, but addresses another one and asks him for the triangle. Figure 5.7 illustrates another situation where the language learner asks an agent for an object which the addressed agent does not even have. This is because in the early stages of training the learner has no or no accurate representation of the context condition for successful use of an utterance and also does not know its conventional effect.

The language understanding of the learner faces the same problems. Similar to the previous simulations during which the value function was trained, all agents, including the learner, are forced to cooperate when being asked for a food item. For the language learning agent this means that he has to comply to the request of another agent as well as his linguistic capabilites allow him to. If he does not understand the request, and shows no reaction, or if he shows the wrong reaction, he is not punished. In early stages of training the learning agent usually shows no reaction when being asked for a food item, because he maps the utterance to the wrong state, i.e. not to the state that is desired by speaker. Often the state that the learner identifies as the one desired by the agent who has addressed him is not accessible from the state in which the utterance takes place. In the rare cases where it is accessible the agent does what is needed to bring about the state he has identified as the desired one. While the training progresses, however, reactions are more and more in accordance with the utterances he is addressed with. Also the learner's utterances are more and more attuned to the situation and his needs. After training, language production and comprehension was optimal and no difference between teacher and learner could be detected any longer. As can be seen in figure 5.8, the prediction accuracy increased very fast to a level close to 100 %.

5.4.3 Learning without Observation

An internal model of utterance effects makes it possible to learn these effects also by observing the language use of other agents. To see the actual benefit of observation-based learning I compared (i) learners who do not learn by observation (i.e. they train their forward model only by observing the effects of their own actions) with (ii) learners who learn by observation as well.

As a first measure of successful learning I counted the number of wrong uses of utterances, i.e. utterances which did not match the context conditions (e.g. an utterance in which another agent was asked for a certain object which he did not have). The ratio of wrong utterances for both types of agents is 0.5 in the beginning of the simulation (see figures 5.9 and 5.10). For agents who do not use learning by observation, the ratio stays at this level until 10 million time steps. At 50 Million it is still around 0.1 and it finally reaches 0.0 at 80 million time steps. For agents which do use observation learning, the ration drops very fast in the beginning, being below 0.2 after 10 million time steps. After 50 million time steps the ratio for this type of agents is already below 0.05, reaching 0.0 before 80 million time steps. Although the agents who use observation learning learn a lot faster, both types of agents are free of errors after approximately the same number of time steps.

I also compared the performance (in terms of score) of the two learning types. As can be seen in figure 6.1 the performance of the observation learners increases immediately and very fast, while the agents without observation learning have a slower start and perform worse for about 170 million time steps. After that, there is no significant difference between the two learning types.

5.4.4 Muted Agents

Although I constructed the simulation environment to make it beneficial for the agents to use language, I wanted to test how much agents gain by communication. Therefore, I modified the simulation in such a way that agents could no longer select verbal actions to fulfill their desires. I ran 10 simulations for 10 million time steps with and without language and computed the mean value of reward received by all agents per time step. With language, agents received an average reward of 0.843, without language they received -1.913 (figure 5.12). Using the values of the 10 runs, I performed a paired t-test of the significance of the difference. It revealed a highly significant p-value of 0.0000003928.

5.5 Discussion

5.5.1 Summary of Simulation Results

I have demonstrated that in a game environment where only certain accomplishments are rewarded, agents equipped with a value function (trained with reinforcement learning), a (rule-based) forward model, and a set of verbal and non-verbal actions can learn to behave

in an optimal way, employing language and other actions whenever appropriate. I have also demonstrated that in such an environment an agent equipped with an optimal value function and a rule-based model for non-verbal actions can learn to use language to achieve his goals by expressing his desires and to understand the desires of other agents.

Furthermore, I have compared this model with one that does not learn by observation. The small difference between value function learning with and without observation can be explained by the extreme simplicity of the language used in this study. I predict that the more complex the language used by the agents is, the higher the advantage of observation learning would be.

In general the number of time steps required for the training of both, the value function and the forward model, is very high. For the forward model the high number can be explained by the fact that the model learns the effect of utterances without prior knowledge or subfunctions. Further, the function $F()$ uses the complete state of the world as arguments and result (maps context states and utterances onto the next state). To create a more realistic model it would be necessary to incorporate conceptual knowledge and mechanisms such as shared attention to a particular part of the world. These changes might reduces the learning rate immensely, because instead of mapping complete worlds onto complete worlds, the model would just have to learn a change in propery (location) of the object of significance. However, neither for the forward model nor for the value function was optimization of the learning time the purpose of this study. So I have not investigated how changing the values of various parameters reduces the number of time steps necessary for learning. An actual comparison of the speed of word learning by a model of this type and the learning rate of children would have to pay more attention to these kind of factors. Such a study would be very interesting, but is beyond the scope of this thesis.

5.5.2 Basal Ganglia and the Value Function

The theoretical framework of language use introduced in this paper proposes two major components: a value function to predict the value of states and an internal model to predict which states can be brought about by which utterances. Neural correlates to both components can be found in the literature.

To find brain areas involved in the computations described by the value function, functional imaging experiments need to be considered, which identified regions for *reward prediction*. In experiments where subjects need to act to maximize the reward, they often need to predict the reward (e.g. as the consequence of a certain action). The task of the value function is also to predict a reward - the reward for a certain state. Brain activity correlating to reward prediction was found in the striatum in fMRI studies (Breiter et al. 2001, Seymour et al. 2004, Haruno et al. 2004, Tanaka et al. 2004). Also, reward predictive modulation of striatal neuron firing was shown by Watanabe et al. (2003) and Kawagoe et al. (2004).

5.5.3 Cerebellum and the Forward model

In our architecture, the effects of verbal and non-verbal actions are predicted by the internal model. Based on neurophysiological data already Ito (1970) suggested that internal models predicting the effects of motor commands are represented in the cerebellum. Recent neurophysiological and imaging data supports this theory (Kawato 1999, Imamizu et al. 2000). Also, current neural models of linguistic processing identify the cerebellum as the neural site representing an internal model used in speech motor control (Guenther 2001). To regard the function of the cerebellum within linguistic processing as limited to motor control is a result of the traditional view of the cerebellum. In contrast to this tradition, I suggest that the cerebellum is a main component of the complex mechanisms involved in predicting context-dependent utterance effects that are required for appropriate language use. Several kinds of evidence support this view.

The first kind of evidence, summarized by Doya (1999), are anatomical, physiological and computational data, suggesting that major brain structures are not to be distinguished mainly by their cognitive domain (motor control, language, memory, attention etc.), but by their method of learning (their learning algorithm). For the cerebellum the method of learning indicated by this data is supervised learning. Supervised learning, again, is the most suitable algorithm for the acquisition of internal models (Kawato 1999). If it is the case that the cerebellum is generally the neural structure acquiring and representing internal models, then it is likely that internal models required for higher linguistic functions are also represented in the cerebellum (at least to some important part).

Second, clinical data also points towards an involvement of the cerebellum in the prediction task modeled in this study. Note that the internal model is used by the agent to select the appropriate utterance for a certain goal and a certain context. A patient without such an internal model would still have some basic linguistic knowledge (e.g. would be able to name objects). However, he would not be able to predict language effects and, therefore, would not be able to use language appropriately. This aspect of use, the so-called pragmatic aspects of language, are known to be impaired in patients with Autism or Asperger syndrome. While patients with Autism often have no or strongly impaired language, individuals with Asperger syndrome, a high functioning[2] variant of Autism, have a normal command of language. Yet they fail to *use* this language capacity to engage in interactive communication.

With autistic patients, anatomical abnormalities have been identified in many brain areas. These include the cerebellum (Courchesne et al. 1994) and the hippocampus, but also other areas, such as frontal lobes, parietal lobes, the amygdale, and the brain stem. Further, decreased Purkinje cell density in the cerebellum is a relatively constant observation across post-mortem studies of autistic patients (Williams et al. 1980, Bauman and Kempner 1985, Ritvo et al. 1986, Bauman and Kemper 1994, Bailey et al. 1998). Purkinje cells, on the other hand, have been shown to be important for internal models (Kawato

[2]By convention, if an individual with autism has an IQ in the normal range (or above), they are said to have *high-functioning autism (HFA)*. If an individual meets all of the criteria for HFA except communicative abnormality/history of language delay, they are said to have Asperger syndrome (AS).

1999).

Besides these findings in Autists and Asperger patients, there is another impairment suggesting an important role of the cerebellum in the neural representation of the internal model: *Cerebellar mutism*, as described by Turgut (1998), is a specific disorder in which a complete but transient loss of speech occurs following resection of intrinsic posterior cranial fossa tumors or cerebellar hemorrhages, or upon trauma. Trauma to the cerebellum is the most common organic cause of mutism (Gordon 2001). As cerebellar mutism is usually followed by dysartria, it has often been regarded as an extreme disorder of speech motor control. However, a systematic study of Riva and Giorgi (2000) showed, that, depending on the site of the lesion, cerebellar mutism can also be followed by higher order language disorders comparable to agrammatism, and disorders comparable to those of autism.

The third kind of evidence comes from functional imaging and other studies showing that the cerebellum also plays a major role in the processing of higher linguistic levels (Desmond and Fiez 1998, Papathanassiou et al. 2000, Marien et al. 2001, Noppeney and Price 2002, Xiang et al. 2003, Stowe et al. 2004, Justus 2004). If our architecture is correct, and the cerebellum is one of the main neural sites of the internal model for language use, then it is not unlikely that these cerebellar activations found in those experiments concern the use-related aspects of linguistic competence.

5.5.4 Involvement of Other Brain Areas

On its own the cerebellum appears to be a short term prediction component which is unlikely to be able to handle the long term effects of utterances. However, loops between cortical areas (such as the cortical language areas) and the cerebellum could handle these long term effects. Functional connectivity between frontal areas (Broca's Area) and the cerebellum shown in a functional imaging study (Tamada et al. 1999a) supports this view. Also, an internal model predicting context-dependent utterance effects is likely to involve many other linguistic skills, which are likely to be represented in cortical areas, such such as BA 44/45 and BA 22. Other cortical areas involved in the understanding of communicative intentions might be those associated with mind-reading (Baron-Cohen 2004, Hill and Frith 2003).

5.5.5 Alternative Models of Language Use

In my approach agents are capable of generating their own goals, based only on rewards given in certain states. Unlike any other approach known to the authors, the modeled system has to learn when it is useful to speak, instead of just keeping silent or performing a non-verbal action. This is an important feature of humans and human communication. Humans are not simply reactive systems. I can initiate conversations, ask questions when I need information, or talk just for social reasons. Models of language use which are not goal-directed lack this important capability. In most current dialogue systems, for example, questions trigger the search for some information, which is then made available in some form to the querier, possibly in the form of an utterance phrased in the same language as

the question. In general, such systems cannot understand speaker's intentions, nor can they form appropriate intentions themselves. The capability to handle the goal-directedness of language in production and perception appears to be one of the most essential features of the language faculty in the human brain. Purely reactive systems (as most systems are until now) cannot serve as plausible models of the cognitive architecture of the human communication system.

Goal-directedness can also be modeled without a predictive component (such as an internal model) with simple reinforcement learning. However, it has been argued by Chomsky (1959), that language cannot be learned by simple reinforcement learning[3]. Further, language is learned to a large degree by observation; and a device is needed which can learn context dependent utterance effects independently of the reward these effects produce. In other words, the language faculty must be independent of the knowledge about which states of the world are desirable, or how much. This independence is upheld by our theoretical framework, which - in contrast to simple reinforcement learning - uses the value function and the internal model as two interacting, but separate components.

With a value function as one component, an internal model of the environmental dynamics is necessary for utterance selection. Although I used forward models in the simulations reported in this thesis, *inverse models* might be involved too. Inverse models can calculate necessary feed-forward motor commands from desired trajectory information (Kawato 1999). Adapting this definition to language, inverse models would map desired states and context states directly into utterances[4], i.e. they would calculate the action necessary to accomplish the most desired state in the present context. The problem of inverse models for this task is that not all desired states (probably only very few of them) can be reached by some action from the current context state. States which cannot be reached by an action need to be ruled out computationally. For an environment of a similar complexity as the one in our study (which is still quite simple in comparison to the real world), such computations are very expensive.

The alternative I have chosen in our approach - the combination of value function and forward model - selects utterances by predicting the effect of all possible actions or utterances and selecting the utterance or action which brings about the effect the agent desires the most. Of course, for natural language it is an impossible task to compute the expected effect of all possible utterances. There are infinitely many of them; therefore, such a computation could not be carried out in finite time. If, however, in the brain, the linguistic constructions for which the effects are computed by the forward model are triggered by context cues and properties of the desired state, then the forward model approach seems to be more plausible than the inverse model approach.

Our view is also supported by research in language acquisition. Children appear to learn the use of utterance in terms of relations between context, utterance, and effect (Tomasello 2000, 2003).

[3]Although his points are still quite controversial.

[4]In contrast to the forward model, which is used to predict the outcome of all possible actions in the present context, so that the action which produces the most desired effect can be selected.

5.6 Conclusion

In this article, I have introduced a neural model of language use which is based on the combination of a value function and an internal model. In this approach, the value function determines which state of the world is the most desirable and the internal model computes which verbal or non-verbal action is the best to reach this state. I showed that such an architecture is capable of deciding when to use (a simple form of) language and selecting the appropriate utterance with respect to the goal and the context. Various kinds of evidence support the view that this is the brain's way of using language and currently I do not see a good alternative to this approach. Furthermore, there is evidence indicating that the basal ganglia are the major component in the brain's implementation of the value function and that the cerebellum plays at least an important role in the representation of the internal model.

5.7 Technical Details

5.7.1 Rules

This section gives the details of the rule-based language. Agents using the rule-based language use the rules both for language production and for language comprehension. Note that all agents (including the rule-based speakers) use a neural-network based value function to select actions.

The algorithm for the rule-based language would first test whether the last action was a verbal action. If this is the case, a reaction is required. If no reaction is required, the standard action selection algorithm would be carried out. To minimize computational and coding effort, I have used the rule-based forward model for non-verbal actions to predict the outcome of verbal actions for the rule-based speakers. To do this I let the rule-based forward model for non-verbal actions predict the outcome of taking objects from other agents (this type of prediction is blocked in my neural network-based agents). If the state to which such an action would lead gets the highest value (using the $argmax_u V()$- function), the non-verbal action is transformed by a function *generate_utterance* into a verbal action. This function transforms the parameters of the non-verbal action into parameters of a verbal action approximately as in equations 5.13 - 5.15.

$$addressee \leftarrow object\_source\_location \tag{5.13}$$

$$speaker \leftarrow object\_target\_location \tag{5.14}$$

$$content \leftarrow object\_type \tag{5.15}$$

If a reaction is required, the rules compute the required reaction as in equations 5.16 - 5.18.

$$object\_target\_location \leftarrow addressee \tag{5.16}$$

$$object\_source\_location \leftarrow speaker \tag{5.17}$$
$$object\_type \leftarrow content \tag{5.18}$$

There is a further set of rules to compute the non-verbal reactions of the neural-network based speakers. When they are addressed, they use the neural network-based forward model to compute the desire of the speaker, i.e. they compute the desired state. However, as the transformation of a desire into an action is not a linguistic problem, our simulation uses a set of rules to compute the action which fulfills the desire. This is done by comparing the desired state with the current state and identifying the object with the changed position. After having computed the position of this object in the current state and the position in the desired state, the transformation into non-verbal action parameters is straightforward.

5.7.2 Parameters of the Simulations

The neural network implementation of the value function maps states of the game to real numbers. The input layer consists of one neuron for every binary value of the vectors and the matrices of s. The output of the network is the linear combination of the weighted binary inputs. As exploration mechanism I used a *soft-max* method. I tested the training of the value function for $\gamma = 0.1$, 0.3. 0.5, 0.7, and 0.9. The lower the γ-value the faster was the decrease of the TD-error and the higher the γ-value, the faster was the increase in performance (with respect to score) and the better the performance in the game in total (see also figs. 5.4 and 5.5). Therefore, I chose to use a γ-value of 0.9. The learning rate of the value function α_V was set to 0.001. The softmax exploration algorithm used a method in which random numbers were added to the values of the states. In the beginning of the simulation, these values were selected from between a maximal value v and 0. Until the end of the simulation v was linearly decreasing to 0. As initial value of v I chose 4.

The neural network implementation of the forward model is a single-layer network with two types of input: the first type of input is the complete state s. The input layer, therefore, has one neuron for every binary value of the vectors and the matrices of s. Further, the input layer has a binary vector to code the used one word sentence. The output of the model is, again, the complete state of s, coded by one neuron for every binary value of the vectors (representing the trees) and the matrices (representing the agents). No hidden layers are used. The learning rate of the forward model α_f was set to 0.00005. The threshold θ for the output neurons was 1 in all networks and all simulations.

The simulation parameter regulating the number of food items in the game, were the minimal number of items in the game ($n_o = 9$), the number of items in the game at the very beginning ($n_i = 9$), the number of objects which grow on a tree when the item number is below n_o ($g_o = 3$) and the number of time steps after which one objects of every agent is digested ($d_o = 5$).

Further parameter defining the game were:
number of agents: 3
number of trees: 3
number of food types: 3

number of possible items on a tree: 5
number of possible items of one type an agent can posses: 3

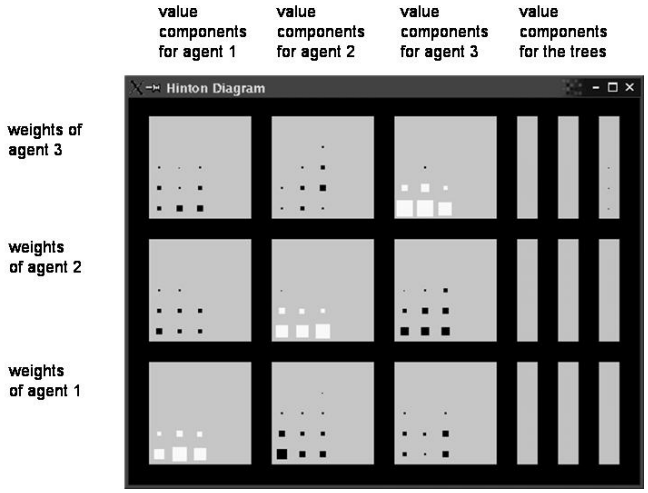

Figure 5.4: This shows the strength of the synaptic weights for the value function after 20 million time steps for $\gamma = 0.9$. Every horizontal row represents the weights of one agent for computing the value function over the game state.

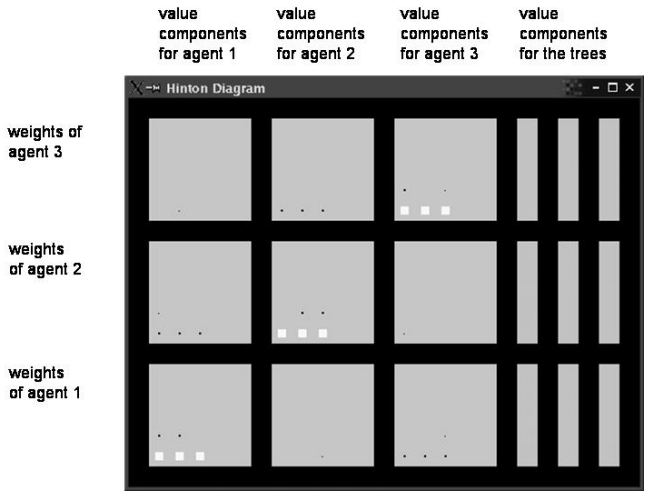

Figure 5.5: This figure shows the strength of the synaptic weights for the value function for $\gamma = 0.1$ (20 million time steps).

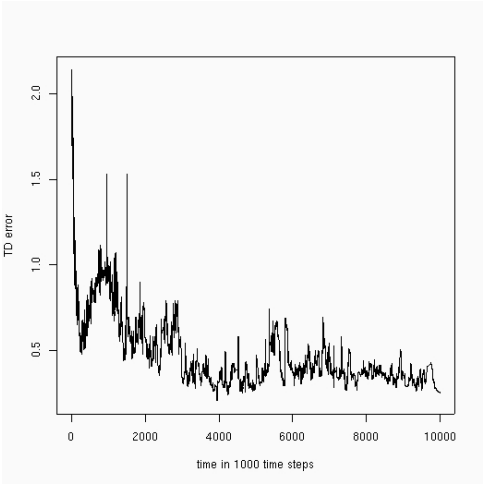

Figure 5.6: This is the development of the TD-error with $\gamma = 0.3$. The figure shows the average error of 5 runs.

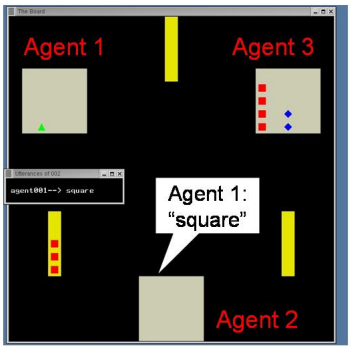

Figure 5.7: This shows an episode during language learning. The language learner (agent 2) asks agent 1 for the square, although agent 1 does not have one. This is an example of the language learner not being able to understand the context conditions and the normal *effects* this kind of utterance has.

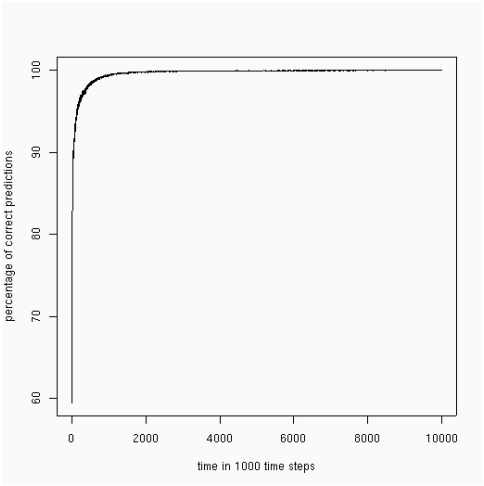

Figure 5.8: The prediction error of language learning changes over time. This graph shows the percentage of correct predictions for five runs (10 million time steps each). The error decreases very fast in the beginning and then slowly reaches 0.

.

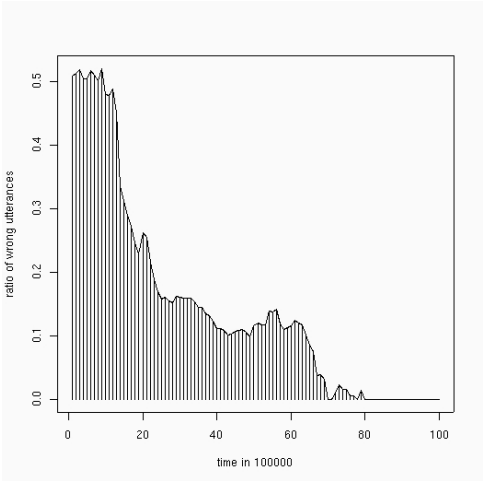

Figure 5.9: This figure shows the average ratio of wrong utterances for five simulations in which only trial-and-error learning and no learning by observation was used.

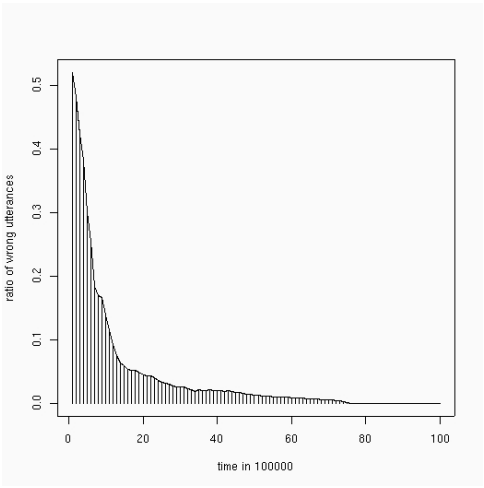

Figure 5.10: This figure shows the ratio of wrong utterances for cases in which learning by observation is used in addition to trial and error learning.

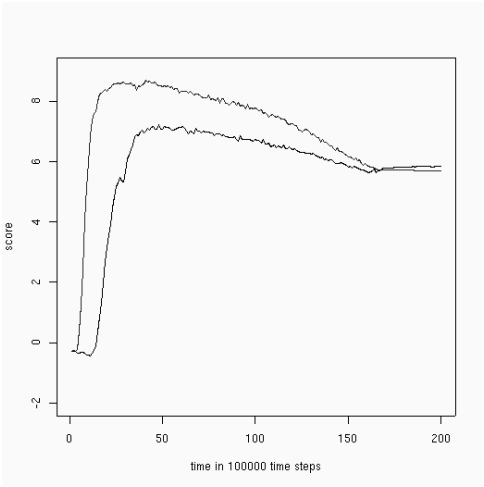

Figure 5.11: This figure shows the performance of a language learner which uses v-function and forward model, but no learning by observation (black line) in comparison to a language learner which uses v-function and forward model, and learns by observation as well (grey line). The lines represent the average development of the score for 5 simulation runs. In these simulations, the reaction of the learner to request for objects is switched off, because, as soon as he understands other agents, he also has to give objects away. This would result in a massive drop of the score.

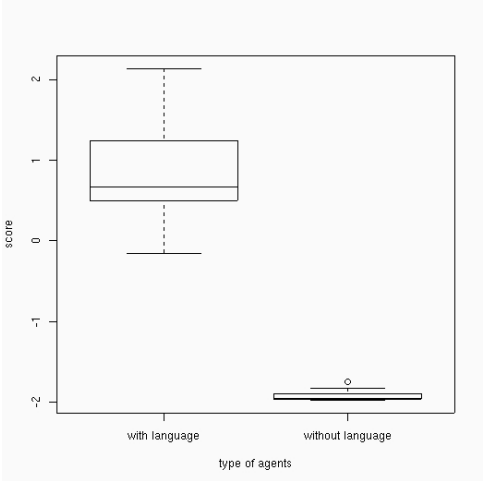

Figure 5.12: This shows the distribution of average scores of 10 runs. The left box shows the values for simulations in which the agents where allowed to communicate. The right box shows the value for simulations during which communication is suppressed.

Chapter 6

Speech Acts

6.1 Increasing the Speech Act Variety

In this thesis, language acquisition and processing in the brain is modeled within the theoretical framework of goal-directed systems. This perspective puts non-linguistic goals - the reasons why we talk in the first place, but also the purpose of each single utterance - into a central position and regards other aspects of language processing as a means to an end. One main strength of this approach is that it allows a natural analysis of how humans are able to generate utterances to accomplish their non-linguistic goals. The acts of pursuing such non-linguistic goals were called perlocutionary acts by Austin (1961). So far, I have only dealt with one sort of perlocutionary act - the act of getting an addressee to give the speaker an object. The corresponding speech act type is a *request*. In this chapter, I will show that my theoretical framework (using internal models and value functions) can also deal with other major types of speech acts: *questions* and *assertions*. The theory can explain how humans select among different types of speech acts depending on the context and their desires. To accomplish this extension of the theory, I will introduce a fundamental distinction which I have ignored so far: the distinction between the states of the world and the *knowledge of the agents about this world*. This distinction is essential for the modeling of meaning processing in the brain, since in the end it is the information about the world that one wants to obtain when asking questions and that one wants to share when making assertions. In other words, the possibility of *not* having information is an important presupposition for a theoretical investigation on these speech acts.

I will use several different game environments in this chapter to illustrate different problems that have to be dealt with while dealing with different speech acts and to describe how they can be solved using internal models and value functions. These different games are all related to the game used in chapter 5, but often more abstract and I will focus only on the particular relevant part of the games.

6.2 Questions

6.2.1 Knowledge

In the model described in chapter 5 a request is triggered if the value function indicates that a state of the game in which the speaking agent would have a certain object has a higher value than the current state in which this agent does not have this object while another agent has it. The intended effect of questions is of a different type: it is not a gain of objects, but a gain of knowledge.

Of course, asking questions to gain knowledge and making statements to provide knowledge only makes sense if there are states of affairs in the environment that are unknown to some of the agents while known to others. The simulations reported in chapter 5 were designed in a way that every relevant property of the environment was directly accessible at every point in time, i.e. the agents always knew everything they needed to know about the state of the environment. In this chapter, I will introduce properties of the environment which are not always directly accessible. To obtain knowledge about these properties agents need to ask questions.

The explanation that agents ask questions because they want to obtain knowledge is not sufficient in a theoretical framework in which all actions are performed in order to increase the total reward - it needs to be explained in what way knowledge increases the reward the agent will receive. The simplest way of doing this would be to directly reward the gain of knowledge. Such an approach is biologically not implausible. It would correspond to an instinct of curiosity. Curiosity is a general drive to obtain knowledge (about certain things) and it gives a person who has just obtained knowledge a positive feeling. In other words, a reward for knowledge gain is hardwired in the brain. This mechanism is considered an essential property of humans (Oudeyer and Kaplan 2004).

Here, I will take an different approach. This approach is based on the assumption that humans can learn that (at least a certain kind of) knowledge has positive long term effects, i.e. it can lead to a future increase of the received rewards. However, if knowledge gain is not directly rewarded, the only way for it to have a value, i.e. (according to the definition of value given in chapter 5) to lead to an increase in expected rewards, is by helping to select verbal or non-verbal actions which then lead to an increase in received rewards. To learn this value of knowledge it is necessary that the state of an agent contains a representation of the knowledge of the agent. The only difference between some states of the agent might then be a difference of the agent's knowledge. This enables the value function to learn to attribute different values to different states of knowledge. In this approach, a piece of information which does not lead to a long term increase in rewards will be judged as being of no value and, thus, not considered information worth obtaining. Humans, however, might be able to extrapolate from the benefits of particular knowledge to a positive effect of knowledge in general, i.e. they might learn that knowledge is usually something that is worth obtaining. In this thesis, however, I will not deal with such a generalization.

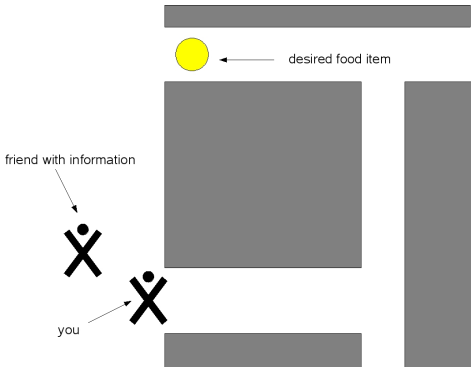

Figure 6.1: This shows a possible situation in which an agent does not have some but not all knowledge about the world. He needs to get a food item, but he does not know whether the items can be found in the east or in the north. He has two possible non-verbal actions to select from (moving east and moving north).

6.2.2 Learning the Value of Knowledge

I will now sketch an environment and an architecture which make it possible to learn the value of knowledge. I do this by introducing situations in which agents do not have the knowledge to decide which of two actions is the better one. They lack information about the state of the world relevant for making this decision, but can obtain this information by asking a question. Having this information they can make informed decisions that result in higher rewards in the long term.

Suppose that in this environment there is (among others) an agent A_1 who, as before, uses a forward model and a value function to select non-verbal or verbal actions.

The environment has 4 states $\{w_0, w_1, w_2, w_3\}$ described by a two-dimensional binary vector $\vec{v}$. In contrast to the setting in chapter 5, there is at least one property in this environment which A_1 cannot directly observe. This property is represented in v_1, i.e. the first dimension of $\vec{v}$. $v_1 = 1$ in w_0 and $v_1 = 0$ in w_1, w_2 and w_3.

$$w_0 = \begin{pmatrix} v_1: & 1 \\ v_2: & 0 \\ v_3: & 0 \end{pmatrix}; w_1 = \begin{pmatrix} v_1: & 0 \\ v_2: & 0 \\ v_3: & 0 \end{pmatrix}; w_2 = \begin{pmatrix} v_1: & 0 \\ v_2: & 1 \\ v_3: & 1 \end{pmatrix}; w_3 = \begin{pmatrix} v_1: & 0 \\ v_2: & 1 \\ v_3: & 0 \end{pmatrix} \tag{6.1}$$

Note that the only difference between w_0 and w_1 is the state of v_1 - and it is exactly this property that agents have no access to. Being part of the state of the environment $\vec{v}$, v_1 also has an effect on the outcome of actions. Among the actions that agent A_1 can select from there are two non-verbal actions a_1 and a_2. The probability of transitions from one state of the environment to another can be read off from the conditional probabilities in equations 6.2 - 6.13. As it can be seen from equations 6.6 - 6.13, only transitions from w_0 and w_1 are dependent on the actions of A_1.

$$p(w_2|a_1 \wedge w_0) = 1 \tag{6.2}$$
$$p(w_3|a_2 \wedge w_0) = 1 \tag{6.3}$$
$$p(w_2|a_2 \wedge w_1) = 1 \tag{6.4}$$
$$p(w_3|a_1 \wedge w_1) = 1 \tag{6.5}$$
$$p(w_0|a_1 \wedge w_2) = 0.5 \tag{6.6}$$
$$p(w_0|a_2 \wedge w_2) = 0.5 \tag{6.7}$$
$$p(w_1|a_1 \wedge w_2) = 0.5 \tag{6.8}$$
$$p(w_1|a_2 \wedge w_2) = 0.5 \tag{6.9}$$
$$p(w_0|a_1 \wedge w_3) = 0.5 \tag{6.10}$$
$$p(w_0|a_2 \wedge w_3) = 0.5 \tag{6.11}$$
$$p(w_1|a_1 \wedge w_3) = 0.5 \tag{6.12}$$
$$p(w_1|a_2 \wedge w_3) = 0.5 \tag{6.13}$$

Since the representation of the environmental state $\vec{v}$ is no longer fully accessible to agent A_1, he can not use $\vec{v}$ as input to his various cognitive functions (e.g. his value function). Instead, he has a dynamical knowledge base $\vec{b}$ which is part of the state of the agent. This knowledge base $\vec{b}$ contains the information about the environment which is accessible at every point in time plus information which is gained by means of communication. A_1 can use $\vec{b}$ as input to his cognitive functions.

The first two dimensions of $\vec{b}$ (b_1 and b_2) represent A_1's knowledge about v_1. These two bits are necessary, although v_1 only has two possible states. The reason is that b_1 and b_2 together need to have at least three possible states - they need to be able to distinguish between three possible states of knowledge: knowing that $v_1 = 1$, knowing that $v_1 = 0$ and not knowing anything about v_1. This is coded in the following way:

$$b_1 = 1 \wedge b_2 = 0 \Leftarrow A \texttt{ knows that } v_1 = 0 \tag{6.14}$$
$$b_1 = 0 \wedge b_2 = 1 \Leftarrow A \texttt{ knows that } v_1 = 1 \tag{6.15}$$
$$b_1 = 0 \wedge b_2 = 0 \Leftarrow A \texttt{ does not knows anything about } v_1 \tag{6.16}$$

There is a correspondence between world states $\vec{v}$ and knowledge base states $\vec{b}$. In equation 6.17 we can see that it cannot be decided whether the world is in states w_0 or in state w_1 if both $b_1 = 0$ and $b_2 = 0$. However, if either $b_1 = 1$ (equation 6.18) or $b_2 = 1$ (equations 6.18) then this decision can be made.

$$\vec{b} = \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix} \Rightarrow \vec{v} = \begin{pmatrix} v_1: & 1 \\ v_2: & 0 \\ v_3: & 0 \end{pmatrix} \vee \vec{v} = \begin{pmatrix} v_1: & 0 \\ v_2: & 0 \\ v_3: & 0 \end{pmatrix} \tag{6.17}$$

$$\vec{b} = \begin{pmatrix} b_1: & 1 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix} \Rightarrow \vec{v} = \begin{pmatrix} v_1: & 1 \\ v_2: & 0 \\ v_3: & 0 \end{pmatrix} \tag{6.18}$$

$$\vec{b} = \begin{pmatrix} b_1: & 0 \\ b_2: & 1 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix} \Rightarrow \vec{v} = \begin{pmatrix} v_1: & 0 \\ v_2: & 0 \\ v_3: & 0 \end{pmatrix} \tag{6.19}$$

$$\vec{b} = \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix} \Rightarrow \vec{v} = \begin{pmatrix} v_1: & 0 \\ v_2: & 1 \\ v_3: & 1 \end{pmatrix} \tag{6.20}$$

$$\vec{b} = \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix} \Rightarrow \vec{v} = \begin{pmatrix} v_1: & 0 \\ v_2: & 1 \\ v_3: & 0 \end{pmatrix} \tag{6.21}$$

The reward function for the states is given in equation 6.47.

$$r(w) = \begin{cases} 1 & \Leftarrow w = w_2 \\ -1 & \Leftarrow w = w_3 \\ 0 & \Leftarrow w = w_0 | w_1 \end{cases} \tag{6.22}$$

This means that our agent always receives a reward of 1 when the world is in state w_2 and a reward of -1 when the world is in state w_3. However, since the value function is no longer trained with state of the worlds $\vec{v}$ but with states of the knowledge base $\vec{b}$, the value function will learn the corresponding values of $\vec{b}$.

$$V \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix} = 1; V \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix} = -1 \tag{6.23}$$

So far, the agent has only learned that certain states of the world, represented in his knowledge base $\vec{b}$ have a specific high or low value. To understand the value of knowledge the agent has to learn that certain states of $\vec{b}$ that contain more information about the state of the world $\vec{v}$ lead to better actions and therefore higher rewards. Actions are selected using the internal model. The internal model we will use in this chapter differs significantly from the one in previous chapters. For one, like the value function, the internal model can no longer use states of the environment $\vec{v}$ as input and output (since they are not accessible), but has to use the knowledge state $\vec{b}$. Further, since the effects of actions are no longer deterministic, the internal model has to be able to acquire probabilistic state transitions. How such a probabilistic model $P()$ is trained will be explained in section 6.2.4.

For now, it is sufficient to know that a probabilistic internal model learns to map triples consisting of an utterance, a context state and an effect state onto probability values.

The transition probabilities of this world are given in equation 6.2 - 6.13. But since the states of the world $\vec{v}$ are not accessible the agent has to learn the transition probabilities of states $\vec{b}$ of the knowledge base. In equation 6.24 the acquired transition probability is the one corresponding to the one given in equation 6.2. The transition probability is one, because the agent, since $b_1 = 1$, is well informed that he is in state w_0

$$P(\begin{pmatrix} b_1: & 1 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_1, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix}) = 1 \tag{6.24}$$

If $b_1 = 0$ and $b_2 = 0$ (with the same value for b_3 and b_4), i.e. the agent does not know whether he is in state w_0 or in state w_1, then the transition probability to w_2 is only 0.5 (equation 6.25) while the transition probability to w_3 is also 0.5 (equation 6.26).

$$P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_1, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix}) = 0.5 \tag{6.25}$$

$$P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_1, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix}) = 0.5 \tag{6.26}$$

Since agents base the selection of actions on the prediction of their internal model, an agent will always select action a_1 if he knows that he is in state w_0 (i.e. if $b_1 = 1$, $b_2 = b_3 = b_4 = 0$) and he will always select a_2 if he knows that he is in state w_1 (i.e. if $b_2 = 1$, $b_1 = b_3 = b_4 = 0$). Both, a_1 done in the context w_0 and a_2 done in the context of w_1 lead to state w_3 and a reward of 1. Therefore, a knowledge state in which $b_1 = 1$ or $b_2 = 1$ (with $b_3 = b_4 = 0$) will also always by followed either state w_3 (and a reward of 1). Hence, the value of these two knowledge states will be increased to 1 as well.

$$V(\begin{pmatrix} b_1: & 1 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = V(\begin{pmatrix} b_1: & 0 \\ b_2: & 1 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = 1 \tag{6.27}$$

To select an action with a probabilistic internal model the value of the outcome for an action a is computed in the following way: for every possible outcome of this action,

the likelihood of a transition to this outcome $\vec{b}$ has to be multiplied by the value of this outcome (equation 6.28)

$$\sum_{\vec{b_{t+1}}} P(\vec{b_t}, a, \vec{b_{t+1}}) \times V(b_{t+1}) \tag{6.28}$$

$$P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_1, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix}) \times V(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix}) + P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_1, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix}) \times V(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix}) \tag{6.29}$$

$$P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_2, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix} \times V(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 1 \end{pmatrix}) + P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, a_2, \begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix}) \times V(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 1 \\ b_4: & 0 \end{pmatrix}) \tag{6.30}$$

Since the result of both terms is 0, the agent has no basis for preferring either action a_1 or action a_2. This means, that if the agent does not know whether he is in state w_0 or in state w_1 he will end up choosing actions more or less at random. This will result in a transition to w_2 in some cases (providing a reward of 1 and in a transition to w_3 in about the same number of cases (providing a reward of -1). Hence, the value of this state will converge to approximately 0 (equation 6.31).

$$V(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = 0 \tag{6.31}$$

$$V(\begin{pmatrix} b_1: & 1 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = V(\begin{pmatrix} b_1: & 0 \\ b_2: & 1 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = 1 \tag{6.32}$$

In contrast, if the agent knows whether he is in state w_0 or in state w_1, he will always select the action resulting in a higher reward. Therefore, the value of the corresponding states of his knowledge base will converge to 1 (equation 6.32). This is the point where the agent has learned to appreciate the value of this knowledge.

6.2.3 Predicting the Effect of Questions

The main reason why it is important to explain how humans learn to appreciate the value of knowledge is that this appreciation motivates an important type of speech act: *questions*. In the previous section, I merely explained how agents could develop a preference for being in a state in which they have more information about the world. But to motivate the speech act of asking a question, the agents need to have the possibility of deciding between a performance of an action that will lead to chance in the environment but is based on

insufficient information and an act that may or will lead to obtaining more information e.g. by asking a question. So far, the agent A_1 can perform the non-verbal actions a_1 or a_2. A question about the property v_1 of the world will be a verbal action (or utterance) u_1. For now, I will not deal with the possible answers that A_1 is maybe getting from other agents, nor will I deal with the problem of how these answers are processed by him so as to obtain the knowledge whether either $b_1 = 1$ or $b_2 = 1$. The only thing that matters at this point is that the question somehow results in a subsequent state in which the required knowledge is present. Put in terms of a probabilistic internal model, the agent learns that asking the question leads to his knowledge that the world is in state w_0 in some cases (equation 6.33) and that it is in state w_1 in about the same number of other cases (equation 6.34).

$$P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, u_1, \begin{pmatrix} b_1: & 1 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = 0.5 \tag{6.33}$$

$$P(\begin{pmatrix} b_1: & 0 \\ b_2: & 0 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}, u_1, \begin{pmatrix} b_1: & 0 \\ b_2: & 1 \\ b_3: & 0 \\ b_4: & 0 \end{pmatrix}) = 0.5 \tag{6.34}$$

To choose between asking a question and performing a non-verbal action it is necessary to compute the value of the outcome of both types of actions. As in the previous section, the outcome of both action a_1 and action a_2 is 0 if the agent does not know whether he is in state w_0 or w_1. If he knows that he is in state w_0 then the outcome of a_1 is 1 and of a_2 is -1, and, vice versa if he is in state w_1. To compute the outcome of asking the question we have to use again the formula described in equation 6.28. As we have seen in equations 6.33 and 6.34, the probability of getting to either of the two fully informed knowledge states (with either $b_1 = 1$ or $b_2 = 1$) is 0.5 and the value of both states is 1 (equation 6.32). Therefore, since the relevant question will necessarily lead to one of these two knowledge states, the value of performing this question is 1 (equation 6.35).

$$0.5 \times 1 + 0.5 \times 1 = 1 \tag{6.35}$$

This means that the agent will not select a non-verbal action if he does not know in which state he is, but instead he will get this information by asking the relevant question. In general this architecture allows agents to learn in which situations it is good to ask questions instead of performing non-verbal actions. Of course, the same principle also works for larger representations of knowledge and more complex worlds and game situations.

6.2.4 A Probabilistic Internal Model

Since a probabilistic internal model $P()$ is used in sections 6.2.2 and 6.2.3, I will briefly sketch in this section how such a model works and how it is trained. A probabilistic forward

model maps triples of utterance, context states and effect states to probability values.

A probabilistic internal model can be trained by storing the co-occurrences of context state, effect state and utterance. Consider for example the world with four possible state representations $\{s_1, s_2, s_3, s_4\}$. Usually if an agent does not say anything, the world moves from s_1 to s_2, from s_2 to s_3, from s_3 to s_4, and from s_4 back to s_1. Now if an agent uses utterance u_1 in the context of s_1 he will bring about s_3. And he can also change the world state from s_4 to s_2 by using utterance u_2. To indicate that an agent is not saying anything, I will use u_0 - the null or empty utterance. It simply means that the agent did not say anything in the given context state.

As a formalization of the storage device, I will use a function $C()$ (for co-occurrence) which maps triples of effect states s_e, context states s_c and utterances u into pairs $< n_c, n_e >$. n_c is the memorized number of times u was made in the context of s_c, while n_e is the memorized number of times a state transition to s_e was observed after u was made in the context of s_c. In our example it might have been the case that the agent has observed a transition from s_1 to s_2 when he said nothing for 11 times (equation 6.36), never heard or used u_1 in state s_1 (equation 6.37), and once used u_2 in s_1 without observing a transition to s_2 (equation 6.38)

$$C(s_2, s_1, u_o) =< 11, 11 > \tag{6.36}$$
$$C(s_2, s_1, u_1) =< 0, 0 > \tag{6.37}$$
$$C(s_2, s_1, u_2) =< 1, 0 > \tag{6.38}$$

The effect of the utterance u_1 in situation s_1 can be stored by the same mechanism.

$$C(s_3, s_1, u_o) =< 0, 0 > \tag{6.39}$$
$$C(s_3, s_1, u_1) =< 6, 6 > \tag{6.40}$$
$$C(s_3, s_1, u_2) =< 1, 0 > \tag{6.41}$$

Now the probability of a transition can be calculated by the formula given in 6.42

$$p(s_e|s_c \wedge u_1) = \begin{cases} \frac{n_e}{n_c} & \Leftarrow s_c \neq 0 \\ 0 & \Leftarrow s_c = 0 \end{cases} \tag{6.42}$$

In our example the probability of a transition from s_1 to s_3 if u_1 was made is 1; if u_2 was made or nothing was said (u_0) the probability is 0.

6.2.5 Learning to ask Questions

In the setting described above, the learning agent is able to learn the effect of questions by trial-and-error, because the pre-programmed agents will answer his questions and thereby change his knowledge base. It is, however, not plausible that children learn the effect of

questions by making an utterance which accidently has the form of a question and thereby learn that such a form leads to an answer. Rather, children learn the effects of questions from their interaction with speakers of the language (such as their parents). It is known from research in language acquisition that parents play question-answer games with their children. The parent asks a question like (1) and if the child does not answer herself, she gives an answers like (2), while pointing out the ball.

(1) Where is your ball?

(2) There is your ball.

Children can use this pattern to learn the connection between questions and answers. On one hand they learn the form of questions and answers, on the other hand they also learn that a question is usually followed by a (verbal or non-verbal) answer, and that this answer supplies the person asking the question with new information. Of course, the child has to understand that the effect of the question is not the *sound* of a verbal answer or the gesture of a non-verbal one, but the knowledge it provides. In other words it has to learn the relation between the content of the question and the information provided by the answer. This relation is a difficult topic which I cannot discuss here in detail. In the approach to language processing with forward model and value function, an agent only has to come to understand that in certain contexts (in which some specific knowledge is lacking) particular questions lead to states with more knowledge (states which, therefore, have a higher value). How the relation between questions and answers is stored in the forward model will be discussed in section 6.4.4.

When the information about the effect of questions is stored in the forward model of the language acquiring child, it knows that a question (such as the one about the ball) requires a verbal answer or some other kind of reaction (such as pointing to the ball) and will, if possible, provide such an answer when it is asked such a question. Furthermore, the child at that point has also learned that it can use questions to obtain information itself. Note the connection between requests and questions: Question too are requests, but requests of a special sort, i.e. requests for information. The difference with other requests usually manifests itself in the kind of reaction that the addressee is invoked to make. In the case of requests other than questions the reaction is normally some non-verbal action. In the case of questions it is usually a verbal answer (and thus a verbal action), although it can also take the form of a non-verbal action, such as e.g. a pointing. It is this way of answering a question, and this way only, which I am investigating here.

To show how the question-answer-pattern can be learned within my theoretical framework, consider a scenario in which agents can point to objects, and can also focus their attention on a certain object, without which the action of pointing to an object would make no sense. Questions are asked about object positions and answers are given in terms of pointing gestures. The answer to a question about the position of a certain object is given by pointing out the object.

Attention is a very important factor in the neural modeling of language since one of the major purposes of language is to direct the attention of addressees to certain objects

or states of affairs. Also gaze following and shared attention are mechanisms known to be essential prerequisites to learning the meaning of words (Tomasello 2003). I will talk more about attention and how it can be dealt with in my theoretical framework in chapter 7.

Attention is considered to be represented by an attention map in the posterior inferior parietal sulcus (Itti and Koch 2000). Such an attention map can be represented as a matrix of binary values. With such a representation and a simulated pointing device the learning of question-answer patterns can be simulated.

In the following environment there is a pre-programmed parent agent (PP) and a neural network - based child agent (NN). The parent agent teaches the pattern to the child agent. There are three (holophrase) questions possible:

(3) question-triangle

(4) question-square

(5) question-diamond

The *conventional effect* of these questions (i.e. the effect programmed into the pre-programmed agents), is to trigger a non-verbal pointing to an object. The object in question is a triangle in the case of (3), a square in the case of (4), and a diamond in the case of (5).

The game works in the following way: It consists of a number of learning episodes. In each episode, the PP-agent asks a question (such as *question-triangle*) and waits one time step for the answer of the NN-agent. If no or a wrong answer is given (the answer is given in terms of a pointing gesture), the PP-agent points towards the object himself. After that, the next episode starts.

To keep the matter simple, the environment consists of only four fields, each of which can contain an object of one of the three object types. These four fields are presented to the agent in form of a 3×4 matrix, where every column-vector codes one of the three object types (triangle, square, diamond) while the position of the object is coded by the position of the vector in the matrix. In equation 6.43, the first vector would mean that the left most position is occupied by a triangle, the second by a square, the third by a diamond, and the fourth is an empty position.

$$w_k = \begin{bmatrix} 1000 \\ 0100 \\ 0010 \end{bmatrix} \tag{6.43}$$

These three objects (one triangle, one square, one diamond) are shifted to a different position at the beginning of every new episode. The pointing gesture of the agents are coded as a 4-dimensional horizontal binary vector $\vec{p}$ (e.g. in equation 6.44, the agent points the the second object). The attention map is coded by another 4-dimensional horizontal binary vector $\vec{a}$ (e.g. in equation 6.45, the agent attends to the fourth object). If a pointing gesture is made by one agent, the attention of the other agent automatically follows the pointing gesture.

$$\vec{p} = \left[\begin{array}{c} 0100 \end{array} \right] \tag{6.44}$$

$$\vec{a} = \left[\begin{array}{c} 0001 \end{array} \right] \tag{6.45}$$

The learning agent uses the input given by the PP-agent to train his forward model. He learns that a specific question is answered by a specific pointing gesture. His forward model is mapping the state of the world w_t and a question to an attention vector $\vec{a}$.

$$a(t \vec{+} 1) = F(w(t), q(t)) \tag{6.46}$$

Similarly to previous simulations, the agent trains his forward model by observation and will, after a certain number of time steps, be able to answer the question by bringing about the intended state of the speaker.

The reward r in this game is given by a reward function $R()$ which maps triples of world state, question, and attention map to a positive real number in case of successful pointing and to a negative real number in case of failure, i.e. pointing to the wrong cell or no pointing at all (equation 6.47).

$$r = R(w_t, q^A_{t-1}, \vec{a}^A_t). \tag{6.47}$$

The three questions can be represented as three vertical 3-dimensional binary vectors as in equations 6.48, 6.49, and 6.50.

$$question - triangle \rightarrow \left[\begin{array}{c} 1 \\ 0 \\ 0 \end{array} \right] \tag{6.48}$$

$$question - square \rightarrow \left[\begin{array}{c} 0 \\ 1 \\ 0 \end{array} \right] \tag{6.49}$$

$$question - diamond \rightarrow \left[\begin{array}{c} 0 \\ 0 \\ 1 \end{array} \right] \tag{6.50}$$

Rewards are given to the child for correctly pointing out the object. Such a reward is a social reward, which the child obtains for cooperation and correct behavior. In this potential simulation, the child is not supposed to ask questions itself and, therefore, is not rewarded for obtaining knowledge, as in the last scenario. The reason for the child to adapt to the behavior which is conventional in this environment - the behavior of pointing out the objects when being asked about them - can be seen in its desire for social recognition.

Examples of triples of world states, questions, and attention maps for which the child receives a positive reward are given in equation 6.51 and 6.52).

$$< \begin{bmatrix} 1000 \\ 0000 \\ 0000 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1000 \end{bmatrix} > \tag{6.51}$$

$$< \begin{bmatrix} 1000 \\ 0010 \\ 0000 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0010 \end{bmatrix} > \tag{6.52}$$

The forward model then needs to be trained to map pairs of questions and world states to the correct attention maps as in equations 6.54 and 6.55

$$F(w(t), q(t)) = a(\vec{t+1}) \tag{6.53}$$

$$F(\begin{bmatrix} 1000 \\ 0000 \\ 0000 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}) = \begin{bmatrix} 1000 \end{bmatrix} \tag{6.54}$$

$$F(\begin{bmatrix} 1000 \\ 0010 \\ 0000 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}) = \begin{bmatrix} 0010 \end{bmatrix} \tag{6.55}$$

To construct a neural network which can be trained to find this mapping is a trivial matter.

6.3 Assertions

6.3.1 Complex Desires

In this section I will introduce assertive speech acts into my theoretical framework. So far, questions and requests have been analyzed as expressions of desired states - in the simulations of chapter 5 the process of understanding a request involved nothing more than the computation of the desired state which the speaker was trying to bring about. In section 6.2 questions have been shown to function along similar lines. Hence, it seems natural to analyze assertions as expressions of the desires and - similar to requests and questions - all an addressee needs to do to understand an assertion is to compute the desired state which triggered the utterance. However, assertions are more difficult to analyze in my theoretical framework, because agents only choose an utterance if it leads to a state with a higher value than states that can be brought about by other utterances or actions. While it is easy to see how requests and questions (since they bring the speaker useful information) can lead to states with a higher value, assertions appear neither to alter the world, nor the knowledge of the speaker himself into a state with a higher value. Assertions can nevertheless bring about states with a higher value for the one who produces them. To explain how this is possible is the main focus of this section. It involves the description of utterance effects on several levels.

On one level, assertions tend to produce new information for the addressee and thus to change his knowledge state. To represent the speaker's desire to cause this effect, it is necessary, to begin with, to represent his knowledge about the knowledge of other agents. So far I have only introduced a device to store knowledge about states of the world. This device would need to be extended accordingly. However, it is not known what kind of neural representations the human brain uses to represent this kind of information. Psychological experiments seem to suggest that representation of this kind might be of a significantly higher level of complexity. Apparently young children acquiring a first language are, according to these experiments, not yet capable of representing the belief states of other persons in case these belief states differ from the child's own belief state (see Breheny (2002) for a summary and discussion of this literature). So if the primary motivation for declarative utterances would be to change the belief state of another agent, small children, as long as they lack this representational possibility, would not be able to make declarative speech acts.

Besides this problem there is also another one which shows that the desire to change the knowledge state of the addressee cannot be the only desire triggering an assertion. Suppose a person A_1 addresses another person A_2 on a sunny day at the bus stop with the utterance (6).

(6) $A_1 \rightarrow A_2$: Nice weather today.

It cannot be said that it is A_1's desire to change the knowledge state of the agent by adding the fact that the weather is nice today, since A_2 already knows that. Rather, the desire of A_1 is to start a conversation and to socially connect to A_2. So the desire of A_1 is to start a conversation and talking about the weather is a standard way of starting a conversation. If the process of understanding an utterance is by means of computing the desire which triggered the utterance, then A_2 will compute from the utterance that A_1 has the desire to start a conversation and it is very intuitive that this is what A_2 will understand. However, the problem with this approach is that the theory so far claims that the process of understanding an utterance consists *only* of the process of computing the desired state that triggered the utterance and of nothing else. Now suppose A_2 answers with (7).

(7) $A_2 \rightarrow A_1$: It will be even better tomorrow.

If the only desire that triggered utterance (6) is the desire to start a conversation and the process of understanding an utterance is only the decoding of this desire, then A_2 does not have the information that the utterance (6) was about the weather. This, of course, is totally absurd.

So is it after all the case that besides the desire to start a conversation, A_1 has the desire to change the knowledge of A_2 in a way that A_2 might after that know that the weather is nice? No, not unless A_2 is without any sense of temperature.

To problem lies in my assumption that the knowledge state of the addressee can only be changed by adding new conceptual information (i.e. by adding new information about an

individual or categorical concept). But consider that although A_2 knows that the weather is nice, he might not be conscious about it at the moment in time or, in other words, he might not have his attention directed to this fact. Just possibly A_1 might have wanted to bring the beautiful weather to A_2's attention. In neural terms, A_1 might have wanted to activate a conceptual representation in A_2 of today's nice weather, i.e. although the knowledge about the weather might already be present in A_2, the activation of a conceptual representation of this state of affairs might still constitute a change in his mental state. This effect of the activation states of concepts can also be regarded as a change of context (at least of a neural representation of context within the brain of the addressee) and in that respect is related to the idea of Stalnaker (1979) or Kamp (1980) to regard the meaning of utterances as context change potential.

The neural activation of existing concepts is not only necessary for somewhat exceptional cases such as utterance (6), but also for most cases of verbal information transfer: If we give an addressee information about a specific individual we need to activate a conceptual representation which is already present. For example, if we want to tell an addressee that the man we just saw is my brother, we need to activate the concept of this person and the concept of brother and make the addressee store this relationship. However, this kind of process uses compositionality, which my theoretical framework so far does not have and I have to put this issue off until chapter 7.

For now it is sufficient to realize that a representation of the *knowledge base* (as I have used it so far) is not adequate to describe the addressee's processing of declarative (and maybe other types of) utterances. However, the necessary neural basis for the extension of the framework was already introduced in chapter 4. There, I explained the distinction between a neural representation of concepts in working memory and of concepts in long term memory. This distinction in my framework can be used to distinguish between knowledge and the mere activation of concepts needed to solve the problem described above.

Since it can be a desire of a speaker to change the conceptual activation of an addressee, utterances like (6) do no longer pose any serious problems for the assumption that declarative utterances can be understood as expressions of desires. Following this analysis the desire state triggering utterance (6) has two components: (i) starting a conversation and (ii) activating a conceptual representation of (6). To understand the relation between these two components, let me compare it to a similar relation in motor control. Consider the following pair of desires: (i) picking up the phone and (ii) moving my hand. Of course, I need to move my hand in order to pick up the phone, just like I activate the conceptual representation in the brain of the addressee in order to start a conversation. However, unlike in the example of the phone, where the relation between the goals of moving the hand the picking up the phone is not separable, the relation between the goal of starting a conversation and the goal of activating exactly the conceptual representation and connection of *nice*, *weather* and *today* is somewhat arbitrary. A_1 might well have started the conversation with (8).

(8) $A_1 \rightarrow A_2$: If the bus is not here in one minute, I will miss my train.

This seem to suggest that there is a primary social goal and a secondary conceptual

goal which is selected as a means of fulfilling the primary goal. However, this is not the case in general. Take for example sentence (9).

(9) $A_1 \rightarrow A_2$: Your shoelace is untied.

Here, A_1 uses (8) in order to help A_2 avoid stepping on his shoelace and stumbling. A_1's motivation to do this might be social - either because, as a human being A_1 has a cooperative nature (e.g. because there is an inborn mechanism that triggers internal rewards for cooperative behavior) or A_1 is cooperative for the sake of gaining a good *social reputation* or increase the *social bondage* with the other person[1]. So here the goal on one level is to prevent A_2 from stumbling down. On a second level, the goal of A_1 is to provide the conceptual information to A_2 that his shoelaces are untied - and to direct his attention to this fact. In this example the primary goal and the secondary goal are tightly connected and cannot be separated. And one cannot say that A_1 first had the primary goal of preventing A_2 from stumbling down and, therefore, selected the conceptual information. Rather, it is likely that the conceptual information was perceived by A_1 and this triggered the goal of preventing A_2 from falling down. These examples show that the relation between the conceptual level of the desire and what Austin (1961) has called the illocutionary and perlocutionary act cannot easily be put in a simple temporal or causal relation. It is not the case that the social desire triggers the desire to give a specific conceptual information. However, a better analysis of this relation will be possible when compositional aspects of language will be dealt with in chapter 7. For now it is important to understand that the desire which triggers the declarative speech act appears to be a complex of desires which bear a relation to each other that is hard to analyze with the current state of the theoretical framework. This complex usually consist of (i) a desire to cause a certain conceptual state in the addressees brain and (ii) some form of social desire.

6.3.2 Communicative and other Intentions

According to Levelt (1989), the communicative intention is the intention which the speaker want the addressee to understand. In my theoretical framework - as I have presented it this far - an agent A_1 uses an utterance u to address another agent A_2 because u brings about a state b_x with a higher value than the states that would come about as the result of no action or other verbal or non-verbal actions. Utterances are understood by decoding this intended state b_x from the utterance u. This means that information which is not given as part of b_x is not perceived by the addressee. In section 6.3.1, I have shown that the information transmitted by an assertion can be modeled in my general theoretical framework by assuming that it is one part of the desire of the speaker to activate a certain conceptual configuration in the knowledge base b_x of the addressee.

So far, I have also been making the assumption that the desire that triggers the utterance is the communicative intention of the speaker, and it it *this* communication intention

[1]There is no space in this thesis to explain and discuss the current theories about the cause and the nature of human cooperation.

that is transmitted to the addressee, i.e. this is the information the addressee will receive. However, this cannot be right, because often utterances are triggered by intentions that we do not want the addressee to understand. For example, A_1 might use utterance (10) with the intention to impress A_2.

(10) $A_1 \rightarrow A_2$ Yesterday, I bought a Porsche.

Is it therefore justified to give up the theory that utterances are selected as described in equation 6.56 and utterances are understood as described in equation 6.57 - because in that case the desire triggering the utterance will always be the information transmitted to the addressee?

$$u_t = argmax_u V(F(s_t, u_t^*)) \tag{6.56}$$

$$s_{t+1} = F(s_t, u_t) \tag{6.57}$$

I think it is not justified to give up this theory. Even in the case of this example it might in fact happen that A_1 unintentionally transmits his intention to impress A_2, because A_2 knows that he himself would want to impress someone by such an utterance. The only possible way of not transmitting this goal by disguising the original goal - in other words by pretending to have another goal. And pretending, just like lying, adds a layer to the modeling of information transmission that might violate the general principles developed so far in many other ways as well, and I simply save myself from dealing with these issues in this thesis.

However, is it really the case that every assertion is understood as an expression of a desire to cause a change in the mental state of the addressee? Consider for example the utterance (12) made by $A1$ - should it always be interpreted as (13) or can A_2 directly decode it into (14)?

(12) $A_1 \rightarrow A_2$: *Here is a triangle.*

(13) A_1 wants me to believe that there is a triangle at the position of A_1.

(14) There is a triangle at the position of A_1.

In case utterances are understood as (14), then the given information can be directly added to the knowledge base unless it is inconsistent with existing knowledge. In the case of such an inconsistency, A_2 could, by using an additional inference step, realize that although (14) is maybe not true A_1 wants A_2 to believe it. Thus A_2 would only generate a representation of (14) in case of inconsistencies. In case utterances are generally understood as (13), then first the consistency of (12) can be checked first and in case there is no inconsistency, the information is added to the knowledge base (as in Heim's File Change Semantics).

In the light of Grice's insights I will model the transmission of descriptive information as embedded in an intention. This also allows me to keep the general framework for modelling

comprehension as a process of transforming every utterance into the expression of a desired world state, only that now this world state includes the knowledge base of the addressee. A neural model that includes a representation of the knowledge base of the addressee is introduced in chapter 7.

6.3.3 Reasons to Change Knowledge

In correspondence to the two kinds of effects of declarative acts: (i) the effect on the addressee's state of conceptual activation and (ii) the effect on some form of social reputation for the speaker, the value of the effect state also has to vary along those lines. To model social desires triggering assertions, the theory needs to involve a component in which the social status of a speaker is incorporated in the overall value of the state, so that a social action leads to an increases of the value of the state.

Besides social reason, assertions can also be triggered by the need to change the knowledge base of the addressee for a purely selfish reason. The reason is that the speaker often needs to give the addressee information he needs in order to accomplish a goal of the speaker. For example, if the addressee does not know the location of a certain object, which the speaker would like to request, the speaker needs to tell him the location of this object. Often it might not be possible to give such information in a single utterance. For example when a speaker wants an addressee to cook a certain dish, he has to explain s sequence of steps to him involving several utterances allowing him to create the knowledge about a more complex process.

6.3.4 The Social Value of Declaratives

Although assertions often do bring the addressee into a state of knowledge which may lead to a long term increase in rewards for the speaker (because they are relevant for the hearer to fulfill a desire of the speaker), many (or maybe most) assertions, especially in the early stages of language acquisition, seem to have a primarily (and maybe instinctively) social motivation. As observed by Dunbar (1996), the social bondage can be increased already just by the general act of talking. But assertions can have more specific social effects. As already mentioned above, they are often made for the purpose of improving or maintaining the social status of the speaker. Answering questions is among the best examples of assertive speech acts which are made for that purpose. But information can also be shared for the very same reasons, without a triggering question.

What is the benefit for the speaker, if he answers questions or shares relevant information with another person? At first sight, it appears to lead to a state with a lower value. For example, by voluntarily pointing out the location of a food source, an agent might lose potential food, because the addressee might eat some of the food at that location. However, if the addressee stores this benevolent act in his memory (stores which persons performed such social acts for his benefit) and in turn might give relevant information to this agent in the future, it is not unlikely that the speaker will increase his overall reward by such an act. This means that the increase in reward is indirectly obtained through a

change in social status. To allow the agent to learn that such an improved social status has a high value (i.e. might lead to a long term increase in reward) the social status must be a part of the state representation (i.e. be among the arguments of the value function and part of the results predicted by the forward model). However, the desire to gain and maintain a good social reputation might not only be motivated by its power of increasing the obtained long term reward. The social reputation (instinctively or by education) is something that humans desire for itself.

Also the sharing of (personal) information enhances the social relation among the members of a group for the following reason: such information allows other persons to better understand the emotional life of their fellow agents. Through a better access to the value function of another person, sympathy and empathy can increase and therefore also enhance the emotional connection to this person. Although this effect can, of course, be learned to a certain degree, it also appears to be to some degree a genetically predispositioned pattern.

Another pattern of sharing information, which appears to be clearly instinctive, and, therefore, can be observed already in small children, is to share information that is (emotionally) significant. The biological reason for humans to share such kinds of information is that from an evolutionary perspective humans shared significant stimuli (food sources, predators etc.). Stimuli pointed out by the child (emotionally significant for it) might however not be emotionally significant for the adult. This is simply because the child has less experience than the adult and generally regards novel or noisy or other objects with outstanding features as significant.

To simulate children and adults pointing out emotionally significant objects the theory needs to be able to detect such objects. This can be done with the value function. Objects involved in state changes which correlate with a significant change in value can be regarded as significant objects. For a child of a certain age, it might probably be everything that is new especially if it is in a certain way spectacular (such as colorful or noisy things) or socially relevant, such as new persons.

Adults often point out objects to children and give them a verbal label simply because they want to teach the name of the object to the child. This might be because the adult consciously wants to teach the child the name or because he simply follows some parental instinct.

Such a situation could be simulated by making an adult agent pick out an emotionally significant object and name it. His goal would be to make the child share the attention towards the object. This would give the child the opportunity to learn the name of the object. Then, when it is the child's turn to name the object it would be rewarded if it could successfully get the adult to share the attention towards objects which are emotionally significant to him by using a correct verbal label. I will not discuss technical details of such a simulation, since there are already many simulations of this kind (e.g. Steels (1996) or Steels (2001)). However, the selection of significant objects by means of a value function would be a new feature of such simulations.

To conclude this section, let me mention that truth is also an important social aspect when assertions are made. In a framework in which people are speaking to maximize their

rewards, there are occasions in which it might be to the advantage of a person not to answer a question truthfully or even to initially provide false information to lead somebody astray. But, as already said, these aspects are beyond the scope of this thesis.

6.4 Selecting Among Different Speech Acts

6.4.1 The Environment

With this theoretical approach to what I believe to be the major types of speech acts, it is now possible to model the selection of different speech acts within this framework. As before, agents are able to decide whether to perform non-verbal action or verbal action, and the verbal actions are either declaratives, imperatives, or interrogatives.

Given that we have triangles, squares and diamonds and three types of speech acts and assuming that we have only single-word utterances without any compositional features, we have 9 possible forms of utterances:

(1a) req-triangle

(1b) req-square

(1c) req-diamond

(2a) quest-triangle

(2b) quest-square

(2c) quest-diamond

(3a) assert-triangle

(3b) assert-square

(3c) assert-diamond

Since it is hard enough to create an environment in which each of those speech act types is useful in a different manner to increase the total reward, I will only use pre-programmed agents for the time being and do not concern myself with the problem of learning.

I will assume that (i) agents use utterances of the form of 1a-c because they get a direct reward for getting a certain object in a certain situation; (ii) agents use 2a-c because they will obtain knowledge about the location of objects which will lead to an increased reward in the long run; and (iii) agents use 3a-c because that will improve or maintain their social status. This also helps them to obtain a higher reward in the long run, since other agents decide whether or not to cooperate with an agent based on the social status of this agent.

Agents in this simulation move on a 10×10 grid world. New objects are placed at random locations. Agents have an internal map (a short term memory map) of the

environment in which object locations can be memorized for $n = 10$ time steps after an agent has encountered such an object. Objects can be perceived only if the object is in the same field as the perceiving agent. The map also shows the position of the agent himself and the position of the other agents, but not which objects these agents have. Every agent has an internal representation of the world including a representation of the objects about which they know. This representation can be thought of as the knowledge base $\vec{b}$ introduced in section 6.2.2 and is distinct from the current state of the world. Agents learn about object positions by perceiving the object in case they are close enough, or by receiving information from other agents. Often they might not have an object represented at a certain position although it is in fact there, because they have not perceived it yet and received no information about it. Or they might have an object represented at a position which no longer can be found there, since it was taken by another agent.

An agent can carry a maximum of $maxt = 3$ objects and needs to digest one object in each 10 time steps. Every 10 time steps the agent has to digest a different type of object, but always a triangle after a diamond, a square after a triangle, and a diamond after a square. If the object which is supposed to be eaten at a time step of digestion is not in the possession of the agent he gets a punishment of -1; otherwise he gets a reward of 1.

The non-verbal actions of the agents are:

- movement in one of four possible directions
- taking an object
- dropping an object

The possible movements of the agents are a single step in either of the four directions (north, east, west, south) in the grid. The effect of such a movement is that the agent leaves his current position and moves to the next position in the direction of the movement. If the border of the grid world is reached, a movement in the direction of the border has no effect.

Taking an object results in the agent having this object unless the agent has already $maxt$ objects; then it has no effect.

Dropping an object results in the agent no longer having this object. The object is then at the position in the grid at which the agent is at the time of the dropping action. A dropped object made in a field in which there is also another agent can be picked up by this other agent.

6.4.2 The Social Score

The social score is essential in this formal framework. It is defined for any two agents in both directions, i.e. for two agents A and B, $Soc_A(B)$ gives the past experience A had with B and B had with A. The score is a number between 0 and 10. For any cooperation (dropping of a requested object) the number is increased (+ 1), and every time an agent does not cooperate (i.e. does not drop a requested object), the number is

decreased (- 1). For every answered question, the number is also increased (+ 1) and for an unanswered question, the number is decreased (- 1). The score is also increased if agent share information about object positions with another agent (assertions) even if they are not asked for this information. Agents cannot improve their social status by pointing out an object more than once to the same agent.

Although agents will be engaged in social acts, their behavior is egoistic in the sense that all their verbal and non-verbal actions have the aim of maximizing their overall reward. To make social acts pay off in the long run, unsocial behavior needs to have a strong disadvantage. This disadvantage is accomplished by making the agents punish non-cooperative behavior of other agents by letting previous acts of cooperation of an agent determine their willingness to cooperate with this agent. More precisely, they determine whether or not to cooperate with another agent, i.e. whether to drop a requested object or answer a question by weighing up against each other the cost of cooperation, the benefit of cooperation, and the social status of the speaker (what one knows about that agent's willingness to cooperate).

6.4.3 Verbal Actions

Requests

An agent has a desire of degree 3 for an object if this object is the next one he needs to eat and he does not have it. If it is not the next, but the one after the next, his desire is 2, and if it is the last object that he needs, his desire for that object is 1.

A request is made if one agent (henceforth called A) desires a particular object of another agent (henceforth called B). Note that an agent A can only see that another agent has a certain object B if the two agents are in the same position of the grid world. B will drop the object if the calculated ratio between the social score and his own desire of the object is below a certain threshold. This threshold is either fixed before the game or estimated during the game. In case it is estimated during the game, all other variables should be fixed.

Questions

A question is asked by an agent A, if he desires a certain object which he does not have and he also cannot see whether another agent has it. A question is answered by an (affirmative) assertive utterance if the addressee B has the object e.g. question (2a) of A might be answered by the assertion (3a) of B in case B has the object; no answer is given otherwise (since there are no negations in this language). Since A cannot know whether B's silence means that he does not have the object or whether it is an act of non-cooperation, B needs a way to indicate his willingness to cooperate. He can do that by answering the question with information about another object he has. E.g. he could use utterance (3b).

(2a) quest-triangle

(3a) decl-triangle

(3b) decl-square

Assertions

Besides assertions that are used to answer questions, assertions can also be made if an agent A discovers an object (i.e. he is on the same field as this object) and his desire to improve his social status with agent B is stronger than his desire to keep the object himself. This allows agents to improve their social status even if they are not asked questions or asked to drop objects.

6.4.4 The Mathematical Framework

The Agents

Every agent A is given by a 4-tuple $A =< O_A, M_A, S_A, D_A >$, where O_A is the object store of agent A, M_A is a short term memory map, S_A is the representation of the social score, and D_A represents the order of object digestion.

The object store is a 3×3 binary matrix, in which the number of objects is coded which the agent carries at the moment. The number of 1's in the first column vector represents the number of triangles the agent has; the number of 1's in the second column vector represents the number of squares and the number of 1's in the third the number of diamonds. E.g. in equation 6.58, agent A has two triangles, no square, and one diamond.

$$O_A = \begin{bmatrix} 000 \\ 100 \\ 101 \end{bmatrix}, \tag{6.58}$$

In the short term memory map M_A the agents A have located their own position, the position of other agents and they can store location of objects which they have encountered for n time steps. After n time steps the representation of an object vanishes. The agent can learn about object positions by being in the same location as the object or by getting the information about the object location from another agent. M_A is a $10 \times 10 \times 6$ array of binary values. The 10×10 - array gives the position of the agent and the objects, the third dimension codes the object type or the identity of the agents. In equation 6.59, I will use a $3 \times 3 \times 6$ array of binary values to save space while still able to show the principle.

$$M_A = \begin{bmatrix} 000 \\ 010 \\ 000 \end{bmatrix}, \begin{bmatrix} 000 \\ 010 \\ 000 \end{bmatrix} \begin{bmatrix} 000 \\ 000 \\ 001 \end{bmatrix} \begin{bmatrix} 000 \\ 000 \\ 000 \end{bmatrix}, \begin{bmatrix} 001 \\ 000 \\ 000 \end{bmatrix}, \begin{bmatrix} 000 \\ 000 \\ 000 \end{bmatrix} \tag{6.59}$$

Here, the left most matrix codes the position of the agent in a 3×3 space. The next matrix codes the position of the second agent, which, as it can be seen is on the same position as the agent A. The position of the third agent is coded in the third matrix. The next matrix codes the position of triangles - there is no triangle position known to agent

A. The next matrix, coding positions of squares, shows that there is a square in the upper right corner. The final matrix, coding positions of diamonds, shows no diamond.

The social score Soc_{A_i} of an agent A_i is coded by a set $\{soc_{A_i}(A_1)...soc_{A_i}(A_n)\}$, where $A_1...A_n$ are the other agents in the game. Every element $soc_{A_i}(A_j)$ of the set Soc_{A_i} is represented by a pair of natural numbers $< s_{ij}, s_{ji} >$ with values between 0 and 9 with s_{ij} representing the history of j's social behavior towards i and s_{ji} represents the history of i's social behavior towards j

The representation of the digestion order, D, is given by a 3×3 matrix in which the first column vector represents the object which is needed at the next digestion step, the second column represents the one needed after that and the third the one after that. Within the column, a 1 in the first row represents a triangle, a 1 in the second row a square and a 1 in the final row represents a diamond.

$$D = \begin{bmatrix} 100 \\ 010 \\ 001 \end{bmatrix} \tag{6.60}$$

D is changed automatically when objects are digested. You can think about it as a form of appetite or hunger, thirst, freezing etc. - some form of natural body signals.

Utterance Effects

Since we are using only pre-programmed agents, the conventional utterance effects need to be defined. The other types of speech acts introduced in this chapter have made the state quite complex. Therefore, utterance effects in the current model are far more complex than in previous models. Effects have to be defined in terms of agent states, i.e. context and effect states are based on the state of the agent $A =< O_A, M_A, S_A, D_A >$.

Request A request is made when the speaker A_1 desires an object o that is in the possesssion of another agent A_2. If the agent is willing to cooperate he drops the object and the position of the object is then indicated in M_{A_1}. Nothing happens if the addressee does not cooperate.

Declaratives For declaratives, agents *name* the objects which are at their current location to another agent. An agent can do this instead of picking the object up or even after dropping an object. The effect for agent A_1 is that his social status with the agent A_2 to whom he points out this object improves. This effect happens only if the speaker really has an object in his current position and the word he uses is the correct description.

The effect for the addressee A_2 is that the described object appears at the position of the speaker A_1 in his map M_{A_2}. Since the speaker has no representation of the knowledge of the addressee, he cannot know about this effect. He only speaks in the hope of improving his social status. This ignorance of the knowledge about the knowledge of other speakers is of course a major simplification. I have briefly sketched above how the framework should

be changed in the long run to model the knowledge change of the addressee. A first effort to do this is made in chapter 7.

Questions The effect of a question is an increase in knowledge. A question is answered by an assertion, and question and answer are performed and evaluated in a single time step. The question about a specific object, e.g. *triangle-quest*, is answered with *triangle-decl* if the triangle is at the position of the addressed agent and with no answer in case there is no such object. Therefore, the effect of the question is the knowledge about the specific object in case it is answered with the declarative. The mechanisms of reward are the same as with declarative utterances in general. If a question is not answered, the agent is punished. Note that the agent who does not have the object also does not answer. Therefore, in a more realistic extension of the present language another utterance content corresponding to English *no* should be introduced.

Utterance Selection

With this model, agents can select the correct and most useful imperative, interrogative, or declarative act.

Requests Having certain objects, especially the objects which currently have a high value (objects that are needed) increases the value of the state of the agent. Requesting such an object results in having such an object and, therefore, requests will be selected because they increase the total value of the speaker's possessions, and thus the value of the state he is in.

Declaratives Pointing out objects *voluntarily* or answering questions increases the social score, which increases the value of a state. If requests do not result in the addressee giving the object away, an agent can improve his social score by pointing out objects to him.

On the other hand, an agent has to decide in which cases he will give the information about the current object to another agent, and in which cases he won't. He needs to calculate with which agents it is good to increase his social score, and also which objects he might need soon himself. For such a decision to make sense, the agent also needs to include D and S into his considerations.

Questions Questions should be asked if objects of a specific type are needed and their location is not known or is too far away. They result in an increase of knowledge. The value of this knowledge has to be dependent on D as well, because knowing about objects which are needed soon has a higher value than knowing about objects that are not needed immediately.

Suppose an object of a certain category is needed and no object of this category is indicated in M. Then the (probabilistic) forward model should indicate the possibility of a state in which the object is at the position of the addressee and this state should have a very high value.

6.5 Speech Acts and Reference

So far, this chapter was only a theoretical description of how different speech acts can be modeled in the general theoretical framework using forward models and value functions. Of course, it would be nice to have a working simulation, but there is a good reason not to implement the framework as far as it is developed so far. The reason are the speech act holophrases used in our model. Apparently children never use them. They learn at a very early stage of language acquisition, usually before they even speak their first words, to use different intonational patterns in ways comparable to the intonational patterns coding different speech acts. Then, when they use their first real words, they can use them with these different intonational patterns to perform different speech acts with the same content word, i.e. they can ask about an object, request it, or just point it out (Tomasello 2003). Although these utterances of children are also called holophrases, in contrast to the holophrases in my model so far, they code the information about the speech act and the referential information in different components of the speech signal.

To account for this we would need a language with some form of speech acts markers (corresponding to the English grammatical rules and intonational patterns coding the basic speech acts). This requires significant changes to the model. The value function does not need to be changed, since it does not make any difference whether the states are accomplished by compositional or non-compositional utterances. But fundamental changes need to be made to the internal model. How such an internal model is constructed mathematically and implemented as a computer program will be explained in chapter 7.

Chapter 7

Compositionality

7.1 Definition

Humans can produce and understand utterances they have never heard before. They have this ability, because they can derive the meaning of a complex utterance from the meanings of its parts. This property of language is called *compositionality*.

There is considerable disagreement among researchers over the exact nature of compositionality. In this thesis, however, I do not want to involve myself in this discussion. Since no one would doubt that some degree of compositionality is one of the single most important features that makes language the powerful tool which it is, any neural theory of meaning needs to deal with it. In this chapter, I will show how my theoretical framework can be extended to deal with a compositional language.

7.2 A Simple Extension of the Architecture

7.2.1 Overview

So far, the major theoretical framework of this thesis (using value functions and internal models for goal-directed communication) was only capable of processing *one word holophrases*. In this chapter, I will extend this theory so that it can also deal with simple compositional languages. To do this I will take up the theory of conceptual representations developed in chapter 4 and integrate in into the theory of goal-directed communication.

Before I do this, I will present a very simple extension of the cognitive architecture used in the simulations presented in chapter 5. In this extension, I use requests consisting of a color adjective and a noun (a complex noun phrase), instead of single nouns (as in chapter 5). To process these slightly more complex utterances, the linguistic input vector now represents combinations of two possible words instead of simple words. Although, after superficial observation, the model seemed to be able to deal with this primitive form of compositionality, a simple compositionality test (explained below) demonstrated that the model was not *really* able to grasp the compositional nature of this language.

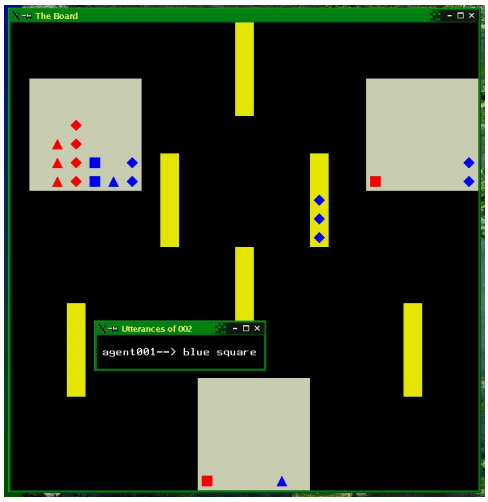

Figure 7.1: This is the more complex environment which is required to allow the simple form of compositional utterances to make sense.

In reaction to the failure of this first simple approach, I will present a more complex forward model, which uses an additional layer of semantic features extracted from the utterance. This second approach turns out to be able to handle this primitive form of compositionality exemplified in this language.

7.2.2 The Simulation

Environment

The language I will use in the first simulation of this chapter has the following simple grammar:

$S \rightarrow A\ N$
$A \rightarrow blue \mid red$
$N \rightarrow triangle \mid square \mid diamond$

The acquisition of this language was simulated in an environment similar to the one used in the simulations of chapter 5. For the current simulation, however, this environment needed to be changed, because even a very simple compositional language does not make sense, if the world does not offer the possibility to use compositional expressions. In particular, it would not make much sense to use an expressions such as *red triangle*, if all triangles were red. Therefore, there is a need to have shapes of at least two colors. Hence, I created an environment in which there are three objects types (triangle, square, and diamond) of two colors (red and blue). The environment looks as in figure 7.2.

The goal of the agents - defined by the reward function - was the same as in the

simulations of chapter 5. Also, the only possible speech act type were requests as in chapter 5, i.e. every utterance describing an object of a certain type was to be interpreted as the request for this object.

To deal with the two-word utterances I modified the value function and forward model accordingly. The input argument of the value function was adapted so that it could learn to compute a value for the more complex environment. For the forward model it was not sufficient to adapt the state-argument to the more complex environment, the input structure for the utterance had to be changed as well. In addition to a vector for the shape word (up to now this was a 3-dimensional binary vector to code one of the three shape words), another 2-dimensional vector is used to code the color word.

Appearent Results

Training the value function with the two-word utterances took about as many time steps as with the one word holophrases. After the training, the mapping of world states into values enabled the agent to select the best actions in all situations.

The training of the modified forward model did not take longer than the training of the one-word forward model. After training, the mapping of world states and two-word utterances into the next world state was 100% accurate.

This superficial description of the model's performance seems to suggest that it was actually able to handle the compositional language. However, to be able to learn the compositional dimension of a language it is not sufficient that complex utterances can be produced or understood. The necessary capability is that the meaning of a complex utterance can be derived from the meaning of its parts. Therefore, I designed a simple *test of compositionality* and I applied it to this simple extension of the model.

7.2.3 A Test of Compositionality

The compositionality test which I will out does not test whether a language is compositional, but whether a system can learn a language, given that this language is compositional.

To test whether a model can derive the meaning of a complex utterance from the meaning of its parts, it needs to be shown whether the model can understand or produce utterances it has never experienced before. Of course, it needs to be exposed *sufficiently* to the meaning of the atomic parts of the language. It is, however, already a crucial question how often and in what combinations these atomic parts of the language need to be experienced by the model.

I will leave the question about frequency unanswered for the time being since this is an empirical question which should be tested in experiments with children. In terms of modelling, what needs to be found is the maximally high learning rate which still allows the network to generalize. To the question of which parts and their (two-word) combinations, on the other hand, I will try to give a tentative answer. To understand the meaning of

the following utterances, described as ordered pairs in the set U_1 (equation 7.2), it is not sufficient for an agent to experience the use of E_1 in equation 7.2.

$$E_1 = \{< a_1, b_1 >, < a_2, b_2 >\} \tag{7.1}$$
$$U_1 = \{< a_1, b_2 >, < a_2, b_1 >\} \tag{7.2}$$

The reason is that the agent has no way of knowing which part of the effect of an utterance (such as $< a_1, b_1 >$) can be attributed to which lexical item. To derive the meaning of a lexical item from its usage in a complex utterance, the system either needs to know the meaning of the other parts or it needs to encounter the item in at least two different contexts[1]. For example, if it is exposed to both $< a_1, b_1 >$ and $< a_1, b_2 >$ then the system can learn what part of the meaning is attributed to a_1 and therefore might be able to attribute which part should be attributed to b_1 and b_2.

In this light, the test situation can be more correctly described as that where a system experiences the meaning of set of sequences $E = \{< a_1, b_1 >, < a_1, b_2 >, < a_2, b_1 >\}$ but never encounteres the sequence $U = \{< a_1, b_2 >\}$. It is nevertheless able to compute its meaning due the meaning representation it has developed from processing the first sequence.

To apply this test, certain combinations (in our case $< a_1, b_2 >$) must not be in the training set. If the model is nevertheless able to understand or even produce such an utterance, then that shows that it is able to handle compositionality.

7.2.4 Application of the Test

Following the test procedure just described, I tested the extended model by blocking the rule-based dialogue systems to produce (and, therefore, also to react to) the utterance: $< red, square >$. The language learner encountered the words *red* and *square* only in the utterances: $< red, triangle >$, $< red, diamond >$, and $< blue, square >$.

After training, the other agents were again allowed to produce the utterance $< red, square >$. However, the learning agent was not able to react to this utterance in the correct way. Also, no production of this utterance could be observed. In other words the architecture did not pass the test. The model was not able to produce or comprehend this utterance by deriving the meaning from its parts.

To check whether this problem is generally unsolvable or whether it is just a problem of learning, I performed a mathematical analysis of the weights, i.e. I checked whether the weights could be set in any way that would solve this problem. I did this by stating the conditions on the weights in terms of a set of mathematical inequalities. It turned out that there was no solution for these inequalities, so the problem cannot be solved by this model.

[1]This condition has to hold only for so called associativist theories of meaning acquisition, in which the referent of a word is determined by properties in the world with which it is correlated most often. The framework I am describing at the moment is such an associativist framework. Very soon in this chapter, I will move more towards what might be called an *intentionalist* approach to meaning acquisition, in which these conditions do longer have to hold.

7.3 Using Conceptual Binding

In chapter 4, I have already explained what I assume as the basic organization of concepts in the brain. One of the main features of this organization was that complex concepts are actually *composed* from more basic concepts or features. If this is the case, then compositionality exists already on the conceptual level. So a red square is likely to be represented as a combination of the feature *red* and the feature *square*.

Given such an organization, it is straightforward to add a conceptual layer to the model in which features such as *red* and *square* are represented. These basic features are assumed to be either inherited or learned before their respective linguistic labels.

The essential step for the learning model is then to associate the labels with the concepts. This pushes the essential problem of compositionality to the level of concepts (where I believe this issue actually belongs).

If the association of features with labels is regarded as the main learning step the model has to perform in a simulation environment such as the one in the previous section, then there are two more tasks which the model has to solve. The first one is the problem of object identification, i.e. how does the model know for which object it actually has to find the features it can associate with the respective labels. The second is the problem of what action actually has to be performed with the object if it is identified. This second problem can be generalized to the problem of *speech act compositionality* and I will discuss it in detail in the remaining sections of this chapter. The first problem, I believe, is the problem of understanding intentions and directing attention respectively.

Children have the (probably instinctive) capability of following the gaze of a parent. Eventually, this capacity enables them to develop a sense of shared attention. Shared attention appears already before language in children and it is regarded as an important prerequisite of word acquisition. Attention also bridges the dispute between associativist and intentionalist accounts of language acquisition. The children have to solve the problem which object they should attend to (intentionalist account), but then the features of this object have to be associated with the verbal label (associativist account).

Bottom-up control mechanisms of attention make us humans (including language learning children) attend to things which are prominent for some reasons and the mechanisms of shared attention make children attend to the objects their caretakers attend to. It is, therefore, reasonable to assume that the language learning agent in my model will attend to the object which actually changes its owner, since (i) it is prominent by way of changing position and (ii) it is attended to by both, the giving and the receiving agent.

7.3.1 Simulation

Taking these two ideas (label-feature-association and attention control mechanisms) into account, I designed a complex forward model which works in the following way: In a first step, a feature representation (color, form, movement) is extracted from the utterance. This feature representation is then used to identify and locate the relevant object in space. Attention is directed on this particular object and a change of its location is predicted.

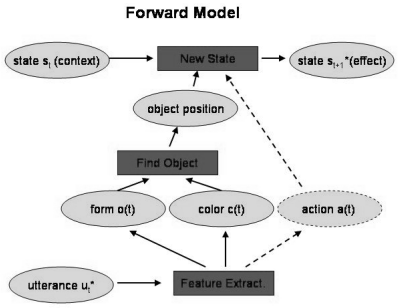

Figure 7.2: This is the compositional forward model. Like the simple forward model it maps a state (context) and an utterance on the next state.

7.3.2 Results and Discussion

I have shown that agents can learn to use compositional utterances of the present fragment to express their desires. They also learn to understand the desires of other agents expressed by compositional utterances. Learning takes significantly longer than in the simulation with the non-compositional model, but in both studies training times have not been used to evaluate the models[2]. After training the agents show correct reactions when they are addressed and produce the appropriate utterance in every situation.

Also, agents using the *compositional* forward model, were able to understand and produce sentences which they have never heard before, i.e. they derive the meaning of the complex utterance from the meaning of its parts.

Whether this was achieved by having an *innate* representation of features, or whether it was mainly because of the extra layer cannot be said at the moment. An interesting experiment to test that would be to use a multi-layer network (such as a multi-layer perceptron trained with backpropagation) and see whether it can aquire the necessary feature representation in the extra layer. The reason why I felt justified in choosing to prespecify the features in the model of this section is that there is sufficient evidence showing that certain basic features are indeed innate.

This can be achieved simply having a third layer in the network or requires in addition preprogramming of innate feature is something that must be left open.

[2]I have not attempted to fit the training times to empirical language acquisition data. However, it seems that in both cases, training takes significantly longer than in the language acquisition of children. This, however, might change in models in which the referential aspect and the usage aspect of meaning are treated independed from each other, as they are in the theoretical models that are introduced below.

7.4 Unifying Theoretical Strands

In this section, I will finally take up the theoretical strand developed in chapter 4 and use it to formalize the insights gained about the relation between utterances, current states and desired states.

Using the example of the previous section, an utterance such as $< green, triangle >$ can be chosen in a *current* state C_{A_1} such as the one described in equation 7.3).

$$C_{A_1} = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} triangle\_f : & < \{w_1\}, \{l_1\} > \\ speaker\_f : & < \{w_2\}, \{\} > \\ addressee\_f : & < \{w_3\}, \{\} > \end{array} \right] \\ VCA : \left[\begin{array}{ll} green\_f : & < \{w_1\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_3, w_1\}, \{\} > \\ position[x_2, y_2] : & < \{w_2\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'triangle' : & < \{\}, \{l_1\} > \\ 'green' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.3}$$

The desired state D_{A_1} of agent A_1 might be such as described in equation 7.4. The only difference between C_{A_1} and D_{A_1} is that the green triangle changed its position from the position of the addressee to the position of the speaker.

$$D_{A_1} = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} triangle\_f : & < \{w_1\}, \{l_1\} > \\ speaker\_f : & < \{w_2\}, \{\} > \\ addressee\_f : & < \{w_3\}, \{\} > \end{array} \right] \\ VCA : \left[\begin{array}{ll} green\_f : & < \{w_1\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_3\}, \{\} > \\ position[x_2, y_2] : & < \{w_2, w_1\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'triangle' : & < \{\}, \{l_1\} > \\ 'green' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.4}$$

The core of a request for objects is then the pattern described in equations 7.5 and 7.6.

$$C_{A_1} = \left\{ \begin{array}{l} BrainAreas\_1 - n \left[\begin{array}{ll} speaker\_f : & < \{w_1\}, \{\} > \\ addressee\_f : & < \{w_2\}, \{\} > \\ identified\_object\_f : & < \{w_3\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} speaker\_position : & < \{w_1\}, \{\} > \\ addressee\_position : & < \{w_2, w_3\}, \{\} > \end{array} \right] \end{array} \right. \tag{7.5}$$

$$D_{A_1} = \left\{ \begin{array}{l} BrainAreas\_1 - n \left[\begin{array}{ll} speaker\_f : & < \{w_1\}, \{\} > \\ addressee\_f : & < \{w_2\}, \{\} > \\ identified\_object\_f : & < \{w_3\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} speaker\_position : & < \{w_1, w_3\}, \{\} > \\ addressee\_position : & < \{w_2\}, \{\} > \end{array} \right] \end{array} \right. \tag{7.6}$$

Note that an utterance is defined by speaker, addressee, and content, i.e. the representation of speaker and addressee is also triggered by the utterance. The utterance has the form described in 7.7

$$speaker \rightarrow addressee : w_1, w_2, ...w_n \tag{7.7}$$

At this point, all utterances are requests for objects. To extend the types of speech acts possible within this new compositional framework, I also need some way to indicate the speech act type. Requests for an object will be indicated by the marker *obj_req* at the end of the word sequence. In English such requests are usually expressed by the syntactical structure of the sentence (V,S,O), or even with an explicit *give me ...* or polite variations of this (such as indirect speech acts).

With this speech act marker, we get an utterance of the form described in equation 7.8

$$speaker \rightarrow addressee : w_1, w_2, ...w_n, obj\_req \tag{7.8}$$

Now it is possible to introduce other speech act markers and examine systematically how they contribute to the context conditions and the effect of the utterance. However, before I do that, I need to introduce some more structure to allow the analysis of the compositional effects of *assertions*. In particular, I need to introduce a way of representing the mental states of other agents. I will then tie up the loose end of chapter 4, where I postponed an analysis of the process of selecting topic and comment predications in production.

7.5 Compositionality in Assertions

7.5.1 Theory of Mind

In section 7.5.3, I have discussed the variety and the different levels of goals that agents can persue by using an assertion. But all these goals depend on the basic goal of changing the conceptual activation in the knowledge base of the addressee. To model this central goal of an assertion, both the current state C_{A_1} and the desired state D_{A_1} of an agent A_1 have to include some way of representing the belief state of another agent A_2.

Unfortunately the empirical data needed to base such a theory on does not offer a solid foundation. There are some functional imaging studies about tasks involving theory of mind (Hill and Frith 2003) which give us some idea about which brain areas might be involved in the representation and processing of other people's mental states. However, close to nothing is known about the representations of a person's knowledge about other people's mental states in terms of neurons. Fortunately, there is some behavioural data which allows us to speculate on the nature of this representation.

Apparently children at first believe that all humans share the same knowledge. They have trouble understanding that another person might not know something which they know. But it appears that a child can realize that another persons might know s.th. that

the child does not know, because they start asking questions, such as *What's that?* and *Where go?*, at a very early stage - clearly before they are able to pass the Sally-Ann Test. Do such questions already represent a conscious realization that other persons have knowledge which they do not have, where it is only later that children grasp that different persons might have different knowledge? How do they know the knowledge of other people differs from their own[3]? Of course, language is one way to make people (including children) aware of differences in knowledge, but the ability to contrast my knowledge with other people's knowledge must be independent of language. This is illustrated by our ability to hide. It is possible for me (and for children, too) to hide myself in a way that person A_2 cannot see me. To do this I need to take the perspective of that person and calculate the areas that are hidden from his sight.

Since there is evidence that the child's first attitude is that all knowledge is shared by all people and that it becomes only later aware of the differences, it is justified to assume that the brain represents knowledge about other person's knowledge as a different from our own knowledge. One way in which such difference could be represented is as follows. By default, an agent A_1 will assume that the knowledge of any other agent A_2 is identical to A_1's own knowledge. Only if there is a reason to assume that A_2's knowledge differs in certain respects, then this difference will be represented. This also makes it unnecessary for the brain of A_1 to represent the same knowledge items several times - once as the knowledge item of A_1 and then again for A_2 and again for every other agent to whom A_1 attributes this bit of knowledge.

There are two main reasons to infer that another agent's knowledge differ one's own: (i) language, i.e. someone telling us s.th. that is not in accord with our own beliefs and (ii) something I will call *simulation of experience*, i.e. I simulate the experience of the other person in my own brain to derive the knowledge this person has. To do this, I need to simulate the perspective of another person and infer what kind of knowledge the other person has gathered. Here are some examples:

- If someone was already in my room, I know that he has *seen* my room. Therefore, I know that the person knows that I have a piano.
- If I have tasted the food the waiter has just brought in, and my dinner companion did not taste it yet, I know that she does not have knowledge of its taste.
- If I watch out of the window and see a butterfly while I am on the phone with my friend, I know that he does not see this butterfly.
- I assume that a certain person does not have my phone number, because I have never given it to him.

Further, the problem of representing knowledge of another agent is in a sense similar to the problem of representing an action of another agent: It needs to be represented as the

[3]Maybe the first step is to realize and become conscious of the fact that they themselves do not have all knowledge.

knowledge of that particular agent - in other words, the representation of the knowledge needs to include some form of link to a person.

But how is a person represented in the brain? It seems justified to assume that the representation of a person involves several levels of representation, among them at least the following: (i) the person's perceptual features (e.g. visual information about the person), (ii) memories of episodes that involved this person (events involving the person which we remember), and (iii) what we know about the mental state of the person (e.g. what we know about the person's knowledge). It is unlikely that we link our representation of the knowledge of an agent with everything we know about the agent - just in order to indicate whose knowledge it is. This would be to much information which might impede the brain in its effort to keep apart the features of the many concepts and episodes it has memorized. It is, however, not unlikely that the brain develops some form of higher level representation of a person that abstracts from detailed information such as particular episodic memories and perceptual memories about the appearance of that person. I will assume such a representation and I will call it a *tag* - and if this tag is used to mark the knowledge of a person, I will call it a *knowledge tag*.

Also, it seems reasonable to assume that representations of concepts such as categories, individuals (events, agents, etc.) are shared between the agent's own knowledge and her knowledge about the mental states of other agents. Using the knowledge tag, such a sharing of concepts is possible.

Like all other conceptual representations, the representation of the mental states of another agent can be either in working memory or in long term memory with the usual interaction between the two types of memory.

7.5.2 A First Speculative Representation

In this section I will use the idea of *knowledge tags* to enrich the theoretical framework with a way of representing the mental states of other agents. So far, the conceptual knowledge of an agent A_1 is represented as in equation 7.9.

$$C_{A_1}(t) = \begin{array}{ll} Modality\_1: & \left[\begin{array}{ll} feature\_1: & <\{l_i...\},\{w_k...\}> \\ ... & ... \\ feature\_n: & <\{l_l...\},\{w_m..\}> \end{array}\right] \\ & ... \\ Modality\_z: & \left[\begin{array}{ll} feature\_1: & <\{l_o...\},\{w_p...\}> \\ ... & ... \\ feature\_n: & <\{l_q...\},\{w_r...\}> \end{array}\right] \end{array} \tag{7.9}$$

Now suppose that an agent A_1 knows that a certain banana is sweet while another agent A_2 does not know that. A_1's knowledge could be represented as in equation 7.10.

$$C_{A_1}(t) = \begin{array}{l} Sight: \left[\; banana\_f: \; < \{w_1\}, \{\} > \; \right] \\ Taste: \left[\; sweet: \; < \{w_1\}, \{\} > \; \right] \\ Location: \left[\; position[x,y]: \; < \{w_1\}, \{\} > \; \right] \end{array} \tag{7.10}$$

In contrast, A_2 does not know that the banana is sweet. His knowledge is represented in equation 7.11.

$$C_{A_2}(t) = \begin{array}{l} Sight: \left[\; banana\_f: \; < \{w_1\}, \{\} > \; \right] \\ Taste: \left[\; sweet: \; < \{\}, \{\} > \; \right] \\ Location: \left[\; position[x,y]: \; < \{w_1\}, \{\} > \; \right] \end{array} \tag{7.11}$$

To represent A_1's knowledge of A_2's mental state, I will use the *knowledge tags* introduced above. It is, however, not clear where in the brain those tags might be represented. I will, therefore, put the tags into a hypothetical area that I will call *knowledge*. To which area in the real brain this functional area corresponds is subject to experimental research.

Using the tag, the knowledge of other agents A_2 is represented by linking a conceptual configuration to the tag of A_2. If the tag is not active (as in equation 7.12), the represented knowledge is that of the representing agent himself (A_1). Equation 7.12 describes a state of A_1 in which he knows that a particular banana is sweet.

$$C_{A_1}(t) = \begin{array}{l} Sight: \left[\; banana\_f: \; < \{w_1\}, \{\} > \; \right] \\ Taste: \left[\; sweet: \; < \{w_1\}, \{\} > \; \right] \\ Location: \left[\; position[x,y]: \; < \{w_1\}, \{\} > \; \right] \\ Knowledge: \left[\; A_2: \; < \{\}, \{\} > \; \right] \end{array} \tag{7.12}$$

However, if the tag is active, it generates the contrasting information about the concepts leading to the representation of the knowledge of A_2. Equation 7.13 describes a state of A_1 in which he knows that A_2 does *not* know that this particular banana is sweet.

$$C_{A_1}(t+1) = \begin{array}{l} Sight: \left[\; banana\_f: \; < \{w_1\}, \{\} > \; \right] \\ Taste: \left[\; sweet: \; < \{\}, \{\} > \; \right] \\ Location: \left[\; position[x,y]: \; < \{w_1\}, \{\} > \; \right] \\ Knowledge: \left[\; A_2: \; < \{w_1\}, \{\} > \; \right] \end{array} \tag{7.13}$$

7.5.3 Transmission of Information

Within this enriched framework I am not only able to describe A_1's representation of A_2's knowledge (equation 7.13), but also the knowledge that A_1 *wants* A_2 to have. To do this, I simply use the tags also in the representation of the desired states D (equation 7.14).

$$D_{A_1}(t) = \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{w_1\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x, y] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A\_2 : & < \{w_1\}, \{\} > \end{array} \right] \end{array} \tag{7.14}$$

Since the outline of my theoretical framework of goal-directed communication in chapter 3, it has been the central assumption of my theory that utterances are selected by means of an internal model, i.e. either by a forward or an inverse model, or by both. In chapter 5, I favoured the assumption that utterances are selected with a forward model, by simulating the effects of all possible utterances and choosing the one with the effects which are desired the most. While this approach has its advantages (discussed in chapter 5), it also has a major problem. Because of the richness of language it is absolutely impossible for a speaker to go through *all possible* utterances or even a large subset of them. Given, however, that the language producing agent knows what the desired state is, an inverse model can be used to select the utterance. In what follows, I will sketch an inverse model which uses several processing steps to compute an utterance u which is likely to cause the transition from C_{A_1} to D_{A_1}.

To accomplish this transition by means of language, it is, of course, essential that some of the concepts, such as *banana* and *sweet* are lexicalized, as in equations 7.15 and 7.16.

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x, y] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{w_1\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.15}$$

$$D_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{w_1\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x, y] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{w_1\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.16}$$

Given these two states, producing the utterance which will bring about the transition involves several processing steps. The first one is to find the difference between the current state and the desired state (this is probably the case for other speech acts, too). This is *processing step 1.*

$$C_{A_1} = \left\{ \begin{array}{l} Taste : \left[\begin{array}{ll} sweet : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.17}$$

$$D_{A_1} = \left\{ \; Taste : \left[\; sweet : \;\; < \{w_1\}, \{l_2\} > \; \right] \right. \tag{7.18}$$

Given that this difference is found, *processing step 2* is to identify the individual concept of which this difference is part. In this case, it is the representation bound by the working memory label w_1.

Processing step 3 is then to select one or a combination of lexicalized concepts which can be used to refer to this individual. Technically, it is the concept or concept combination which has features that in combination are part of the individual concept and of no other one. Also it has to be part of the individual concept in the knowledge representation of A_2 in the current state C_{A_1} and not in the desired state D_{A_1}. Since A_2 does not know these additional properties, they can, of course, not be used to refer to this object.

In principle there are various options here. One option that is found, it seems, in all natural languages is that of referring to the individual represented by one's individual concept by means of a proper name. I conjecture that names are associated with individual concepts in essentially the same way in which a word like *sweet* is associated with a categorical concept. For the case under discussion, however, a proper name seems an unlikely communicative vehicle. For in order that a proper name can function effectively in communication it is necessary that those who use it are members of a group within which there exists a certain amount of common knowledge regarding the entity named. Such knowledge presupposes that there has been considerable interaction within the group pertaining to this entity. Not only must there be enough members who have had an opportunity to acquire an individual concept representing the entity, but they must also have had the chance to establish that they share a concept for it. If those conditions are fulfilled, and moreover members of the group come to associate the same name (i.e. the same phonological string) with the individual concepts they have for the entity, then use of the name by one of them in the presence of another member can be expected to have the effect of triggering in that other member an activation of the individual concept with which the name is linked. In general, however, it takes time for those preconditions to be fulfilled, and it also requires an explicit act of *naming* on the part of somebody or some people within the group to get the name-individual concept association going. It is something that as a rule will happen only if the entity is long-lived and important enough to sustain this fairly complex social process. Thus, in the normal course of things a banana neither has the individual significance nor the longevity needed to become the bearer of a shared name.

A_1 will thus have to make use of another way of referring to his banana - or, in our terms, another way of activating A_2's individual concept for this banana. Among the alternative means that English has available for this purpose there are complex definite descriptions and demonstratives, in which a noun (with or without additional adjectives, relative clauses and/or prepositional phrases) is preceded by *the* or by one of the demonstratives *this* and *that*, respectively. For present purposes I won't distinguish between these three options - there is much that linguistics has to say about the differences between them, but that is not central to our present concern. Instead, I want to focus on the choice of the other part of the phrase, i.e. of the noun (with or without satellites). How is A_1 to decide on the noun

or nouns he is to use?

The choice will have to depend on what A_1 assumes A_2 knows about the individual to which he wants to draw A_2's attention. There are two cases that should be distinguished here. The first is the one where A_1 assumes - and let us suppose he assumes correctly - that A_2 already has an individual concept for the thing he wants to refer to. (For instance, A_1 may have noted that A_2 was observing him while he was tasting the banana.) In that case A_1 will have to decide on one or more categorical concepts that (i) are part of A_2's individual concept; (ii) whose activation will trigger the activation of A_2's individual concept for the individual to which A_1 means to refer; and (iii) for which A_1 and A_2 share a lexicalization (i.e. *sharing* the conviction that they associate the same word forms with those concepts).

Suppose that A_1 settles on a set of one or more such concepts and utters a definite description or a demonstrative in which the word forms for these concepts are conjoined. (Again, I ignore the question exactly what form this conjunction will take, e.g. whether any of the concept words will turn up in a relative clause and so on.) In order that this phrase (description or demonstrative) has the intended effect on B, the activation of the concepts which A_2 associates with the concept words in this phrase must trigger the activation of A_2's individual concept for the banana of which they are part. Intuitively, the likelihood that this will happen would seem to depend on two different factors, (i) whether the lexically triggered concepts play a sufficiently salient part in the network which implements A_2's individual concept, and (ii) whether they play a comparable role in other individual concepts. (In order that the right individual concept be activated by the activation of its part concept or concepts there shouldn't be too much competition from other individual concepts to which the lexically activated concept or concepts also belong.)

The lexicalized concept that A_1 will choose in our case is the one associated with the word *banana*. Processing step 3 leads to an activation of the lexicalized concept and also the word form *banana* (equations 7.19 and 7.20, working memory label w_2). This step is what I call the *topic predication* with *banana* as its *topic predicate*.

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1, w_2\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x, y] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{w_1\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{w_2\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.19}$$

$$D_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{w_1\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x, y] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{w_1\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{w_2\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.20}$$

Given this state of activation in the speaker's brain, she now has to give the addressee the new information about the individual the representation of which he has just triggered in working memory. *Processing step 4*, therefore, is to find a lexicalized concept which can activate this information. In our example, the concept is the one associated with the word *sweet* (long term memory label l_2). This step is called *comment predication* and *sweet* is the *comment-predicate*.

Since the goal of the whole utterance production is the change of the knowledge state of the addressee, the speaker also has to mark the utterance as an *assertion*, leading to an utterance of the form (syntactical issues and function words are ignored):

$$[[banana_T][sweet_C]_{assert}] \tag{7.21}$$

Note that the theoretical framework described so far explains only how an assertive speech act is produced given that the desired state of the speaker is a state in which the knowledge of the addressee is changed. It does not explain why the speaker should have such a desired state in the first place. To select a state as the desired state it needs to be accessible from the current state and the value function needs to map it onto a higher value than all other accessible states. To be mapped on a high value by the value function, the state with the changed knowledge of the addressee needs to include some features which can motivate the selection of an assertive speech act (described in detail in section). The central feature here is in the representation of the knowledge base of the addressee. The desired state needs to include a representation of the fact that the speaker has cooperated with him - a memory that has a high value since it might trigger cooperative behavior from the addressee in the future. An estimation of how important the information that the assertion will provide is for the addressee might determine his degree of gratefulness. However, such an estimation would require considerable development of the possibilities of the theory. Another factor that could determine whether the speaker would provide information to the addressee is whether he has memories of past cooperative behavior of the addressee (e.g. his sharing of important information). Such information could be stored in the current state.

7.6 Decomposing Speech Acts and Reference

As the *coda* of the theoretical development in this thesis, I would like to apply this framework to the production of three types of speech acts (assertions, questions, requests). In

the last section, I have already described the structure of the desired state D (equations 7.16) and the current state C (equation 7.15) serving as input to the inverse model which has to determine the structure of an appropriate assertion.

One possible input (in terms of C and D) to the inverse model for requests can be the one described in equations 7.22 and 7.23. The main point expressed in these equations is that in the current state the sweet banana has a different position than in the desired state (in which it has the same position as the speaker).

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \\ self\_f : & < \{w_2\}, \{\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{w_1\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_1\}, \{\} > \\ position[x_2, y_2] : & < \{w_2\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.22}$$

$$D_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \\ self\_f : & < \{w_2\}, \{\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{w_1\}, \{l_2\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{\}, \{\} > \\ position[x_2, y_2] : & < \{w_2, w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.23}$$

This input leads to the following utterance type:

$$[[banana_T][sweet_C]_{req}] \tag{7.24}$$

To compute a question, the inverse model of the speaker needs to take into account that a question might trigger several possible answers. This means that the question can lead to several possible states of the speaker (characterized by the information about the world he has received and integrated into the overall state). In the following example we have two possible states, called D^1 and D^2, described in 7.26 and 7.27. The current state is given in equation 7.25.

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{\}, \{l_2\} > \\ not\_sweet & < \{\}, \{l_3\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.25}$$

$$D^1_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{w_1\}, \{l_2\} > \\ not\_sweet & < \{\}, \{l_3\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.26}$$

$$D^2_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{\}, \{l_2\} > \\ not\_sweet & < \{w_1\}, \{l_3\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_1\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.27}$$

The utterance type this input leads to is described in equation 7.28.

$$[[banana_T][sweet_C]_{quest}] \tag{7.28}$$

Given these three input patterns and the three utterance types (characterized by the different speech acts they perform), the question is, how does the system learn which input to transform into which speech act type. Looking at the structure of the input pattern shows that in the case of assertions it is the knowledge of the addressee that needs to be changed, while it is the position of objects in the case of requests, and the speaker's knowledge about the world in the case of questions.

7.6.1 Computational Generalizations

So far, the input to the inverse model is a current state and a desired state (in the case of questions a list of desired states). Note that the state descriptions I have used to explain the nature of the representations and processes are very small elements of the very large conceptual states that we must assume to be present in the real brain. The desired states

have the same complexity as the current states. This is of course a very inefficient way of coding (especially for questions), since they are usually - apart from some important but small differences - equal to the current state. What we need instead is to specify a way of reducing the desired states to some description of differences they make to the current state (i.e. the required changes). I do this simply by specifying only those feature(s) which differ from the current state.

For the current and desired states of the previous section, the desired change (Δ) can be found below.

Assertion ($[[banana_T][sweet_C]_{assert}]$) :

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & <\{w_1\},\{l_1\}> \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & <\{\},\{l_2\}> \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x,y] : & <\{w_1\},\{\}> \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & <\{w_1\},\{\}> \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & <\{\},\{l_1\}> \\ 'sweet' : & <\{\},\{l_2\}> \end{array} \right] \end{array} \right. \tag{7.29}$$

$$\Delta = \left\{ Taste : \left[\begin{array}{ll} sweet : & <\{w_1\},\{l_2\}> \end{array} \right] \right. \tag{7.30}$$

Request ($[[banana_T][sweet_C]_{req}]$) :

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} banana\_f : & <\{w_1\},\{l_1\}> \\ self\_f : & <\{w_2\},\{\}> \end{array} \right] \\ Taste : \left[\begin{array}{ll} sweet : & <\{w_1\},\{l_2\}> \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1,y_1] : & <\{w_1\},\{\}> \\ position[x_2,y_2] : & <\{w_2\},\{\}> \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & <\{\},\{\}> \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & <\{\},\{l_1\}> \\ 'sweet' : & <\{\},\{l_2\}> \end{array} \right] \end{array} \right. \tag{7.31}$$

$$\Delta = \left\{ Location : \left[\begin{array}{ll} position[x_1,y_1] : & <\{\},\{\}> \\ position[x_2,y_2] : & <\{w_2,w_1\},\{\}> \end{array} \right] \right. \tag{7.32}$$

Question ($[[banana_T][sweet_C]_{quest}]$):

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\; banana\_f : \; < \{w_1\}, \{l_1\} > \; \right] \\ Taste : \left[\begin{array}{ll} sweet : & < \{\}, \{l_2\} > \\ not\_sweet & < \{\}, \{l_3\} > \end{array} \right] \\ Location : \left[\; position[x_1, y_1] : \; < \{w_1\}, \{\} > \; \right] \\ Knowledge : \left[\; A_2 : \; < \{\}, \{\} > \; \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.33}$$

$$\Delta^1 = \left\{ \; Taste : \left[\; sweet : \; < \{w_1\}, \{l_2\} > \; \right] \right. \tag{7.34}$$

$$\Delta^2 = \left\{ \; Taste : \left[\; not\_sweet \; < \{w_1\}, \{l_3\} > \; \right] \right. \tag{7.35}$$

The problem which still needs to be addressed is whether there are general processing steps leading from the current state C and the desired change Δ to an utterance, irrespective of what speech act type this utterance should belong to.

Processing step 1 is to determine the speech act type. For the examples discussed above the following algorithm can be used:

- If some *Knowledge-feature* is activated in the current state ⇒ assertion.
- If Δ contains *Location-features* ⇒ request.
- If there are several Δs ⇒ question.

This is, of course, a first approximation that only works for the specific speech acts and utterances discussed so far. I will point out some problems with this approach, as well as possible solutions to overcome these problems at the end of this section.

Processing step 2 can be analogous to the one in section 7.5.3: identification of the individual to which the changing feature belongs. The same holds for *processing step 3*, viz. the topic predication.

Using *processing step 4* from section 7.5.3 I could handle assertions in the way specified there. To deal with questions the general algorithm can be changed to repeat processing steps 2, 3 and 4 for every Δ. Instead of the question specified in equation 7.28, this would produce a sequence such as in equation 7.36.

$$[[banana_T][sweet_C][banana_T][not\_sweet_C]_{quest}] \tag{7.36}$$

Such a sequence is a meaningful sequence, since it is an accurate description of the information state of the person asking the question. I will therefore assume it to be the produced utterance form of a question.

For requests, Δ contains two features, one new feature activation and one feature *deactivation*, i.e. the object which was activated at the location of the addressee is no longer active there (this is the deactivation) and is not active at the speaker's location

(this is the new activation). Since the main goal of the request is to get the object to the position of the speaker and it is a necessary consequence that the object will no longer be in the position of the addressee, the algorithm for request can be designed in such a way that the newly *activated* feature of Δ determines the comment predicate. In this case the feature is the speaker's position, for which generally the word *here* is used. We get the following utterance, instead of the one in equation 7.24.

$$[[banana_T][here_C]_{req}] \tag{7.37}$$

To conclude this section, let me point out one important problem concerning processing step 1 (the determination of the speech act type) concerning the difference between questions and requests. While assertions have a pretty clear criteria of selection (change of knowledge of the addressee) that distinguishes their desired state from the desired state of other speech acts, in the current framework, questions and request both change the representation of reality. While requests change the representation of reality because they (might) lead to an event that actually changes reality (and, therefore, its representation), questions change the representation of reality because the speaker gets knowledge about something he does not know before. The only difference in the representation so far is that questions have (at least) two alternative Δs.

These two Δs are not a problem if utterances are selected with a probabilistic forward model as described in section 6.2, since such a model can compute probability values for more than one possible outcome. However, for an inverse model this means that more than one state has to be selected by the value function, and, at this moment, it is not clear how that can be possible.

Assuming, however, that this problem can be solved, then it might also be possible that an agent desires the world to be in one of certain alternative states, e.g. he might wants a certain object, say his jacket, to be on one of his chairs, i.e. either on his chair number 1 or his chair number 2. How does the configuration of desired state and current state then differ from asking whether his jacket is on chair 1 or on chair 2? The crucial difference seems to be that in the case of the question the agent already knows that his jacket is on either on chair 1 or on chair 2, while in the case of the request, he probably know that the jacket is neither on chair 1 nor on chair 2. Both facts can be represented in the theoretical framework as it is. In other words, if the current state and the desired state contradict each other, then the selected speech act is a request, while if they don't the selected speech act is a question.

It might be that in the long run the solution of this problem needs to use some sort of meta-representation of knowledge, i.e. an agent's knowledge about his knowledge. But for the time being, it seems that the distinction can already be made to a sufficient degree by using the existing first-level representations of reality.

There is a further problem concerning questions: How does the speaker know who the question should be addressed to. In case of requests for objects, the addressee simply is the person with the object of desire and in the case of assertions, it is the person, whose knowledge the speaker wishes to change. In the case of questions the current and desired

state only represent the knowledge that is desired, but not the potential source of this knowledge. To determine the addressee, a representation of the possible source of this knowledge has to be included in the current state using the existing framework for the representation of the knowledge of other agents. However, such a representation needs to use higher level concepts such as *location of*, which still need to be introduced into the theory.

7.7 Final Theoretical Remarks

7.7.1 Determining Desired State

As pointed out in chapter 5, inverse models have the disadvantage that the desired state (or the desired change) needs to be found before the utterance selection process can take place. Since it is not possible to compute the value of all possible states at every point in time, it is necessary to find a way to compute the most salient and potentially most desired possible changes instead of all possible changes.

One way to do that is by associative learning. The agents need to learn which current (context) states are usually followed by which future states in case of no action. Then they can learn by trial and error or by imitation what can be done to prevent this natural flow of events. Also states can be associated with certain desires that in the past have been fulfilled in the context of these states. These desires would then trigger the activation of the desired states.

For forward models the problem is that not all possible actions can be checked, so either the current state or the desired state needs to trigger some constraint on action selection. In fact, as Hommel (1998) has shown, the desired state actually triggers the activation of the action which has led to it in the past. This is evidence that indeed associations are made between states and actions. So they could be used to constrain the possible actions.

These considerations need to be taken into account before experiments can be set up to test hypotheses about the nature of the utterance selection process.

7.7.2 Syntax and Sequence Processing

Currently, the utterances represented are of a fixed length and with a fixed structure which has fixed slots for the possible words: There is a slot for the first word (the topic predicate), another slot for the second word (the comment predicate), and a final slot for the speech act marker. However, in human languages, the number of words and the syntactical structure and the semantic types in an utterance are much more flexible. Therefore, to bring the language which these types of model are able to handle closer to real natural language, the structure of the utterances must not be predefined by the programmer. Instead, the agents should be able to learn which sequences produce which effects in the same way they learn the effects of words and word combination. A model with a vocabulary consisting of

red, *blue*, *square*, *diamond*, *triangle*, *give*, and *eat* should be able to produce and to receive utterances such as (1) and (2).

(1) red red red red red

(2) square eat give square

Only through experience (imitation as well as trial-and-error) the model should learn which sequences are an effective means of achieving goals and which are not. This also sheds a new light on the function of syntactic structure. Either the sequence of words determines the meaning (*John kisses Mary.* vs. *Mary kisses John* or it facilitates the processing of the utterance, i.e. it is a conventionalized trigger for standardized (and therefore faster) processing.

One possible way to accomplish the processing and generation of free sequences might be to use a simple recurrent network (Elman 1990) as an internal model. Such networks are able to generate some form of internal representation of syntactic and semantic categories.

Another and probably more interesting way would be to introduce a verbal working memory (VWM) as a part of the agent's internal state. A VWM has to contain an ordered list of *word forms*. With such a device the production and perception of utterances could be restricted to one single word, which would then be processed in the context of the previous sequence. The effect of some words would simply be to change the state of the VWM.

Of course, words will not stay in the VWM for very long. For one, there are constraints on memory - only a specific number of items can be stored and also items can be stored only for a limited time. But also, it is likely that items will not be held for long in VWM if they have already been processed and are no longer needed.

Take the previous case of the assertion that a particular banana is sweet, given that the current state is as in equation 7.38 with the desired state given in equation 7.39.

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\; banana\_f : \; < \{w_1\}, \{l_1\} > \; \right] \\ Taste : \left[\; sweet : \; < \{\}, \{l_2\} > \; \right] \\ Location : \left[\; position[x, y] : \; < \{w_1\}, \{\} > \; \right] \\ Knowledge : \left[\; A_2 : \; < \{w_1\}, \{\} > \; \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \\ VWM : [] \end{array} \right. \tag{7.38}$$

$$\Delta = \left\{ \; Taste : \left[\; sweet : \; < \{w_1\}, \{l_2\} > \; \right] \right. \tag{7.39}$$

The first step, as we know, is to determine the speech act type, but nothing is produced at that stage. After the third processing step the topic-predicate *banana* is produced with the goal of transmitting this word into the VWM of the addressee. In equation 7.40 the

changed representation of the mental state of the addressee shows an activation of the word *banana* in VWM as the first word of a sequence.

$$C_{A_1} = \left\{ \begin{array}{l} Knowledge : \left[\begin{array}{ll} A_2 : & < \{w_1, w_2\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{w_2\}, \{l_1\} > \\ 'sweet' : & < \{\}, \{l_2\} > \end{array} \right] \\ VWM : \left[\begin{array}{ll} 1st\_word : & < \{w_2\}, \{l_1\} > \end{array} \right] \end{array} \right. \tag{7.40}$$

This representation of the word *banana* is now - in the speaker's representation of the mental state of the addressee - activating the conceptual representation of *banana*. If the information that *banana* is the topic predicate is also already available at that stage, and if it is used as a definite description denoting an individual of the category *banana*, then processing step 3 can also be executed. But depending on what information becomes available at what stage of the utterance, word forms can be stored in VWM until processed. This shift of the processing of the sentence structure to the processing of states with the internal model and the value function allows (i) the production of arbitrary sequences, but also enables (ii) the aquisition of the meaning of certain syntactic structures. Using this design an agent can learn several things about the structure of utterances:

- He can acquire an understanding that word repetitions are useless in most cases. E.g. after producing the word *banana*, the production of another word *banana* does not have any useful effect on the addressee (in most cases).
- It allows the acquisition of conventional sentence structures. E.g. an utterance of the form $[[banana_T][here_C]_{req}]$ is more often understood (has the desired effect) than an utterance of the form $[[here_C][banana_T]_{req}]$.
- Meaningful differences in sequences can be understood. E.g. the agents can learn that *John kissed Mary* and *Mary kissed John* produce different changes of knowledge in the addressee's knowledge representation.
- When the model is extended to handle more complex syntactic structure, it can learn the meaning of certain function words. E.g. the agents can learn that the generation of the sequence *put the banana on the table* can produce a different effect from the production of the same sequence with *a* instead of *the* (in the first case the most salient banana has to be put on the table, in the second case it does not matter which banana).

Let me finally point out that such a design does not require multiple representations of word forms as the current slot representation. To represent both utterances *John kissed Mary* and *Mary kissed John*, in the previous design, the model needs one neuron the represent *Mary* in the first position of the utterance and one neuron the represent *Mary* in the third position. In the design just sketched, one representation of the word form is sufficient.

7.7.3 Learning Word Meanings in Context

Suppose a child already has a conceptual representation associated with the word form *window* as well as a representation of the categorical concepts *closed* and *open*, but no words associated with those. It further knows how to produce and comprehend the speech act *request*. Now it observes a speaker who uses the request 7.47 to a third person. The production of the request is based on the representations described in equations 7.42 and 7.43.

$$u = [[window_T][open_C]_{req}] \tag{7.41}$$

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} window\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} open : & < \{\}, \{l_2\} > \\ closed : & < \{w_1\}, \{l_3\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'window' : & < \{\}, \{l_1\} > \\ 'open' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.42}$$

$$\Delta = \left\{ Location : \left[\begin{array}{ll} open : & < \{w_1\}, \{l_2\} > \\ closed : & < \{\}, \{l_3\} > \end{array} \right] \right. \tag{7.43}$$

Since the child understands the topic-predicate *window*, as well as the general way a request works, it knows that something will happen with the window. It can observe that the window changes its position (from being closed to being open). Since the only new word in the utterance was *open*, the child can associate the word with the concept *open* (equations 7.45 and 7.46 show the representations of the child after linking word and concept in working memory).

$$u = [[window_T][open_C]_{req}] \tag{7.44}$$

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} window\_f : & < \{w_1\}, \{l_1\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} open : & < \{\}, \{l_2\} > \\ closed : & < \{w_1\}, \{l_3\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'window' : & < \{\}, \{l_1\} > \\ 'open' : & < \{w_1\}, \{\} > \end{array} \right] \end{array} \right. \tag{7.45}$$

$$\Delta = \left\{ Location : \left[\begin{array}{ll} open : & < \{w_1\}, \{l_2\} > \\ closed : & < \{\}, \{l_3\} > \end{array} \right] \right. \tag{7.46}$$

This should then result in a long term connection with the word and concept (equation 7.48 show the representation of the child after linking word and concept in long term memory).

$$u = [[window_T][open_C]_{req}] \tag{7.47}$$

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} window\_f : & < \{\}, \{l_1\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} open : & < \{\}, \{l_2\} > \\ closed : & < \{\}, \{l_3\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'window' : & < \{\}, \{l_1\} > \\ 'open' : & < \{\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{7.48}$$

7.7.4 Learning the Processing Steps

We are now equipped with a general inverse model which is able to compute utterances of different speech act types from a current state and a desired change. A final question to ask here would be how these processing steps are executed by a neural model, and also whether they can be learned by one or several types of neural learning algorithms.

Unfortunately I cannot answer this question here in sufficient detail, but I would nevertheless briefly sketch out some first ideas.

One of the first things agents have to acquire are categories of effects. They have to learn that effects of a certain category can be accomplished with utterances of a certain type. E.g. changing the state of an object in the world (in the examples in this thesis it is usually the object's position) is one category and the utterance type is a request.

Also, agents have to learn that words can be used to bring a certain object to the attention of an addressee. They can already learn this in the single word stage. The comment information can remain implicit, and has to be guessed by the addressee from the situation. E.g. an utterance *milk* might have an implicit comment-predicate *give*.

Further, agents have to learn that if an addressee is attending to a certain object, some information can be given about it (e.g. by the word *give*, indicating that the object should be given to the speaker). This is also possible in the single word stage. The attention can be drawn to an object by other means, such as gaze following, or pointing.

Then eventually the agents have to learn that verbal actions can accomplish both these goals. At this point they have to acquire the skill to produce an utterance with a topic - comment structure.

However, agents need to use subgoals for an implementation of this principle. For example a desired state of an agent A_1 could be to get an addressee A_2 to open a certain door. This means that A_1 has the representation of a certain door in his brain and this door is represented as closed. To produce the utterance which might cause A_2 to open the door is, according to the theoretical frame developed up to this point, done in several major steps. First, A_1 has to select the speech act type. Second, A_2's attention needs to be directed towards the door; this leads to a decision on the topic predicate. Third, the comment predicate needs to be decided on. These processes are probably formed in that order, but this would be a question of empirical research. However, it can be assumed that the decision regarding the speech act type and the topic-predicate and the comment-predicate are made before A_1 starts to actually produce the words of the utterance.

Given that the utterances I have dealt with so far are direct expressions of speech act type, as well as topic-comment-structure, the translation into a sequence of words using

the model described in section 7.7.2 is straightforward. However, in real natural languages the surface form of the utterance often stands in a more complex relation to this structure.

Chapter 8

Predictions and Empirical Questions

8.1 On Theory and Data

8.1.1 Theory and Model

Before talking more about the relation between theory and data, I would like to draw a clear distinction between *theory* and *model*. In the literature these two words are often used interchangeably or without explaining the difference.

I use the terms in the following way: A theory is a set of formal or informal propositions that are claimed to be true (i.e. cognitively realistic). Those propositions might be simplified, but are claimed to capture essential properties of human information processing. A model of this theory is a computational implementation in accordance with those propositions. However, the model is allowed (and often required) to use additional constructions (mechanisms and representations) to be able to implement the theory and deliver the required results. Those constructions, as far as they are not claimed to capture real properties of human information processing, are not part of the theory. In almost all modelling work this distinction is not worked out in sufficient detail. Often those additional contructions are not really critical, since they are easily identified as justifiable or irrelevant simplifications. For example, depending on the exact goal of the model it might be justified (and necessary) to simplify the representation of speech sounds in a model about meaning processing in the brain, while it might not be justified in a model about speech recognition. This is because a more exact representation might be relevant for the goals of the first, but not for those of the second model.

To simplify cognitively realistic processing is possible as long as the simplifications capture those features of the process and the representations that are relevant for the goals of the simulation. In some way or another however, simplifications have an effect on the processes which are described by the theory. In my simulation (described in chapter 5), the visual input to my model is simplified, but it seemed to be perfectly clear that I do not claim this to be the cognitively realistic input. However, the choice of the representation was made so that the essential features of the learning algorithm (which I *do* claim to be cognitively realistic) could be implemented more effectivley and transparently. Because

of the learning algorithm the visual input has to capture some essential properties of the cognitively real visual representations, however, since otherwise a cognitively realistic algorithm might not work as well.

Often a theory describes the relation between two variables in terms of a mathematical function F. In the model such a function, especially if it describes a complex cognitive process, needs to be implemented as an algorithm. This algorithm does not need to be claimed to be cognitively realistic, which would mean that it is an additional construction and not a part of the theory. If, however, the choice of algorithm has an effect on the behaviour of the model, then the algorithm becomes theoretically relevant and should be discussed and included in the theory. Many backpropagation models, for example, are not clear about the exact status of the learning algorithm they use. Do they claim that the brain uses backpropagation? Do they claim that the brain uses s.th. similar to backpropagation? Or is it just a way of arriving at distributed representations which *are* claimed to be cognitivly realistic? What exactly is the claim here?

And why is this so important? Because clearly stated claims are the only way of deriving clear predictions from a theory. And only clear predictions can be tested. Therefore, only a precisely worked out theory can lead to real scientific progress. Making neural models, since they are computational, forces the researcher to be mathematically precise and algorithmic, and, therefore, is a very suitable method of arriving at precise cognitive theories. However, since most neural models (or modellers) do not clearly distiguish between theory and additional constructions, i.e. the modellers haven't really worked out their theoretical claims, it is hard to see to what extent the data is in accord with their ideas. It is not so hard to fit data to a model, when the model is full of algorithms, representations, and parameters, and it is left in limbo which of those are essential and meaningful (and true). Therefore, it should be made perfectly clear for every computational neural model which parts of it are *realistic* - simplified maybe, but claimed to capture essential properties of the process or representations, and which parts are mere additional constructions introduced just for the sake of making the model work. Everything else leads to confusion and little scientific progress.

8.1.2 Data Types

In this thesis, I have developed a theory about meaning acquisition, representation, and processing in the human brain. Whether the claims made by this theory are true is a question I am not able to answer at this point. But what I will do in this section is to point out how they can be tested by empirical research. A computational theory of meaning processing in the brain cannot be tested directly, since no direct evidence is available. However, there is indirect evidence available on several levels of observation. The strength of the model developed so far is that it can be tested on several of those levels and, therefore, can serve as an integrative platform to be constrained by data of different levels of observation.

These levels of observation can be classified into two broad categories: the *behavioral* level and the *neurophysiological* level.

On the behavioral level, I will classify the data into two main types: The first type I will call *output* data. Output data consists of the language (utterances) humans produce. It includes linguistic data, i.e. data which is generally used by theoretical linguists to evaluate their theories. These are the utterances which normal speakers, who have fully acquired their native language, produce under normal circumstances. But output data also includes the utterances (or pre-linguistic sounds) of children in their early stages of development, as well as the transition between the utterances in earlier and later stages. Further, it includes the errors children and adults make under normal circumstances. Finally, it includes all output under experimental conditions (e.g. in the presence of distractors, noise, etc.).

The second type I will call *performance* data. I will distinguish two types of performance data. The first one consists of the results of *the* standard measurement in behavioral experiments: the reaction time (RT). RT experiments can give a good insight into the complexity and the difficulty of brain processes. The second one is learning rate.

Both types of data can be obtained either in patients (e.g. people with brain diseases or brain lesions which impair their performance) or in (so-called) normals (i.e. people without such impairments). Behavioral data of patients already give insights into the locations of the brain mechanisms involved and the neural architecture of the cognitive processes.

More information about the biological implementation of those processes can be obtained by measurements on the *neurophysiological* level (the second broad category). Here, it isn't the behavior of the tested person that plays the central role, but some measured physiological brain response.

One possible way to categorize data of the neurophysiological level is by considering the number of neurons they measure. Small scale measurements are the recording or stimulation of single or small numbers of neurons by introducing an electrode directly into the brain. This type of research can in some rare cases be applied to humans, e.g. in case they undergo some sort of epilepsy surgery, but in general this type of data is obtained with animals. Therefore, it cannot be directly used to measure and stimulate while a person is performing a linguistic task. However, as animal brains and human brain are very simular in terms of their general mechanisms of computation, the insight this research provides into the general learning and processing mechanisms of the brain can be applied also to design biologically plausible constraints of a model of language processing. I will call the insights into information processing in the human brain provided by this kind of data *design constraints*.

Large scale measurements are measuring methods that measure the activation of large numbers of neurons (or synapses). The most frequent are fMRI and EEG (others are PET, SPECT, MEG, and TMS). Functional imaging data, such as fMRI give us clues about the location of certain brain processes with a fairly good space resolution, but pretty bad time resolution. EEG on the other hand gives us insights into the time response of neural activation, but the space resolution is very low. EEG (and MEG data) can also give us information about activity in different frequency bands (such as the γ-band). Both functional imaging data and EEG data can be simulated with neural models, although in relation to language this has hardly every been done.

Finally, connectivity data, i.e. data about the connection between brain areas, can be

used as another design constraint.

8.1.3 Dependent and Independent Variables

The terms *dependent* and *independent* variable apply mostly to experimental research. Independent variables are those that are manipulated whereas dependent variables are only measured or registered. Dependent variables are called dependent variables because they depend on the manipulation (the experimental conditions) of the independent variable. In that sense, the data types described in the previous section are different kinds of dependent variables. The purpose of a model or a theory is to predict the value of such a dependent variable for different conditions (different parameters or manipulated independent variables). In that sense a computational neural model can be regarded as a function M which computes a dependent variable y from an independent variable (parameter)[1] p (equation 8.1).

$$y = M(p) \tag{8.1}$$

A computational neural model usually computes primarily output data (i.e. linguistic responses). Such a model can be regarded as an output function $O()$ that computes a certain output y from input x. If the input x is different for different experimental conditions then x is a manipulated independent variable; if it stays the same for all the conditions, it is not. In the latter case different parameters $p_1...p_n$ have to function as independent variables $p_1...p_n$ (equation 8.2).

$$y = O(x; p_1...p_n) \tag{8.2}$$

Besides the output data, a neural model can also be used to predict reaction time. To do that, the neural network has to be a function over time t (equation 8.3) while the input x becomes a parameter. To predict the reaction time a thresold function (equation 8.4) is needed. The reaction time is the value t for which the threshold function and the output function intersect (equation 8.5).

$$y = O(t; x, p_1...p_n) \tag{8.3}$$
$$y = \theta \tag{8.4}$$
$$\theta = O(t; x, p_1...p_n) \tag{8.5}$$

It is desirable not only to predict output and behavioural data with the model, but to test it on the basis of physiological data as well. The activation $b^A(t)$ of a certain brain area A at t can be computed as the sum of the activation of all neurons m belonging to that area in the model (equation 8.6).

[1] Of course, there can be more then one independent variable which can be manipulated independent of each other.

$$b^A(t) = \sum_{n=1}^{\omega} m_n^A(t) \tag{8.6}$$

All these major types of data can be computed by one and the same model based on one theoretical framework. The framework I have developed in my thesis can be used as a basis for such functions M that relate independent variables or parameters with dependent variables of various types (output, behavioral, physiological). In the remainder of this chapter, I will try to be more specific about where and how such functions can be extracted from my theory.

8.1.4 Relating Behavioral and Physiological Data

A neural model can be used in a very straightforward way to directly relate observed behavioral data to the areas which are simulated in the model. I will discuss two examples of how this can be done: simulated lesions and synthetic brain imaging.

Simulated Lesions

Evidence for the involvement of a brain area in a cognitive function traditionally comes from lesion studies (Broca 1865, Wernicke 1874).

Recently, there has been a growing interest in constructing neural models to study how specific pathological neuroanatomical and neurophysiological changes can result in various clinical manifestations, and to investigate the functional organization of the symptoms that result from specific brain pathologies.

To simulate lesions, neurons or connections can be destroyed or disabled. Ruppin and Reggia, for example, simulate a focal structure lesion by permanently clamping the activity of the lesioned units to zero. According to Ruppin and Reggia such *structural* lesions (involving neurons which are actually defect or destroyed) cause *functional lesions* of areas which to a large degree depend for their input on the estroyed units in the area of the structural lesion.

For the computational model described in this thesis there are many simulations of lesions possible. One of the most interesting ones would be to destroy those neurons in the internal model which have the function of the Purkinje cells in the cerebellum and to see whether this produces malfunctions in the behavior of the model comparable to the behavior of Autists. This would be an interesting simulation because it is a common finding in Autistic patients that they suffer from a deficiency of Purkinje cells in the cerebellum. These cells are important for the cerebellum in the acquisition of a forward model. Therefore, a functional connection between those two phenomena might be found by producing similar lesions in the neural model.

Synthetic Brain Imaging

Neural models have been used to synthesize functional imaging data for more than 10 years now (Arbib et al. 1995, Tagaments and Horwitz 1998). To use this approach for linguistic tasks is complicated, but possible (Klein and Billard 2001).

Brain activity can be simulated by computing a value of activation of the neurons in every defined brain structure and every discrete point in time while the behavior is simulated. To synthesize brain activation with the model described in this thesis it is necessary to attribute a brain region to those units that have not yet been correlated with such an area.

The neural network representing the value function has an input layer for the perception of the state, and an output layer consisting of only one neuron. The network representing the forward model has an input layer for the perception of the state, an input slot for the word and an output layer consisting of a neural representation of the state. The input layers of both networks should be considered to be part of the perceptual system, and, therefore, be considered to simulate cortical sensory areas. The output layer of the forward model, on the other hand, can be used to simulate the cerebellum. The output neuron of the basal ganglia represents activity of the basal ganglia.

To generate synthetic fMRI data, a computer simulation needs to be designed that can also be executed as an experiment. To predict fMRI data for a classical subtraction design, for example, two conditions that differ in one parameter have to be specified (e.g. p_1 in equation 8.3 can have the value v_1 in condition 1 and v_2 in condition 2).

8.2 Predictions

8.2.1 Forward or Inverse Models

One central claim of the theory developed in this thesis is that internal models play an essential role in the selection of utterances. Recall that an internal model is an internal representation of the dynamics of a system (e.g. an arm, a billiard ball, a human being, the world etc.). If an agent knows how such a system in certain context conditions reacts to a certain input parameter set (e.g. and action), then he can be said to have an internal model of this system. To illustrate this, imagine that you are trying to lift a watering can full of water from the ground. As it turns out the watering can had a little hole and the water is all gone. Since you expect the watering can to be full of water and very heavy, your movement is far too powerful for the empty can and you pull it into the air in a fast uncontrolled movement. The reason, why your movement is so uncontrolled is that you computed a different outcome for your motor action parameters by using your internal model of the dynamics of the watering can.

While the involvement of internal models in the selection of verbal action has not been tested so far, their involvement in motor action (involving speech motor control) is already tested and widely accepted. In motor control, forward models appear to be necessary,

because movements are too fast to be produced on slow sensory feedback alone. With internal models, motor commands can be executed in a pure feed-forward manner.

Utterance selection can't be based on sensory feedback either, because the consequences of utterances cannot be perceived until a reaction of the environment has taken place. Therefore, for utterance selection it is very likely that the reaction of the environment is predicted by the speaker using an internal model of the environmental dynamics.

In motor control the existence of forward models has been demonstrated with experiments in which subjects had to undertake point-to-point arm reaching movements. Such reaching movements, if they are fast enough, need to use an internal model of the dynamic characteristics of the arm, i.e. the brain musst have learned what kind of motor command is necessary to bring the arm and the hand into a desired position with a fast movement. In the experiment the dynamics of the arm was changed using a force field (Shadmehr and Mussa-Ivaldi 1994, Lackner and Dizio 1994). As action selection is done using a internal model, motor commands were still selected due to the relation between them and effects stored in the internal model (which was no longer appropriate for the changed dynamics), and it took some time till the model was adapted to the new dynamics. Therefore, movement was distorted in the beginning and adapted to the new dynamics after a short while. When the force field was then removed, the movement was again distorted for a short time.

The same type of experiment would be possible with language. The usual effects of utterances could be altered in a simulated environment. Subjects would select utterances due to the utterance - effect relation they are familiar with. Subjects are then likely to change their utterances with respect to the new dynamics of the environment.

These experiments could also be executed in a PET or an fMRI scanner. As opposed to what is found in studies investigating neural activation correlating to internal models in motor control (Tamada et al. 1999b, Imamizu et al. 2000), experiments using language are likely to activate some of the known language areas (e.g. those where lexical meaning might be stored). It would be interesting to see which brain regions would be activated when the subject adapts to the new utterance - effect relationship. Our model predicts that activation correlating to the processing of the difference between the predicted and the perceived effect will be found in the cerebellum.

In case internal models *do* play an essential role in utterance selection, it is also necessary to find out whether this internal model is a forward model $F()$, predicting the next state $s(t+1)$ from the current state $s(t)$ and the verbal action u (equation 8.7) or an inverse model $I()$, computing the required verbal action u to get from the current state $s(t)$ to the next state $s(t+1)$ (equation 8.8). Utterance selection with the forward model is described in equation 8.9 and equation 8.10 describes utterance selection with the inverse model.

$$s_{t+1} = F(s_t, u) \tag{8.7}$$

$$u = I(s_t, s_{t+1}) \tag{8.8}$$

$$u = argmax_u V(F(s_t, u)) \tag{8.9}$$

$$u = I(s_t, argmax_s V(s_{t+1})) \tag{8.10}$$

To use the forward model for utterance selection as in equation 8.9, the outcome of all (or at least several) possible utterances needs to be predicted because the algorithm has to compute the outcomes s_{t+1} of all possible u in the current state s_t to find the outcome with the highest value. This necessarily leads to an activation of those utterances. Since this activation would make it easier to retrieve those sentences, a priming experiment could be used to test whether or not a forward model is at work.

However, a systematic computation of all possible utterance effects and the expected values is hardly possible. It is, therefore, very likely that certain context situations, certain desires, objects and activities trigger certain words, verbal patterns or utterances simply because they are associated with them over time. In other words, utterance selection by internal models is aided by associative links between words, concepts, context states, desired states and sentence patterns. This means that, for example, if a person knocks at my office door, I will not use a forward model to predict the effects of all possible utterances including *can you please pass me the salt* and *your plants really need water*. It is more likely the case that I associate this knock at my door with (i) the most likely future state, i.e. a person standing on my door step as well as with (ii) the most frequent verbal reaction to this situation, i.e. *yes* or *come in*. The forward model can then be used to predict the effect of *come in* and if that effect is okay, I will just produce it and not bother predicting the effect of other possible utterances. Or my desire not to be disturbed (triggered by the knock) can be used as input to an inverse model producing a *wait* or a *not now* or maybe even the reaction of not saying anything at all to make the person go away. But how exactly these associative connections can restrict or bias the input of the utterance selection algorithm needs to be described in more detail before a good experiment could be set up to test this idea.

8.2.2 Reward Prediction

For the reinforcement learning component, there are already brain imaging studies testing reward prediction (Breiter et al. 2001, Seymour et al. 2004, Haruno et al. 2004, Tanaka et al. 2004). However, in most of these studies, the reward prediction has been tested for actions and not for states, i.e. subjects had to predict the reward they would get for a certain action and not necessarily for the state which would be brought about by the action. Therefore, it would be interesting to see whether the activation in the striatum could be reproduced in settings where the reward is given as dependent on the states instead of the actions. This could be done by using actions for which the consequent state, but not the reward has been experienced by the tested subject. In a similar study, verbal and non-verbal actions could be used to bring states about. This would allow us to see whether the activation correlating with reward prediction is the same for verbal and non-verbal actions (which is predicted by our model). One of the most interesting questions to be answered by such an experiment would be whether these reward predictions are language specific and or whether the language faculty of the brain employs general (non-language specific) reward mechanisms.

8.2.3 Acquisition of First Words

As we have discussed in the previous chapter, children use their first words in so called holophrases which are composed of two components: a content word and an intonational pattern indicating the speech act (declarative, imperative, interrogative). Here, I will describe the acquisition of this capability by means of my theoretical framework. I consider it to be important to distinguish how the child actually comprehends the communicative intention of the utterance from how it stores the relation between utterance form and meaning, as well as how it can use the knowledge it has gained.

Comprehending the Intention

Suppose a specific learning situation in which a child learns the word *ball* from its mother. The situation is the following: The mother points her finger to a ball and by doing this directs the attention of the child towards the ball. While doing this she says *ball* with a declarative intonation. I will ignore the problem that the child may have when it must single out the relevant word when the mother uses a full sentence.

So far I have mostly focused just on the current state and the desired state of the speaker. But for language acquisition it is necessary to describe the state of the addressee (i.e. the child). Nevertheless, I will start with the current state and the desired state of the speaker (the mother), since she uses the representations and algorithms which the child actually needs to acquire. Note that in the following representation of the state of the mother the described attention is the attention of the child. The desire of the mother is to make the child focus its attention on the ball. To do this, the mother needs to go through the processing steps described in chapter 7.6.1. First the speech act type is selected. So far, declarative intonation was selected when the speaker wanted to change the knowledge of the addressee. Manipulating his attention is very similar to the manipulation of his knowledge and declarative intonation seems to be appropriate for this goal, too. Due to the simplicity of the sentence, the mother can select the utterance in the first two processing steps: 1. selection of declarative intonation and 2. selection of the content word. It seems straightforward to assume that the goal of manipulation of attention or knowledge is associated with declarative intonation resulting in a brain state of the mother described in equations 8.11 and 8.12.

$$C_{A_1} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} ball\_f : & < \{w_1\}, \{l_1\} > \\ mother\_f : & < \{w_2\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_1\}, \{\} > \\ position[x_2, y_2] : & < \{w_2\}, \{\} > \end{array} \right] \\ Attention : \left[\begin{array}{ll} A_2 : & < \{w_2\}, \{l_2\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{l_2\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'ball' : & < \{\}, \{1_1\} > \\ decl : & < \{\}, \{l_2\} > \\ imp : & < \{\}, \{\} > \\ quest : & < \{\}, \{\} > \end{array} \right] \end{array} \right. \tag{8.11}$$

$$\Delta = \left\{ \begin{array}{l} Attention : \left[\begin{array}{ll} A_2 : & < \{w_1\}, \{l_2\} > \end{array} \right] \end{array} \right. \tag{8.12}$$

Now for the addressee A_2 (i.e. the child). The initial state is represented in equation 8.13 (with the attention of the child towards the mother).

$$C_{A_2} = \left\{ \begin{array}{l} Sight : \left[\begin{array}{ll} ball\_f : & < \{w_1\}, \{\} > \\ mother\_f : & < \{w_2\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[x_1, y_1] : & < \{w_1\}, \{\} > \\ position[x_2, y_2] : & < \{w_2\}, \{\} > \end{array} \right] \\ Attention : \left[\begin{array}{ll} A_2 : & < \{w_2\}, \{\} > \end{array} \right] \\ Knowledge : \left[\begin{array}{ll} A_2 : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'ball' : & < \{\}, \{\} > \\ decl : & < \{\}, \{\} > \\ imp : & < \{\}, \{\} > \\ quest : & < \{\}, \{\} > \end{array} \right] \end{array} \right. \tag{8.13}$$

From the pointing gesture of the mother the child understands her intention to direct the child's attention to the ball.

$$C(A_2) = \left\{ \begin{array}{l} Intention : \left[\begin{array}{ll} A_1 : & < \{w_3\}, \{\} > \end{array} \right] \\ Attention : \left[\begin{array}{ll} A_2 : & < \{w_3, w_1\}, \{\} > \end{array} \right] \end{array} \right. \tag{8.14}$$

This explanation, of course, is only valid if language acquisition can build on the capability of the child to understand non-verbal communicative intentions, such as pointing gestures and gazes. Evidence that such capabilites exist in children is summarized by Tomasello (2003).

Acquisition

Let us assume that the child understands that a basic declarative use of the word *ball* has the same communicative intention as a pointing towards a ball. The child thus needs to

learn that this kind the declarative usage is like pointing and the word denotes the object which is pointed out. How can the child learn that? Some researchers think that the basic intonation pattern for different types of situations are innate. Other opinions are that the child observes that different intonation patterns correlate with certain intentions of the adults e.g. a rising intonation is used when some form of reaction is required (question), while a more flat pattern is used when the utterance is merely a comment (or assertion). In our case, the child might be able to observe a certain intonation pattern that is frequently accompanied by a pointing gesture and it understands that this intonation means that it is supposed to direct its attention towards some thing. Having understood this, the child can also understand that the word varies in relation to the particular object to which the reference is made.

To test this hypothesis is not easy. There are many studies describing the use of intonation in different situations with (apparently) different goals. But to test where this use of intonation is coming from is difficult. To investigate the very cries that use a specific pattern might not help since by then the child has already had verbal interaction with its parents and has been exposed to different intonational patterns. One possibility would be to design a new intonation pattern (e.g. a very high pitch beginning which drops linearly to very low one). This intonation pattern could be used for example with new words for new objects and a gesture (or a gentle push) indicating that the child is supposed to move to that object. This experiment could also be done with adults. However, it is not yet clear what the dependent and the independent variables (the manipulations) of such a experiment should be. These aspects of such an experiment might be more apparent when the acquisition part of the theory is worked out in more mathematical and computational detail.

Usage

To use its newly acquired communication skills the child needs to be in a situation where it needs to point out different objects to another person. Assuming that it has acquired the association between speech act and intonation, as well as the associations between object and word, the child can now select both accordingly (using the same mechanisms as the mother).

8.2.4 Comprehension of Complex Utterances

This section will nicely show how the theory developed in this thesis makes predictions on the behavioral and physiological level, while connecting nicely to ideas of theoretical linguistics and philosophy of language. The theory distinguishes states of the systems in terms of information in long term memory, which is stored in terms of long term memory binding, i.e. increase in synaptic strength from information in working memory, which is stored by working memory binding, i.e. increase in synchronization between neurons. It further describes the process of retrieving information from long term memory into working

memory and the other way round. It also distinguishes episodic from semantic memory. All these distinctions and processes are described in terms of neural and synaptic behavior.

Here I will describe the comprehension of sentence (1). I will describe the different processing steps and the possible manipulations of the independent variables to test the theory. Note that in this example I have ignored the idea that utterances are always causing a representation of the intention of speaker in the brain of the addressee, which is another aspect of the theory which should be tested experimentally.

(1) The banana is sweet.

Equation 8.15 describes the initial state of the addressee.

$$C_{A_2}(t) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \\ Lexicon: \left[\begin{array}{ll} 'banana': & <\{\},\{l_2\}> \\ 'sweet': & <\{\},\{l_3\}> \end{array} \right] \end{array} \right. \tag{8.15}$$

Processing Step 1

The first processing step is the activation of the word *banana*, i.e. the activation of the lexical representation from the acoustic and phonetic features (which I did not include in the model).

$$C_{A_2}(t+1) = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} banana\_form\_feature\_1 : & < \{\}, \{l_1, l_2\} > \\ other\_form\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ V4 : \left[\begin{array}{ll} banana\_color\_feature\_1 : & < \{\}, \{l_1, l_2\} > \\ other\_color\_feature\_2 : & < \{\}, \{\} > \end{array} \right] \\ TA : \left[\begin{array}{ll} sweet\_feature\_3 : & < \{\}, \{l_3\} > \\ other\_taste\_feature\_4 : & < \{\}, \{\} > \end{array} \right] \\ Location : \left[\begin{array}{ll} position[1,1] : & < \{\}, \{l_1\} > \\ position[1,2] : & < \{\}, \{\} > \\ position[2,1] : & < \{\}, \{\} > \\ position[2,2] : & < \{\}, \{\} > \end{array} \right] \\ Time : \left[\begin{array}{ll} t[1] : & < \{\}, \{l_1\} > \\ t[2] : & < \{\}, \{\} > \\ t[3] : & < \{\}, \{\} > \\ t[4] : & < \{\}, \{\} > \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & < \{w_1\}, \{l_2\} > \\ 'sweet' : & < \{\}, \{l_3\} > \end{array} \right] \end{array} \right. \tag{8.16}$$

Processing Step 2

The second step is the activation of the lexicalized concept associated with the word. It is a retrieval of information from long term memory to working memory. In working memory it is represented as a synchronized activation of the features of the concept.

$$C_{A_2}(t+2) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \\ Lexicon: \left[\begin{array}{ll} 'banana': & <\{w_1\},\{l_2\}> \\ 'sweet': & <\{\},\{l_3\}> \end{array} \right] \end{array} \right. \tag{8.17}$$

Processing Step 3

The third step is the activation of the individual concept to which the activated category concept applies from long term into working memory. The activated features of the lexicalized concept activate the features of the individual concept.

$$C_{A_2}(t+3) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{w_1\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{w_1\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \\ Lexicon: \left[\begin{array}{ll} 'banana': & <\{\},\{l_2\}> \\ 'sweet': & <\{\},\{l_3\}> \end{array} \right] \end{array} \right. \tag{8.18}$$

Processing Step 4

The understanding of the comment predicate then triggers the activation of the categorical concept of *sweet* (equation 8.19).

$$C_{A_2}(t+4) = \left\{ \begin{array}{l} VFA : \left[\begin{array}{ll} banana\_form\_feature\_1 : & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2 : & <\{\},\{\}> \end{array} \right] \\ V4 : \left[\begin{array}{ll} banana\_color\_feature\_1 : & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2 : & <\{\},\{\}> \end{array} \right] \\ TA : \left[\begin{array}{ll} sweet\_feature\_3 : & <\{w_2\},\{l_3\}> \\ other\_taste\_feature\_4 : & <\{\},\{\}> \end{array} \right] \\ Location : \left[\begin{array}{ll} position[1,1] : & <\{w_1\},\{l_1\}> \\ position[1,2] : & <\{\},\{\}> \\ position[2,1] : & <\{\},\{\}> \\ position[2,2] : & <\{\},\{\}> \end{array} \right] \\ Time : \left[\begin{array}{ll} t[1] : & <\{w_1\},\{l_1\}> \\ t[2] : & <\{\},\{\}> \\ t[3] : & <\{\},\{\}> \\ t[4] : & <\{\},\{\}> \end{array} \right] \\ Lexicon : \left[\begin{array}{ll} 'banana' : & <\{\},\{l_2\}> \\ 'sweet' : & <\{\},\{l_3\}> \end{array} \right] \end{array} \right. \tag{8.19}$$

Processing Step 5

The attribution of the property *sweet* to the individual banana takes the form of synchronizing the activity of the individual concept with the categorical concept, leading to a uniform representation of the sweet banana (equation 8.20) in working memory.

$$C_{A_2}(t+5) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{w_1\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{w_1\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{w_1\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \\ Lexicon: \left[\begin{array}{ll} 'banana': & <\{\},\{l_2\}> \\ 'sweet': & <\{\},\{l_3\}> \end{array} \right] \end{array} \right. \tag{8.20}$$

Processing Step 6

This can lead to a binding of the feature sweet to the individual concept in long term memory (equation 8.21).

$$C_{A_2}(t+6) = \left\{ \begin{array}{l} VFA: \left[\begin{array}{ll} banana\_form\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_form\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ V4: \left[\begin{array}{ll} banana\_color\_feature\_1: & <\{w_1\},\{l_1,l_2\}> \\ other\_color\_feature\_2: & <\{\},\{\}> \end{array} \right] \\ TA: \left[\begin{array}{ll} sweet\_feature\_3: & <\{w_1\},\{l_3\}> \\ other\_taste\_feature\_4: & <\{\},\{\}> \end{array} \right] \\ Location: \left[\begin{array}{ll} position[1,1]: & <\{w_1\},\{l_1\}> \\ position[1,2]: & <\{\},\{\}> \\ position[2,1]: & <\{\},\{\}> \\ position[2,2]: & <\{\},\{\}> \end{array} \right] \\ Time: \left[\begin{array}{ll} t[1]: & <\{w_1\},\{l_1\}> \\ t[2]: & <\{\},\{\}> \\ t[3]: & <\{\},\{\}> \\ t[4]: & <\{\},\{\}> \end{array} \right] \\ Lexicon: \left[\begin{array}{ll} 'banana': & <\{\},\{l_2\}> \\ 'sweet': & <\{\},\{l_3\}> \end{array} \right] \end{array} \right. \tag{8.21}$$

Experiments/Dependent Variables/Possible Manipulations

Spelling out the process of understanding utterances in my theoretical framework at this level of detail immediately leads to several interesting empirical predictions which can be tested experimentally. I grouped these predictions with the processing steps which they are testing.

Processing Step 2

- Depending on the feature types different brain areas are predicted to activate. This can be tested with fMRI or with EEG (in which case the activation is predicted to be in the gamma range).
- Higher feature load should lead to a stronger activation.

Processing Step 3

The comprehension of sentences that refer to an individual by means of category information (e.g. *the president*) include processing step 3, while sentences that do not refer to an individual do not include this step. Sentence (2), for example needs to use the category concept triggered by step 2 to activate an individual concept, while to understand sentence (3) step 3 is not required.

(2) A president has a lot of power (unspecific reading).

(3) The president has a lot of power.

A difference in processing steps suggests a reaction time experiment in which the processing time of the two condition are compared (less processing steps is likely to result in faster reaction time). Since the reaction time depends on many other factors as well, such an experiment needs to carefully control those factors.

Processing Step 4

Processing step 4 is like step 2.

Processing Step 5

Depending on whether the utterance takes the form of an attribution or not, a difference in coherence is predicted.

(4) The yellow object and the triangle I saw yesterday in the lesson.

(5) The yellow object is a triangle I saw yesterday in the lesson.

In utterance (5), in which *a triangle I saw yesterday in the lesson* is attributed to the yellow object, form and color areas of the brain should synchronize while they should not for utterance (4), which describes the yellow object and the triangle as two distinct entities.

Processing Step 6

Depending on whether the information is considered relevant for the future, it will be stored in long term memory. A different pattern of EEG and fMRI activity can be expected for information stored in long term memory than for information which is not.

Chapter 9

Conclusion

In this thesis, I have developed a theory about the acquisition, processing, and representation of meaning in the human brain in terms of neural computations. While the theory is very far from dealing with the large variety of complex sentences possible in natural language (at the moment it deals with a simplified pseudolanguage) it explains what I consider to be the most important ingredients of natural language: conceptual representations, compositionality, topic comment structure, goal directed utterance selection, and the usage of different speech acts.

The strength of the theory is that it bridges ideas of theoretical semantics and philosophy of language with psychological phenomena (attention, long term and working memory, episodic and semantic memory, concepts, language acquisition) and explains both in terms of physiologically realistic neural computations involving references to anatomical regions of the brain.

The major work that remains to be done is to experimentally evaluate the theory. In chapter 8, I have started to work out its exact predictions so it can be tested with appropriate experimental methods, such as EEG, fMRI, patient studies, behavioral studies of language acquisition with children and language use with adults. Because the complete theoretical framework is very complex, it might be necessary to evaluate small parts of the theory in different experiments. Also, to arrive at more exact predictions, and also to combine the different empirical levels this theory integrates (behavioural, physiological) it will be necessary to implement more parts of it in form of computational neural models. This would also make it possible to simlulate lesions and synthesize fMRI data.

While on the neuroscientific side, it will be important to evaluate the theory with experimental means, for the linguistic side of the theory it will be important to develop it to a level of detail which allows it to deal with natural language sentences. This would make it more relevant to linguists.

Summary

This thesis deals with the question how the human brain acquires, represents, and processes the meaning of natural language expressions. A computational neural theory of meaning is developed with the goal of overcoming the strong prevalence of empirical results over theoretical understanding that is currently present in the neuroscience of language. In this context, the brain is regarded as a goal-directed system, which acquires language and meaning as one means for achieving its goals. In this framework, learning of language is an optimization process, in which the performance of the language learner in handling the environment increases with a higher level of linguistic capacity. To accomplish complex learning tasks, such as acquiring a language, the brain uses subsystems, which differ especially with respect to their learning strategies, but interact so as to achieve the global goals of the system. One of the main subsystems is concerned with prediction: we learn language by observing the changes in the environment that utterances can bring about in certain contexts. In other words, we learn to predict the context-dependent effects of utterances. As trained speakers, we can construct and select utterances in accordance with how much we desire their predicted effects. In the introduction I discuss these motivations that triggered the development of this theory. In particular, I discuss the need for massive integration of data types by means of a computational neural model, learning in the human brain, the question of innateness, and why it is important to consider the goals of speakers from the very beginning of theory development.

In the second chapter, I describe the view that major brain structures differ especially in respect of their learning strategies (Doya 1999). I explain three major types of learning algorithms in mathematical detail: reinforcement learning, supervised learning and unsupervised learning and review the evidence for the thesis that each of those learning methods is the learning method of one major structure of the brain (unsupervised learning is the method of the cortex, reinforcement learning is the method of the basal ganglia, and supervised learning is the method of the cerebellum).

On the basis of these different subsystems of the brain, the third chapter describes how these components interact and jointly function as a goal-directed communication system. I describe what the function of each of these subsystems in communication is. In particular, I describe the use of the cerebral cortex in unsupervised concept acquisition, the role of the basal ganglia in goal-directed utterance selection using reinforcement learning, and the role of the cerebellum in acquiring internal models for predicting the effect of potential utterances.

In the fourth chapter, I present a computational framework of concept acquisition and representation in the cerebral cortex. This theory deals with the distinction between individual and categorical concepts, and with the transmission of information about certain individuals from one person to another by means of the *topic comment* structure of an utterance.

In the fifth chapter, I present simulation experiments that provide evidence for my theoretical framework. In a multi-agent language game, agents use reinforcement learning to train a value-function. This value-function is used along with a forward model to select actions. The agent can select verbal and non-verbal actions, depending on whether speaking or manipulating the environment *directly* is more likely to bring about the change which the agent desires. Moreover, the learner can take his clues from the behavior of the two other players, who strictly follow the rules the learner is to learn. The learner trains a forward model of context-dependent utterance effects, which he then uses to express his desires as well as to understand the verbal expressions used by other agents to convey their desires.

The sixth chapter extends the theoretical framework, which so far could only deal with requests, to deal with more speech act types, in particular, assertions, and questions. It discusses the exact modification of the value function and the internal model that are necessary to deal with these other types of speech acts. It also describes a possible simulation environment to test those extensions.

In the seventh chapter, I discuss aspects of compositionality. First, I present some simulation work, in which the forward models of the agents are extended with a layer of features (color, form, movement) and an attention component. This extension enables the agents to learn and use compositional aspects of language. After showing the results of these simulations, I provide a theoretical extension of the framework that includes the descriptive formalism developed in chapter 4. I use this theoretical extension to decompose the speech act information (coded in the intonation of the utterance) and the reference information (coded in the word form).

In the eighth chapter I show how the theory developed in this thesis can be tested empirically. After discussing several types of data, such as linguistic, behavioral, clinical and physiological data that can be used to evaluate my theory, I outline one acquisition and one comprehension experiment and describe which exact theoretical predictions can be tested with what kind of experimental methods.

The final chapter concludes about what has been achieved in this thesis and sketches future direction of research in line with the research in this thesis.

Deutsche Zusammenfassung

Diese Arbeit behandelt die Frage, wie das menschliche Gehirn die Bedeutung natuerlicher Sprache erlernt, repraesentiert und prozessiert. Das derzeitige Uebergewicht an empirischen Daten gegenueber theoretischem Verstaendniss in der neurowissenschaftlichen Erforschung der Sprache soll durch die Entwicklung einer computationellen neuronalen Theorie der Bedeutung ausgeglichen werden. In diesem Zusammenhang wird das Gehirn als ein zielgerichtetes System betrachtet, welches Sprache und Bedeutung erlernt als ein Mittel, um sein Ziel zu erreichen. In diesem Rahmen wird das Lernen der Sprache zu einem Optimierungsprozess, in welchem die Faehigkeit des Sprachlerners seine Umwelt zu meistern sich verbessert je mehr er seine Sprache beherrscht. Um die komplexen Aufgaben des Sprachlernens zu meistern benutzt das Gehirn Teilsysteme (Strukturen) die sich in ihrer Lernstrategie unterscheiden, aber zusammenarbeiten um das globale Ziel des Systems zu erreichen. Eines der wichtigsten Teilsysteme beschaeftigt sich mit Voraussagen: Wir lernen die kontextabhaengigen Effekte von Aeusserungen vorauszusagen indem wir die Veraenderungen beobachten, die Aeusserungen in bestimmten Kontexten verursachen. Wenn wir unsere Sprache beherrschen, koennen wir die Aeusserungen auswaehlen (und produzieren) deren vorausgesagten Effekt wir erstreben. In der Einleitung dieser Arbeit diskutiere ich die Motivationen die zur Entwicklung dieser Theorie gefuehrt haben. Im Besonderen diskutiere ich die Notwendigkeit massiver Integration von Datentypen durch computationelle neuronale Modelle, Lernen im menschlichen Gehirn, die Frage welche Aspekte der Sprache angeboren und welche erlernt sind, und warum es essentiell ist die Ziele des Sprechens bei der Entwicklung einer solchen Theorie von Anfang an in Betracht zu ziehen.

Im zweiten Kapitel beschreibe ich eine Theorie, die sagt, dass die wichtigsten Hirnstrukturen sich insbesondere dadurch unterscheiden wie sie lernen (Doya 1999). Ich erklaere drei wichtige Arten von Lernalgorithmen in mathematischem Detail: (*reinforcement learning*, *supervised learning* and *unsupervised learning*). Danach gebe ich einen Ueberblick ueber die Evidenz fuer die These, dass jede dieser Lernmethoden die Lernmethode eine der wichtigen Gehirnstrukturen ist (Unsupervised Learning ist die Methode des Kortex, Reinforcement Learning ist the Methode der Basal Ganglia und Supervised Learning ist die Methode des Zerebellums).

Auf der Grundlage dieser verschiedenen Teilsysteme des Gehirns beschreibt das dritte Kapitel wie diese Komponenten interagieren und als zielorientiertes Kommunikationssystem zusammenarbeiten. Danach erlaeutere ich was die Funktion jedes dieser Teilsysteme

im Gesamtsystem ist. Ich beschreibe insbesondere die Aufgabe des zerebralen Kortex im unueberwachten Konzepterwerb, die Aufgabe der Basal Ganglien in der zielgerichten Aeusserungswahl und die Rolle des Zerebellums in der Aneignung von *internen Modellen* zur Voraussage von potentiellen Aeusserungseffekten.

Im vierten Kapitel praesentiere ich eine computationelle Theorie des Konzepterwerbs und der Konzeptrepraesentation im zerebralen Kortex. Diese Theorie behandelt unter anderem die Unterscheidung zwischen Individuen- und Kategorienkonzepten, sowie die Uebertragung von Informationen von einer Person zur naechsten mit Hilfe der *topic - comment* Struktur einer Ausserung.

Das fuenfte Kapitel beschreibt Simulationsexperimente welche Evidenz fuer meine Theorie liefern. In einem Multi-Agenten-Sprachspiel benutzen Agenten *reinforcement learning* um eine *Wertefunktion* zu trainieren. Diese Wertefunktion wird gemeinsam mit dem internen Modell benutzt um Handlungen zu waehlen. In Abhaengigkeit davon, ob es den Agenten seinem Ziel naeher bringt zu sprechen oder die Welt direkt zu manipulieren, kann er zwischen verbalen und nicht verbalen Handlungen waehlen. Darueberhinaus kann der Lerner seine zwei Mitspieler imitieren, welche streng den Regeln folgen, die der Lerner lernen muss. Der Lerner trainiert sein *forward model* (eine Art von internem Modell) der kontextabhaengigen Aeusserungseffekte und benutzt es um seine eigenen Ziele auszudruecken, aber auch um die Ausdruecke zu verstehen, die andere Agenten benutzen um ihre Wuensche zu vermitteln.

Im sechsten Kapitel erweitere ich den theoretischen Rahmen, welcher bis dahin nur *Forderungen* behandelt hat, um die zwei weiteren zentralen Sprechakttypen *Fragen* und *Aussagen*. Das Kapitel diskutiert die Veraenderungen die fuer das interne Modell und die Wertefunktion vorgenommen werden muessen so dass diese auch mit den anderen Sprechakttypen umgehen koennen. Auch eine moegliche Simulation um die Erweiterungen zu testen wird beschrieben.

Im siebten Kapitel diskutiere ich Aspekte der Kompositionalitaet. Zunaechst praesentiere ich ein paar kleinere Simulationen in welchen das *forward model* des lernenden Agenten durch eine Schicht an Merkmalen (Farbe, Form, Bewegung), sowie einer Aufmerksamkeitskomponente bereichert wurde. Diese Erweiterungen ermoeglichen dem Agenten die kompositionellen Aspekte der Sprache zu lernen und zu benutzten. Danach erweitere ich auch den theoretischen Rahmen dieser Arbeit in dem ich dem Formalismus der in Kapitel vier entwickelt wurde, in die Gesamttheorie integriere. Mit Hilfe dieser Erweitung ist es mir dann moeglich Sprechakte in einerseits die Sprechaktinformation (kodiert in der Intonation der Aeusserung) und andererseits Referenzinformation (kodiert in der Wortform) zu dekomponieren.

Im achten Kapitel zeige ich, wie die in dieser Arbeit entwickelte Theorie empirisch getestet werden kann. Nachdem ich verschiede Datentypen (linguistische Daten, Verhaltensdaten, sowie klinische und physiologische Daten) die benutzt werden koennen um meine Theorie zu evaluieren, diskutiert habe, beschreibe ich ein Erwerbs- und ein Verhaltensexperiment und erlaeutere, welche exakten theoretischen Voraussagen mit welchen experimentallen Methoden getestet werden koennen.

Das letzte Kapitel beschreibt was diese Arbeit insgesamt erreicht hat und skizziert kurz

wie die Forschung, die in dieser Arbeit beschrieben wurde, weitergefuehrt werden sollte.

Bibliography

Michael A. Arbib. The mirror system, imitation, and the evolution of language. In C. Nehaniv and K. Dautenhahn, editors, *Imitation in Animals and Artefacts*. MIT Press, 2000.

Michael A. Arbib, A. Bischoff, A.H.Fagg, and S.T.Grafton. Synthetic pet: Analyzing large-scale properties of neural networks. *Human Brain Mapping*, 2:225–233, 1995.

A. Artola, S. Brocher, and W. Singer. Different voltage dependend thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, 347:69–72, 1990.

John L. Austin. *Philosophical Papers*. Oxford University Press, 1961.

A. Bailey, P. Luthert, A. Dean, B. Harding, I. Janota, M. Montgomery, M. Rutter, and P. Lantos. A clinicopathological study of autism. *Brain*, 121:889–905, 1998.

Simon Baron-Cohen. The cognitive neuroscience of autism. *Journal Neurol. Neurosurg. Psychiatry*, 75:945–948, 2004.

M. L. Bauman and T. L. Kemper. Neuroanatomic observations of the brain in autism. In M. L. Bauman and T. L. Kemper, editors, *The neurobiology of autism*, pages 119–145. John Hopkins University Press, 1994.

M. L. Bauman and T. Kempner. Histoanatomic observation of the brain in early infantile autism. *Neurology*, 35:866–874, 1985.

T. V. P. Bliss and T. Lomo. Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthesized rabbit following stimulation of the perforatant path. *Journal of Physiology*, 232:331–356, 1973.

Paul Bloom. *How Children learn the meaning of words*. Cambridge University Press, 2000.

Valentino Braitenberg. *Vehicles - Experiments in Synthetic Psychology*. MIT Press, 1984.

Valentino Braitenberg and Friedemann Pulvermueller. Towards a neurological theory of language. *Naturwissenschaften*, 1992.

Valentino Braitenberg and Almut Schuez. *Cortex: Statistics and Geometry of Neural Connectivity*. Springer, 2 edition, 1998.

Richard Breheny. Theories of communication and theory of mind. *in preparation*, 2002.

H. C. Breiter, I. Aharon, D. Kahneman, A. Dale, and P. Shizgal. Function imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, 30:619–639, 2001.

Paul Broca. Sur la faculte de langage articule. *Bulletin de la Societe d'Antropologie*, 6: 337–393, 1865.

K. Brodmann. *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaus*. Barth und Campell, 1909.

Angelo Cangelosi, A Greco, and S. Harnad. Symbolic grounding and the symbolic theft hypothesis. In Angelo Cangelosi and Domenico Parisi, editors, *Simulating the Evolution of Language*, pages 191–210. Springer, London, 2002.

Franklin Chang. Sybolically speaking: A connectionist model of sentence production. *Cognitive Science*, 26:609–651, 2002.

Noam Chomsky. A review of b. f. skinner's verbal behavior. *Language*, 35(1):26–58, 1959.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.

Noam Chomsky. *The Minimalist Program*. MIT Press, 1994.

Michele Chouinard and Eve V. Clark. Adult refomulations of child's errors are negative evidence. *Journal of Child Language*, 30:637–669, 2003.

E. V. Clark. What's in a word. In T.E. Moore, editor, *Cognitive Development and the Acquisition of Language*. New York: Academic Press, 1973.

Lucien Côté and Michael D. Crutcher. The basal ganglia. In Eric R. Kandel, James H. Schwarz, and Thomas M. Jessel, editors, *Principles of Neural Science*, chapter 42, pages 647–659. Appleton & Lange, third edition, 1991.

E. Courchesne, J. Townsend, N. A. Alkshoomoff, O. Saitoh, R. Yeung-Courchesne, A. J. Lincoln, H. E. James, R. H. Haas, L. Schreibman, and L. Lau. Impairment in shifting attention in autistic and cerebellar patients. *Behavioral Neuroscience*, 108:848–865, 1994.

Antonio R. Damasio. Aphasia. *New England Journal of Medicine*, 326:531–539, 1992.

Terrence Deacon. *The Symbolic Species*. Norton, 1997.

T.W. Deacon. Cortical connections of the inferior arcuate sulcus cortex in the macaque brain. *Brain Research*, 573:8–26, 1992.

J. E. Desmond and J. A. Fiez. Neuroimaging studies of the cerebellum: Language, learning and memory. *Trends in Cognitive Sciences*, 2:355–362, 1998.

Kenji Doya. Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10(6):732–9, 2000.

Kenji Doya. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, 12:961–974, 1999.

R. Dunbar. *Grooming, Gossip, and the Evolution of Language*. Faber and Faber, London, 1996.

R. Eckhorn, R. Bauer, W. Jordan, M. Bosch, W. Kruse, M. Munk, and H. J. Reitboeck. Coherent oscillations: A mechanism of feature linking in the visual cortex? *Biological Cybernetics*, 60:121–130, 1988.

J. Elman, L. Bates, M. Johnson, A. Karmiloff-Smith, and D. Parisi. Rethinking innateness, 1996.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Jerry A. Fodor. *The Modularity of Mind*. MIT Press, Cambridge MA, 1983.

Gottlob Frege. *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Nebert, Halle, 1879.

Gottlob Frege. Ueber sinn und bedeutung. *Zeitschrift fuer Philosophie und philosophische Kritik*, 1892.

S. R. Freidman-Hill, Lynn C. Robertson, and A. Treisman. Parietal contributions to visual feature binding: Evidence from a patient with bilateral lesions. *Science*, 269:853–855, 1995.

Joaquin M. Fuster. *Memory in the Cerebral Cortex*. MIT Press, 1995.

N. Geschwind. Disconnexion syndromes in animals and man. *Brain*, 88:237–294,585–644, 1965.

Claude Ghez. The cerebellum. In Eric R. Kandel, James H. Schwarz, and Thomas M. Jessel, editors, *Principles of Neural Science*, chapter 41, pages 626–646. Appleton & Lange, 3 edition, 1991.

Harold Goodglass. Word-finding deficits in aphasia. In Harold Goodglass and Arthur Wingfield, editors, *Anomia - Neuroanatomical and Cognitive Correlates*, chapter 1. Academic Press, 1997.

Gordon. *Brain Development*, pages 83–7, 2001.

Gray, Koenig, Engel, and Singer. Oscilatory responses in cat visual cortex exibit intercolumnar synchronization which reflects global stimulus properties. *Nature*, 1989.

Stefan Greyer, Massimo Matelli, Guiseppe Lupino, and Karl Zilles. Functional neuroanatomy of the primate isocortical motor system. *Anat Embryol*, 202:443–474, 2000.

Herbert Paul Grice. Meaning. *Philosophical Review*, pages 377–88, 1957.

Herbert Paul Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Speech Acts*, number 3 in Syntax and Semantics, pages 41–58. Academic Press, New York, 1975.

Patrick Grim and Trina Kokalis. Boom and bust: Environmental varialbility favors the emergence of communication. In Jordan Pollack, Mark Bedau, Phil Husbands, Takashi Ikegami, and Richard A. Watson, editors, *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALIFE9) Boston, Massachusetts September 12-15th 2004*, pages 164–169, Boston, Massachusetts, 2004. MIT press.

S. Grossberg. How does a brain build a cognitive code? *Psychological Review*, 87(1):1–51, 1980.

Frank H. Guenther. Neural modeling of speech production. In *speech motor control in normal and disordered speech - 4ht international speech motor conference*, pages 12–15, 2001.

T. A. Hackett, I. Stepniewska, and J. H. Kaas. Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *Journal of Comperative Neurology*, 394:475–495, 1998.

M. Haruno, T. Kuroda, K. Doya, K. Toyama, M. Kimura, K. Samejima, H. Imamizu, and M. Kawato. A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. *Journal of Neuroscience*, 24(7):1660–5, 2004.

Donald O. Hebb. *The Organization of Behaviour*. Wiley New York, 1949.

E. C. Hildreth and Christok Koch. The analysis of visual motion: from computational theory to neuronal mechanisms. *Annual Review of Neuroscience*, 10:477–543, 1987.

Elisabeth L. Hill and Uta Frith. Understanding autism: insights from mind and brain. *Phil. Trans. Royal Society London*, 385:281–289, 2003.

Bernhard Hommel. Automatic stimulus-response translation in dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, 24:1368–1384, 1998.

D. Hubel and T. Wiesel. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

Hiroshi Imamizu, Satori Miyauchi, Tomoe Tamada, Yuka Sasaki, Ryousuke Takino, Benno Puetz, Toshinori Yoshioka, and Mitsuo Kawato. Human cerebellar activity reflecting and acuired internal model of a new tool. *nature*, 403(6777):192–196, 2000.

M. Ito. Neurophysiological aspects of the cerebellar motor control. *Int J Neurol*, 7:162–176, 1970.

Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

Laurent Itti and Christoph Koch. Computational modeling of visual attention. *Nature Neuroscience Review*, 2:194–204, 2001.

P. B. Johnson and S. Ferreina. Cortical networks for visual reaching: intrinsic frontal lobe connectivity. *European Journal of Neuroscience*, 8:1358–1362, 1996.

E. G. Jones and T. P. S. Powell. An anatomical study of converging sensory pathways within the cerbrel cortex of the monkey. *Brain*, 93:793–820, 1970.

Michael Jordan and David E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

Timothy Justus. The cerebellum and english grammatical morphology: evidence from production, comprehension, and grammaticality judgments. *Journal of Cognitive Neuroscience*, 16(7):1115–30, 2004.

Hans Kamp. A theory of truth and semantic representation. Technical report, Center of Cognitive Science, University of Texas at Austin, 1980.

R. Kawagoe, Y. Takikawa, and O. Hikosaka. Reward-predicting activity of dopamine and caudate neurons–a possible mechanism of motivational control of saccadic eye movement. *Journal of Neurophysiology*, 91(2):1013–24, 2004.

Mitsuo Kawato. Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, (9):718–727, 1999.

Michael Klein and Aude Billard. Words in the cerebral cortex - predicting fmri-data. In *Proceedings of the 8th Joint symposium on neural computation - The brain as a dynamical system, San Diego*, 2001.

A. Knoblauch and G. Palm. Scene segmentation by spike synchronization in reciprocally connected visual areas. i.local effects of cortical feedback. *Biological Cybernetics*, 87(3): 151–167, 2002a.

A. Knoblauch and G. Palm. Scene segmentation by spike synchronization in reciprocally connected visual areas. ii.global assemblies and synchronization on larger space and time scales. *Biological Cybernetics*, 87(3):168–184, 2002b.

Christof Koch. *The Biophysics of Computation: Information Processing in Single Neurons*. Oxford University Press, 1999.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernatics*, 43(59-69), 1982.

A.K. Kreiter and W. Singer. Stimulus-dependent synchronization of neuronal responses in the visual cortex of the awake macaque monkey. *Journal of Neuroscience*, 16(7):2381–96, 1996.

J. R. Lackner and P. Dizio. Rapid adaption to coriolis force pertubations of arm trajectory. *Journal of Neurophysiology*, 72:299–313, 1994.

George Lakoff. *Woman, Fire, and Dangerous Things*. University of Chicago Press, 1987.

R. Langacker. *Foundations of Cognitive Grammar*, volume 1. Standford University Press, 1987.

Willem J. M. Levelt. Producing spoken language: a blueprint of the speaker. In Colin M. Brown and Peter Hagoort, editors, *The neurocognition of language processing*. Oxford University Press, 1999.

Willem J. M. Levelt. *speaking - from intention to articulation*. MIT - Press, 1989.

William Levelt, Peter Prammstra, Antje S. Meyer, Paeiri Helenius, and Riitta Salmelin. An meg study of picture naming. *Journal Cognitive Neuroscience*, 10:5:553–567, 1998.

David Levine and Eric Sweet. The neuropathological basis of broca's aphasia. In Michael A. Arbib, David Caplan, and John C. Marshall, editors, *Neural Models of Language Processes*, Perspectives in Neurolinguistics, Neuropsychology, and Psycholinguistics, chapter 14, pages 299–326. Academic Press, 1982.

A. M. Liberman and I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 1985.

O. Lichtheim. On aphasia. *Brain*, 7:443– 484, 1884.

Maria Grazia Maioli, Salvatore Squatrito, Boaz Gedaliahu Samolsky-Dekel, and Eugenio Riva Sanseverino. Corticocortical connections between frontal periarcuate regions and visual areas of the superior temporal sulcus and the adjoining inferior parietal lobule in the macaque monkey. *Brain Research*, 798:118–125, 1998.

Peter Marien, Sebastiaan Engelborghs, Franco Fabbroc, and Peter P. De Deyn. The lateralized linguistic cerebellum: A review and a new hypothesis. *Brain and Language*, 79 (3):580–600, 2001.

B. M. Mazoyer, N. Tzouri, V. Frak, A. Syrota, N. Murayama, O. Levrier, G. Salamon, S. Dahaene, L. Cohen, and J. Mehler. The cortical representation of speech. *J Cogn Neurosci*, 5:467–479, 1993.

A. D. Milner and M. A. Goodale. *The Visual Brain in Action?* Oxford University Press, 1995.

Marco Mirolli and Domenico Parisi. Language, altruism, and docility: How cultural learning can favour language evolution. In Jordan Pollack, Mark Bedau, Phil Husbands, Takashi Ikegami, and Richard A. Watson, editors, *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALIFE9) Boston, Massachusetts September 12-15th 2004*, pages 182–187, Boston, Massachusetts, 2004. MIT press.

Richard Montague. *Formal Philosophy: Selected Papers*. Yale University Press, New Haven, 1973.

M. M. Muller and T. Gruber. Induced gamma-band responses in the human eeg are related to attentional information processing. *Visual Cognition*, 8:579–592, 2001.

U. Noppeney and C. J. Price. A pet study of stimulus- and task-induced semantic processing. *Neuroimage*, 15(4):927–935, 2002.

George A. Ojeman. Brain organization for language from the perspective of electrical stimulation mapping. *Behavioural and Brain Sciences*, 2:189–230, 1983.

P-Y Oudeyer and Frederic Kaplan. Intelligent adaptive curiosity: a source of self-development. In Luc Berthouze, Hideki Kozima, Christopher G. Prince, Giulio Sandini, Georgi Stojanov, G. Metta, and C. Balkenius, editors, *Proceedings of the 4th International Workshop on Epigenetic Robotics*, volume 117 of *Lund University Cognitive Studies*, pages 127–130, 2004.

Guenther Palm. *Pattern Separation and Synchronization in Spiking Associative Memories and Visual Areas*. Springer, Berlin, Heidelberg, New York, 1982.

D. N. Pandya and H.G.J.M. Kuypers. Cortico-corical connections in the rhesus monkey. *Brain*, 13:13–36, 1969.

D. Papathanassiou, O. Etard, E. Mellet, L. Zago, B. Mazoyer, and N. Tzourio-Mazoyer. A common language network for comprehension and production: a contribution to the definition of language epicenters with pet. *Neuroimage*, 11(4):347–57, 2000.

W. Penfield and L. Roberts. *speech and brain mechanisms*. Princeton University Press, Princeton, 1959.

Peterson, Fox, Michael I. Posner, Mintun, and Marcus Raichle. Pet studies of the cortical anatomy of single word processing. *Nature*, 33:585–589, 1988.

D. C. Plaut and C. T. Kello. The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney, editor, *The emergence of language*, pages 381–415. Mahwah, NJ: Erlbaum, 1999.

Karl R. Popper. *Logik der Forschung*. Mohr Siebeck, 1934.

Cathy Price, R. J. S. Wise, E. Wartburton, C.J.Moore, K. Patterson, and D. Howard. Hearing and saying: The functional neuroanatomy of auditory word processing. *Brain*, 119:919–31, 1996.

Friedemann Pulvermueller. Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, 5:517–524, 2001.

Antonio Raffone and Gezinus Wolters. A cortical mechanism for binding in visual working memory. *Journal of Cognitive Neuroscience*, 13(6):766–785, 2001.

J. P. Rauschecker, B. Tian, and M. Hauser. Processing of complex sound in the macaques nonprimary auditory cortex. *Science*, 268:111–114, 1995.

E. R. Ritvo, B. J. Freeman, and A. B. Scheibel et al. Lower purkinje cell counts in the cerebella of four autistic subjects: initial findings of the ucla-nsac autopsy research report. *American Journal of Psychiatry*, 143:862–866, 1986.

Daria Riva and Cesare Giorgi. The cerebellum contributes to higher functions during development. *Brain*, 123:1051–1061, 2000.

Giacommo Rizzolatti and Michael A. Arbib. Language within our grasp. *Trends in Neurosciences*, 21(5):188–194, 1998.

Giacommo Rizzolatti, R. Camarada, M. Fogassi, M. Gentilucci, Giuseppe Luppino, and M. Matelli. Functional organization of inferior area 6 in the macaque monkey: Ii. area f5 and the control of distal movements. *Experimental Brain Research*, 71:491–507, 1988.

Lynn C. Robertson. Binding, spatial attention and perceptual awareness. *Nature Reviews Neuroscience*, 4:93–102, 2003.

Lynn C. Robertson, A. Treisman, S. R. Freidman-Hill, and Grabowecky. The interaction of spatial and object pathways: Evidence from balint's syndrome. *Journal of Cognitive Neuroscience*, 9:295–317, 1997.

Ardi Roelofs and Peter Hagoort. Control of language use: Cognitive modeling of the hemodynamics of stroop task performance. *Cognitive Brain Research*, 15:85–97, 2002.

Eleanor Rosch. Natural categories. *Cognitive Psychology*, 4:328–350, 1973.

Eytan Ruppin and James A. Reggia. Patterns of functional damage in neural network models of associative memory.

Sophie K. Scott, C. Catrin Blank, Stuart Rosen, and J. S. Wise. Identification of a pathway for intelligible speech in the temporal lobe. *Brain*, 123:2400–2406, 2000.

B. Seymour, J. P. O'Doherty, P. Dayan, M. Koltzenburg, A. K. Jones, R. J. Dolan, K. J. Friston, and R. S. Frackowiak. Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–7, 2004.

R. Shadmehr and F. A. Mussa-Ivaldi. Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*, 14:3208–3224, 1994.

Lokendra Shastri and V. Ajjanagadde. From simple associations to systematic reasoning: A connectionist rpresentation of rules, variables, and dynamic binding using temporal synchrony. *Behavioural and Brain Sciences*, 16:417 – 451, 1993.

K. Smith, S. Kirby, and H. Brighton. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386, 2003.

Robert C. Stalnaker. Assertation. In P. Cole, editor, *Syntax and Semantics vol. 9: Pragmatics*, pages 315–322. Academic Press, New York, 1979.

Luc Steels. Perceptually grounded meaning creation. In M. Tokoro, editor, *Proceedings of the International Conference on Multi Agent Systems*, pages 338–344. AAAI Press, 1996.

Luc Steels. Language games for autonomous robots. *IEEE Intelligent systems*, pages 16–22, 2001.

Laurie A. Stowe, Anne M. J. Paansb, Albertus A. Wijersc, and Frans Zwartsd. Activations of "motor" and other non-language structures during sentence comprehension. *Brain and Language*, 2004.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning - An Introduction*. MIT Press, 1998.

M.A. Tagaments and B. Horwitz. Integrating electrophysiological and anatomical data to create a large scale model that simulates a delayed match-to-sample human brain imaging study. *Cerebral Cortex*, 8:310–320, 1998.

Tomoe Tamada, Satori Miyauchi, Hiroshi Imamizu, Toshinori Yoshioka, and Mitsuo Kawato. Cerebro-cerebellar functional connectivity revealed by the laterality index in tool-use learning. *Neuroreport*, 10:325–331, 1999a.

Tomoe Tamada, Satori Miyauchi, Hiroshi Imamizu, Toshinori Yoshioka, and Mitsuo Kawato. Activation of the cerebellum in grip force load force coordination: an fmri study. *Neuroimage*, 6:492, 1999b.

Saori C. Tanaka, Kenji Doya, Go Okada, Kazutaka Ueda, Yasumasa Okamoto, and Shigeto Yamawaki. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7(8):887–893, 2004.

Y. Tanaka, A. Yamadori, and E. Mori. Pure word deafness following bilateral lesions: a psychophysical analysis. *Brain*, 110:381–403, 1987.

Michael Tomasello. First steps towards a usage-based theory of language acquisition. *Cognitive Linguistics*, 11:61–82, 2000.

Michael Tomasello. *Constructing a Language - A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003.

A. Treisman. Features and objects in visual procesing. *Scientific American*, 255(5), 1986.

A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

Mermet Turgut. Transient "cerebellar" mutism. *Child's Nervous System*, 14:161–166, 1998.

L. G. Ungerleider and M. Miskin. Two visual pathways. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *In Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, 1982.

Christoph von der Malsburg. The correlation theory of brain function. Technical Report 81-2, Abteilung fuer Neurobiologie, MPI fuer biophysikalische Chemie, Goettingen, 1981.

Christoph von der Malsburg. Nervous structures with dynamical links. *Ber. Bunsenges. Phys. Chem*, 89:703–710, 1985.

Eike von Savigny. Sprachspiele und lebensformen: Woher kommt die bedeutung. In Eike von Savigny, editor, *Philosophische Untersuchungen*, volume 13 of *Klassiker Auslegen*, chapter 1, pages 7–40. Akademie Verlag, 1998.

K. Watanabe, J. Lauwereyns, and O. Hikosaka. Neural correlates of rewarded and unrewarded eye movements in the primate caudate nucleus. *Journal of Neuroscience*, 23(31): 10052–7, 2003.

Thomas Wennekers and Guenther Palm. On the relation between neural modelling and experimental neuroscience. *Theory in Bioscience*, 116:273–289, 1997.

Carl Wernicke. *Der aphasische Symptomemkomplex*. Cohn und Weigert, Breslau, 1874.

C. M. Wessinger, M. H. Buonocore, and C. L. Kussmaul amd G. R. Mangun. Tonotopy in human auditory cortex examined with functional magnetic resonance imaging. *Human Brain Mapping*, 5:18–25, 1997.

C. M. Wessinger, J. VanMeter, J. VanLare, J. Pekar, and J. P. Rauschecker. Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *JCN*, 13:1:1–7, 2001.

R. S. Williams, S. L. Hauser, D. P. Purpura, and et al. Autism and mental retardation: neuropathologic studies performed in four retarded persons with autistic behaviour. *Arch Neurol*, 37:749–753, 1980.

Richard J.S Wise, Sophie K. Scott, Catrin Blank, Cath J. Mummery, Kevin Murphy, and Elisabeth Warburton. separate neural subsytems within wernicke's area. *Brain*, 124: 83–94, 2001.

Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, 1953.

Daniel M. Wolpert, Kenji Doya, and Mitsuo Kawato. A unifying computational framework for motor control and social interaction. *Phil. Trans. R. Soc. Lond.*, 358:593–602, 2003.

Huadong Xiang, Chongyu Lin, Xiaohai Ma, Zhaoqi Zhang, James M. Bower, Xuchu Weng, and Jia-Hong Gao. Involvement of the cerebellum in semantic discrimination: An fmri study. *Human Brain Mapping*, 18(3):208 – 214, 2003.

R.J. Zatorre, A.C. Evans, E. Meyer, and A Gjedde. Lateralisation of phonetic and pitch discrimination in speech processing. *Science*, 256:846–9, 1992.

S. Zeki. The representation of color in the cerebral cortex. *Nature*, 284:412–418, 1980.

Edgar B. Zurif. The use of data from aphasia in constructing a performance model of language. In Michael A. Arbib, David Caplan, and John C. Marshall, editors, *Neural Models of Language Processes*, Perspectives in Neurolinguistics, Neuropsychology, and Psycholinguistics, chapter 9, pages 203–207. Academic Press, 1982.