The Brain Acquires Internal Models for the Use of Language

Michael Klein and Hans Kamp
Institute for Natural Language Processing, Stuttgart University
Guenther Palm
Department of Neural Information Processing, Ulm University
Kenji Doya
ATR Computational Neuroscience Laboratories

Abstract

This article introduces a biologically motivated neural architecture which is capable of learning to use language by selecting utterances suitable for pursuing a specific goal. We adapted the concept of the internal model (from control theory) to the problem of language use, and, in combination with reinforcement learning, use it to model utterance selection. We also use internal models to approach the problem of understanding intention in language. By means of simulation experiments in a multi-agent environment we show that our architecture is capable of deciding when to speak (instead of performing some non-verbal action), whom to address and what to say, as well as understanding the intention behind the utterance of another speaker. Further, we discuss data from functional imaging and electrophysiological experiments pointing 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, we propose functional imaging experiments which could confirm our hypotheses about the localization of the major components of our model in the brain.

INTRODUCTION

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 involves more than the mere production of an intelligible utterance or the truthful description of a state of affairs (Austin, 1961). Hence, at the beginning of every act of communication there is the speaker's goal and the speaker then applies his linguistic skills to transform this goal into an utterance. Such a transformation needs to involve a process which can compute what kind of linguistic construction is the most useful for pursuing this goal. So far, however, linguistic skills have been investigated in isolation from the non-linguistic goals of language production. Because of this, little is known about the neural bases of this transformation process. Apart from the selection of the appropriate linguistic constructions, such a transformation also must include the decision which goals can be pursued by utterances in the first place (instead of by non-verbal actions or not at all) and, in case of verbal actions, to whom these utterances should be addressed. Furthermore, since the work of Grice (1957) it is a widely accepted view on verbal communication 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 considered successful. So, while the speaker's goal stands at the beginning of the communication process, its recognition by the addressee is the final result of that process. There is also very little knowledge about the neural bases of understanding the communicative goals of other speakers. 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 brain structures.

In this paper, we present a neural architecture using *value functions* and *internal models* for the goal-directed and context-dependent selection of utterances. We 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. Using this architecture we have designed a neural model which can learn to select non-verbal actions and verbal actions (utterances). We tested this model by means of an implementation involving simulated agents in an environment in which they had to sustain themselves by collecting different kinds of food. The agents could use verbal actions (requests) to obtain particular food items from each other. In this simulated environment, agents learned when to speak (instead of performing some non-verbal action), whom to address and what to say, as well as to understand the intention behind the utterance of another speaker. We ran a battery of tests during which agents learned which actions (verbal and non-verbal) are appropriate in which situations, learned which goals can be accomplished by which verbal action in which context. We also tested in how far the efficiency of language learning is changed if agents learn by observing other agents in addition to trial and error learning. Finally, we tested the general benefits of communication in the simulated environment.

So far, value functions and internal models have been investigated as the basis of the goal-directed selection of motor actions (including speech motor control) in the brain (Doya, 1999, 2000). We adapt this usage of forward models and value functions to language because (despite the many differences there might be between verbal and non-verbal actions in terms of their complexity and with regard to how and where they are processed in the brain) in either case selection of a particular action must always be in relation to a goal as well as to a context in which the goal is to be achieved. Therefore, they might also have similarities in terms of the mechanisms and brain structures involved. It has been proposed that such similarities also exist between the processes of understanding people's intentions when they use verbal and when they use non-verbal actions (Wolpert et al., 2003). This proposal is supported by the finding that understanding the intentions behind non-verbal actions ontogenetically precedes the capacity to understand the intentions behind the verbal actions of other people (Tomasello, 2003). There is also evidence supporting the view that this ontogenetic precedence mirrors a phylogenetic precedence (Rizzolatti and Arbib, 1998, Arbib, 2000). There are a number of recent neurophysiological and imaging studies of 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). We discuss these studies in relation to our model and make a suggestion as to where the value function and the internal model are located in the brain. The evidence points towards an involvement of especially the basal ganglia in those computations which the present account attributes to the value function (Breiter et al., 2001, Seymour et al., 2004, Haruno et al., 2004, Tanaka et al., 2004, Watanabe et al., 2003, Kawagoe et al., 2004) and to an involvement of the cerebellum in those which the present account attributes to the forward model (Ito, 1970, Kawato, 1999, Doya, 1999, Imamizu et al., 2000). The model, therefore, is able to 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 that lesion in these areas tend to have on those processes. It can also be used to interpret related functional imaging studies. And, finally, it can be used to design studies to test the proposed relation of the main components of our architecture to real brain structures.

THEORETICAL FRAMEWORK

In this section, we will describe the value function and the forward model in mathematical detail. Then we will show how these two components are combined to select verbal and non-verbal actions. Further, we will explain how the forward model is used to understand the intentions behind utterances. Finally, we will describe the learning algorithms used to train the value function and the forward model.

Value Function

Instead of the term *goal* we use the terms *desire* and *intention*. This is because we want to distinguish states of the world which the agents know to be beneficial for themselves (desired states) from states of the world which they are actually trying to reach by some action or utterance (intended states). In our theoretical framework an agent has many desires. However, only some of these desires actually become intentions[1]. This happens

[1]Our model generally has a high degree of abstraction and idealization. For instance, we use *desire* and *intention* as technical terms here and do not claim these to closely resemble *real* desires and intentions. The analogy with real desires and intentions should nevertheless not be hard to see. Also, the *language*

when the agent tries to fulfill a desire by performing some suitable action. In our model, desires are represented 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 performed in order to bring this state about. The values of the desired states are estimations of how good it is for an agent to be in a certain state. The values are positive or negative real numbers expressing how much an agent desires a particular state s_t. A function V mapping every state onto such a value is called a *value function* (equation 1).

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

$V^{\pi}(s_t)$ is the estimation of the value of state s_t at (discrete) time step t under a *policy* π (Sutton and Barto, 1998). Here, π is a mapping from states s and actions a to the probability $\pi(s,a)$ of performing action a when in state s. $V^{\pi}(s_t)$ is defined in terms of the expected sum of discounted rewards r. The expected value is taken with respect to the Markov chain $\{s_{t+1}, s_{t+2}, ...\}$ where the probability of transition[2] from state s_{t+k} to s_{t+k+1} is given by π. Future rewards are *discounted* by the discount factor γ. The higher the value of γ, the more importance is given to later rewards, i.e. the less they are discounted (see Sutton and Barto (1998) for a more detailed explanation of the formula and the theory that goes with it).

The value function allows us to determine the *desired state* of every agent in every state: the desired state is the state with the highest value. However, not every state can be reached from every other state. In fact, apart from the context state s_t only those few states are accessible which can be produced from s_t through a single action in a single time step. Therefore, the value function only needs to compute the value of those states which can be reached from the current state.

---

used in our simulations is only a simple form of acquired symbolic communication.

[2]In multi-agent systems (which we use in our experiments), the subsequent state of the world does not only depend on the action a_t, of a given agent but also the actions of the other agents. As long as these actions of the other agents are not (yet) deterministic, which is usually the case while they are still learning, the values of the states are unstable; and so is the policy π which depends on the value of states. In general convergence of the value function of an agent depends on convergence of the policies of the others. In case these other policies do converge, then the given agent's value function may converge as well, and with it his own policy.

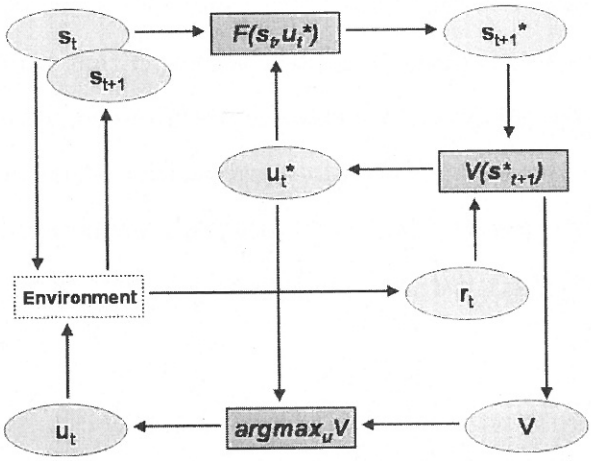

Figure 1: A simplified graph of how utterances (and other actions) are selected in our model: The forward model predicts the context-dependent utterance effects and the value function determines whether the utterance is selected on the basis of how desirable the effect is.

Forward Model

Given the value function, verbal and non-verbal actions can be selected in the light of how much the *consequence* of the action is desired. But to estimate the value of the consequences of actions, another device is required in order to predict these consequences. One such device is what is known as an *internal model*, or, more particularly a *forward* model (Jordan and Rummelhart, 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, we 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 verbal action (i.e. utterances) is not deterministic but probabilistic. This creates additional complications, as it also affects action selection - do we rather select an unlikely effect which is very much desired or a likely effect which is desired to a somewhat lesser degree? However, in this article we limit ourselves to the general principles of the acquisition and utilization of internal models in

language use and do not deal with probabilistic effects. Also, we ignore the fact that the effects of verbal actions on the observable world are indirect. They are achieved by a direct (though only indirectly observable) effect on the mental state of the addressee. To account for this phenomenon 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 various complications. F simply predicts a subsequent state s^*_{t+1} based on a current state s_t (context) and a possible non-verbal or verbal action (utterance) u^*_t.

$$s^*_{t+1} = F(s_t, u^*_t) \tag{2}$$

Part of acquiring a language is to learn what subsequent states an utterance can produce. This learning involves, among other things, observing the utterances of other speakers, the contexts in which they are made and the consequences they produce, as well as the capability to generalize from their utterances to the learner's own utterances. Furthermore, the learner can observe the effects of her own utterance and use this for the fine tuning of her internal model of language use, e.g. she can experience whether or not a certain utterance in a particular context produces the predicted effect. Irrespective of whether the learner observes or speaks herself, in both cases the experienced context-dependent effects of the utterances can be used to train the internal model.

Selection of Verbal and Non-verbal Actions

Given the forward model F, utterances are selected by means of a function $\arg\max_u$ which selects the verbal action that produces the most desirable state (equation 3).

$$u_t = \arg\max_u [r(s_t, u^*_t) + V(F(s_t, u^*_t)] \tag{3}$$

This function returns that one from all possible u^*_t's 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. The immediate reward function $r(s_t, u_t^*)$ of equation 3 which reflects the cost of the given action u_t^* is neglected in our model since this cost is equal to 0 for all possible verbal or non-verbal actions.

Since $\pi(s, u)$ can be determined on the basis of the function described in equation 3, we will, for the rest of this article, no longer talk about π, but only about the forward model and the value function.

Understanding Intentions

Addressees also use their forward model to understand the communicative intentions of other speakers. If the forward model is applied to the utterance with which they were addressed and to the context in which the utterance was made, the result will be 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

The value function is implemented as a neural network. To train this network we used *TD(0) reinforcement learning* (Sutton, 1988). In TD-learning, the so-called TD-error gives the distance from the correct prediction and the direction of the deviation. Thus, it 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 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).

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{4}$$

$$V(s_t) \leftarrow V(s_t) + \alpha\delta \tag{5}$$

The forward model, mapping utterances and game-states (contexts) onto game-states

(consequences) is implemented by a single-layer perceptron. We used 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 (equation 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 7). Then Δw_{ik} is used to update the weight (equation 8).

$$e_k = y_k^* - y_k \tag{6}$$

$$\Delta w_{ik} = \alpha e_k x_i \tag{7}$$

$$w_{ik} \leftarrow w_{ik} + \Delta w_{ik} \tag{8}$$

From this theoretical framework we 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 pre-programmed forward model, and a set of verbal and non-verbal actions can learn to behave in an optimal way, employing verbal and other actions as appropriate. (ii) In such an environment an agent equipped with an optimal value function and a pre-programmed 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.

THE ACQUISITION ENVIRONMENT

We test our hypotheses about language acquisition and communication in a simulation of a multi-agent game. The goal in this game is to obtain food through verbal and non-verbal action. In this simulation, *food* grows in certain intervals on *trees* (how this time interval is calculated is explained in the appendix). There are three trees $T_1 ... T_3$, growing three types of food. Every tree T_i can hold maximally 5 pieces of food. Therefore, T_i can be represented by a 5-dimensional binary vector. The six possible states of such a tree vector, coding the number of food pieces (0 ... 5), is shown in equation 9.

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \tag{9}$$

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 5×3 matrix. The coding of the number of pieces of every food type in the agent's store is analogous to the coding of the trees.

Thus, a state s of the game is represented as a 6-tuple of three 5×3 binary matrices and three 5-dimensional binary vectors (equation 10).

$$s =< A_1, A_2, A_3, T_1, T_2, T_3 > \tag{10}$$

The agents interact with the world through their perceptions and their non-verbal actions and with each other through perception and with actions which can be either verbal or non-verbal. 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 that action a_{t_i} produces in the state s_{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, which is capable of storing game states for a constant number m of time steps (see appendix for the values of parameters).

Within a certain time interval (d_o) invariably one piece of food gets *digested*, i.e. it disappears. Once the total amount of food in the game is below the threshold n_o, 3 pieces of food grow simultaneously on one of the three trees. Because of this design, the agents cannot afford to rest once they have gained a sufficient amount of food items. Agents never starve to death, but for every time step during which they do not have any food they get a very negative reward.

Agents can perform one of the following 16 actions:

- harvest a tree, i.e. collect all its food (3 possibilities)

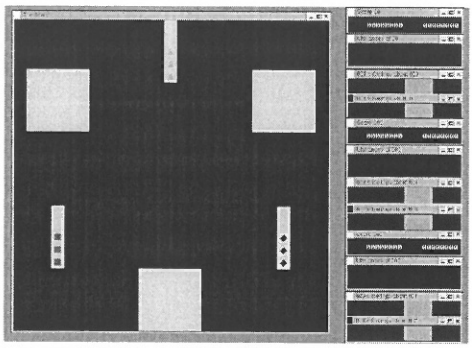

Figure 2: This shows the initial game state. The narrow upright rectangles are the trees, each holding 3 pieces of food. The grey squares are the agents. They have the capacity of storing 5 pieces of each food type. The bar on the right displays scores, utterances, and other useful information.

- give one piece of food to another agent (2 other agents $\times$ 3 food types = 6 possibilities)
- ask another agent for a piece of food of a certain type (2 other agents $\times$ 3 food types = 6 possibilities)
- *no action* (1 possibility)

In certain situations an agent cannot improve his current situation by any of the actions the game permits. Because such situations can arise, we do not require that each agent performs one of the actions open to him at each time step; an option that is available to all agents at all times (unless they are asked for a food item) is that of not performing any action at that time.

At each transition between two successive times, only one agent can perform an action. This agent can perform either 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 while the addressee gives (or fails to give) the desired object to the speaker.

The goal of the agents in the game is to have at least 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.

Utterances in this game are always utterances of a single word. An utterance of an agent is defined by its content (i.e. the word which is used), its speaker, and its addressee. 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*. The words are coded as natural numbers (*triangle* = 1, *square* = 2, *diamond* = 3; see appendix for more details on the coding of the utterance). An agent can only address one of the other agents, never both of them, but all agents 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 context of an utterance is the complete state of the game, as described above, at the time when the utterance is made.

To choose his non-verbal action or utterance, an agent computes the outcome of all possible actions in the present context with his forward-model. The forward model is pre-programmed 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 distinguishes two types of agents. The first type employs a pre-programmed dialogue system for understanding and producing utterances. These agents speak and act according to a set of pre-programmed algorithms which translate their desires into utterances and compute the desires of other speakers from the utterances they produce. The second type uses the neural network-based architecture described above. The first type of agents is used in simulations where the forward model for verbal actions is fixed (e.g. in simulations where the value function is trained). It is also used in our experiments on language acquisition. In these experiments agents of the second type (who make use of a neural network-based forward model) have to learn the context-dependent consequences of different utterances. In our simulations these consequences are assumed to be conventional within the simulated language community. This community has to consist of agents who already know and consistently apply the conventions of verbal behavior which the new agent has to learn. Therefore, the community is made of agents of the first type. Their knowledge about the linguistic conventions of the community is immanent in the pre-programmed algorithms. It is of no concern to these agents whether or how fast the new agent learns the conventions; none of their actions are aimed at teaching the new agent anything.

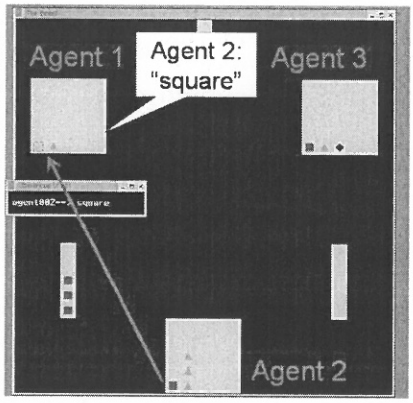

Figure 3: This shows an arbitrary state during the early stages of training. The last action of agent 1 was to ask agent 2 for the square. Obviously, this is not the best move. A better move would be to harvest the *square tree*, as with this action the agent would get 3 squares instead of one.

While agents of the first type form the language community, it is the agents of the second type whose learning and behavior we investigate in our experiments on language acquisition. This second type uses two kinds of data for learning. On one hand they learn by observing the context-dependent effects of utterances made by other speakers. On the other hand they learn by trial-and-error, i.e. by observing the context-dependent effects of their own utterances. Utterances are transformed into a neural representation before they are used as input for the neural network-based forward model (e.g. the word is then coded as 3 dimensional binary vector). The neural network-based forward model trained with this data is gradually improving its ability to map utterance - context pairs onto effects. The better the forward model is trained, the more effectively it can be used in 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). Chosen is that verbal or non-verbal action which will bring about the state with the highest value.

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 speaker's intention from the utterances and the context. In other words, the language learner *understands* the utterance insofar as

he considers what effect, according to his own experience, such an utterance should have in the present context. Using his knowledge of the present state of the game and his estimation of what state the speaker desires, the addressee then uses a pre-programmed algorithm to compute which action would bring this state about.

In the simulations reported in this article we have decided to make it a general policy of the addressee to cooperate. This means that when the addressee has a piece of the kind of food that the speaker asks him for he will give it to him even if it would be in his own interest to keep it. While pre-programmed agents are simply programmed to abide by this policy, the language learning agent cannot do more than act in accordance with what he takes to be the intention of the speaker addressing him. At first, when learning is just beginning the speaker's intention that the learning agent computes bears a purely random resemblance to the intention which the speaker has actually expressed. As learning progresses, the learner gets better at determining the speakers' intentions. But in any case he is programmed to always comply with what he identifies as the speakers' intention.

SIMULATIONS

The first set of simulations was carried out in order to train the value function of all three agents. During these simulations, agents used a pre-programmed forward model for all types of verbal and non-verbal actions, i.e. no forward models were trained during these simulations. The purpose of this study was to test whether by training their value functions agents would learn when to speak (instead of performing some non-verbal action), whom to address and what to say. The trained value functions of this set of simulations was then used in the next set of simulations during which only one agent (in the simulations reported here it is usually agent 2) trained his forward model for verbal actions. These simulations were performed to show that agents which use our architecture are able to acquire the linguistic conventions which are observed in their environment. We wanted to see whether agents could learn to use the appropriate utterances to pursue their goals and learn to understand the intentions behind utterances of other agents. In the next set of simulations we compared agents who trained their forward model for verbal actions only by trial and error learning with agents using who (in addition to trial-and-error-learning) also train it by observing the verbal actions of other agents. This comparison was done to

show the advantage of a model that allows learning by observation over a model that can learn by trial and error only. In the final set of simulations we tested how much agents actually benefit from communication. This was done by comparing the average score of agents in simulations where all agents use language with the average score of agents in simulations where no agents used language.

Value Function

During the training of the value function agents learn which states are desirable. During that training phase they use a pre-programmed forward model which computes the context-dependent consequences of all four kinds of actions (including the verbal actions). By applying the value function to the thus computed consequences, the agent is able to decide in which situations it is better to harvest, donate, speak, or simply do nothing.

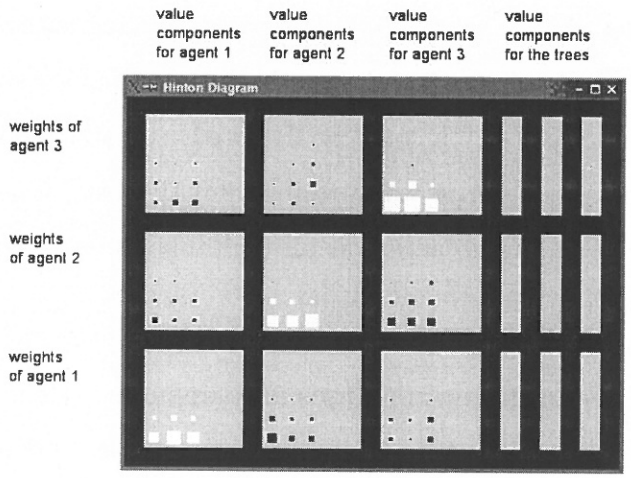

Figure 4: This shows the strength of the synaptic weights for the value function after 20 million time steps for a discount factor γ of 0.9. Every horizontal row represents the weights of one agent for computing the value function over the game state. White squares represent positive values, black squares represent negative values and the size of the square represents the absolute value.

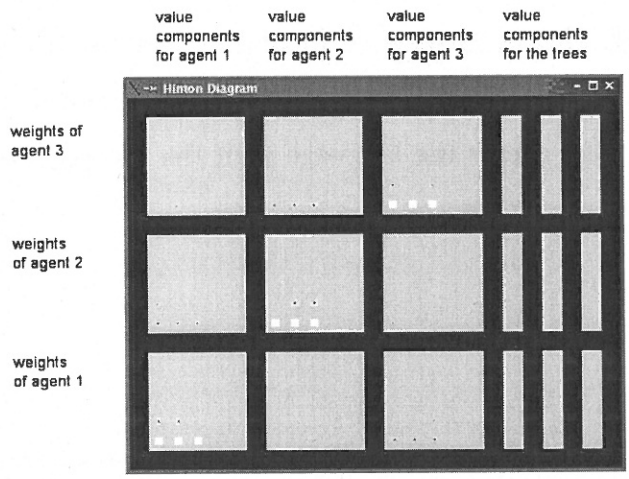

Figure 5: This figure shows the strength of the synaptic weights for the value function for a discount factor γ of 0.1 (20 million time steps).

At the very early stages of the training, agents selected *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 3, did occur in the intermediate stage of training. After training no more suboptimal action could be detected. Agents used language when appropriate, harvested trees whenever it was possible and the best action, and almost never used the option of performing no action at all (suitably so, as in most cases some action or request would improve the state of agent).

Figures 4 and 5 demonstrate the effect of the discount parameter γ. The value function of each agent estimates the value of a state by multiplying for every possible location of an object either 1 (if there is an object at this location) or 0 (if there is no object) with a weight for that location and by computing the sum of these products. The weight is an estimation of the positive or negative contribution of having an object at this location to the total value of the state. After training, as can be seen in figure 4, agents attribute a positive weight to objects in their own store and a negative weight to objects in the store of other agents. In simulations where the γ-parameter was set to 0.9 (and the agents therefore regarded future states as important) the agents were able to learn that it pays of in the longer run to have more than one object of a certain type, even though this is not immediately rewarded. This is shown by the positive weight given to more than the first location for objects of one type in the store of each agent (figure 4). In simulations were the γ-parameter was set to 0.1 (and the agents therefore hardly regarded future states as

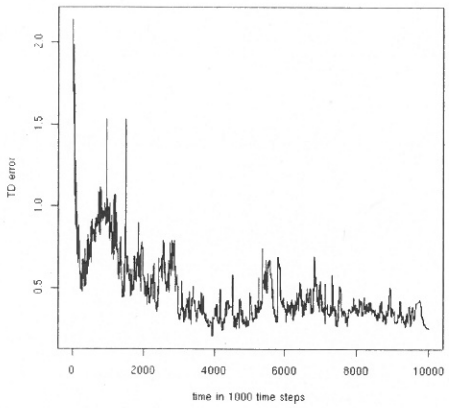

Figure 6: This is the development of the TD-error with $\gamma = 0.3$. The figure shows the average error of 5 runs.

important) a positive weight is given only to the first location of each object type (figure 5). A major difference in behavior that is caused by these differences in weights is that agents do not realize the advantage of harvesting (and thereby getting three objects of a kind) over asking for an object (and getting only one).

The TD-error decreased very fast in the beginning (figure 6) and good 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.

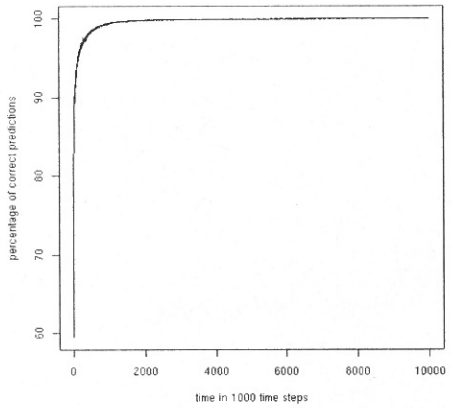

Figure 8: The prediction error of language learning changes over time. This graph shows the percentage of correct predictions for five runs (10 million time steps each). The error rate decreases very fast in the beginning and then slowly goes toward 0.

Language Learning

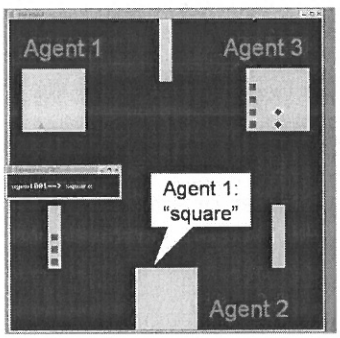

Figure 7: This shows an episode during a very early stage of language learning. The language learner (agent 2) asks agent 1 for the square, although agent 1 does not have one. This is an example of the language learner not being able to understand what effects this kind of utterance has in what kinds of contexts.

In the second type of simulation, a single learning agent is introduced into a community consisting of two other agents who already know and consistently apply the conventions of verbal behavior. The task for the learning agent is to acquire the context-dependent consequences of utterances. By observing the language use of the pre-programmed agents and by trial and error the learner trains his neural network-based forward model using the learning algorithm described above. This network represents the linguistic knowledge

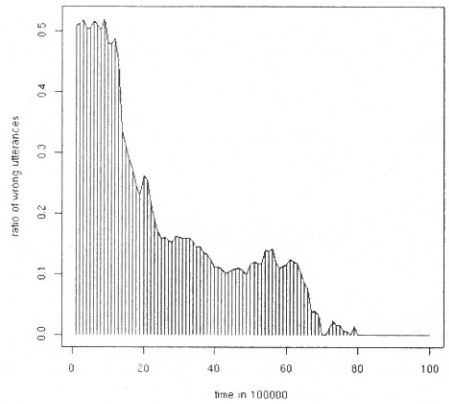

Figure 9: This figure shows the average ratio of wrong utterances for five simulations in which only trial-and-error learning and no learning by observation was used.

of the agent.

We have kept the learning situation simple by assuming that the learner's value function (the neural network which computes values of the states) is already fully trained - from the very start he is able to determine for any two possible states whether the first is preferable, from the perspective of his own interest, to the second. What he needs to learn is which verbal actions bring about which states.

At the beginning of this simulation, the verbal behavior of the language learner can best be described as a production of *random utterances*. The forward model cannot predict the effect of any utterance correctly, so the value function estimates values for the almost random states which the agent *believes* to be able to bring about with his utterances. For instance, an agent may in a situation when it would be the best action to get a square from agent 1 (and his value function could tell him that), address agent 3 and ask him for the triangle. Figure 7 illustrates another situation, where the language learner asks an agent for an object which the addressed agent does not even have. Such things can happen because in the early stages of training the learner has a very inaccurate mapping of context states and utterances onto subsequence states. The agent in this state might be described as knowing neither the conventional effects of an utterances, nor the context conditions for its successful use. But as learning progresses, the learner's utterances are more and more attuned to the situation and to his own needs.

The language understanding problem that confronts the beginning learner is as we de-

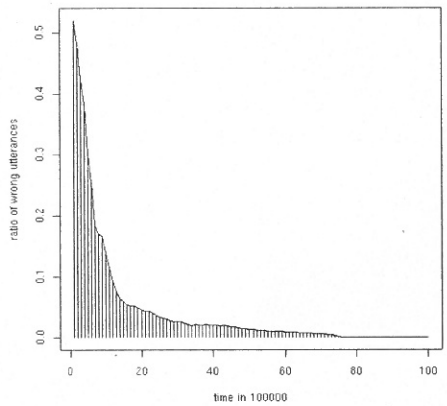

Figure 10: This figure the ration of wrong utterances for cases in which learning by observation is used in addition to trial and error learning.

scribed it earlier: In early stages of training it can be observed that the learning agent shows no reaction to the request made to him. The explanation for this behavior is that he misconstrues the speaker's intention and most of these misconstrued intentions cannot be reached from the current state by a single action. In the intermediate stages of training, when it can be observed that most of his reactions are in accordance with the utterances that are addressed to him, he sometimes shows wrong reactions. The explanation here is that the agent still occasionly misconstrues the speaker's intentions, but that in some rare cases these intended states of the speaker can be reached in one step. The learner therefore reacts with the corresponding action.

After training language production and comprehension was optimal and no difference between teacher and learner could be detected. As can be seen in figure 8, the prediction accuracy increased very fast to a level close to 100 %.

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 we 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 do learn by observation as well

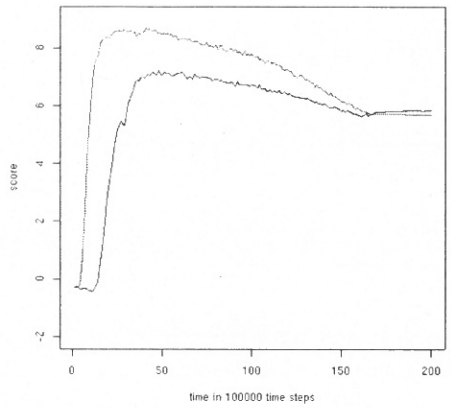

Figure 11: This figure shows the performance of a language learner which uses v-function and forward model, but no learning by observation (black line) in comparison to a language learner which uses v-function and forward model, and learns by observation as well (grey line). The lines represent the average development of the score for 5 simulation runs.

As a first measure of successful learning we counted the number of wrong usages of utterances, i.e. utterances which do not match the context conditions (e.g. an utterance in which an agent is asked for a certain object which he does not have). The ratio of wrong utterances for both types of agents is 0.5 in the beginning of the simulation (see figures 9 and 10). For agents who do not use learning by observation the ratio stays at this level until 10 million times steps. At 50 Million it is still around 0.1 and it finally reaches 0.0 at 80 million time steps. For agents who do use observation learning, the ration drops very fast in the beginning, falling to below 0.2 after 10 million time steps. After 50 million time steps the ratio for this type of agent is already below 0.05, reaching 0.0 before 80 million time steps. Thus, although the agents who use observation learning learn a lot faster, both types of agents are error free after approximately the same number of time steps.

We also compared the performance (in terms of score) of the two learning types. As can be seen in figure 11, 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.

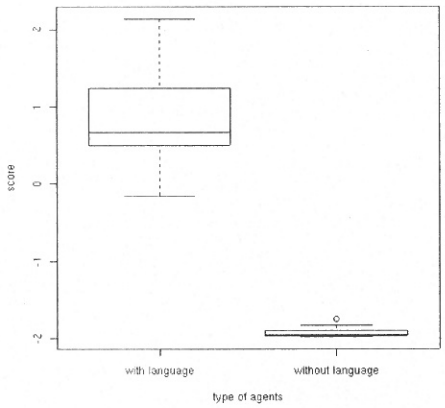

Figure 12: This shows the distribution of average scores of 10 runs. The left box shows the values for simulation in which the agents where allowed to communicate. The right box shows the value for simulations during which communications is suppressed.

Muted Agents

Although we constructed the simulation environment in such a way that it is to the agents' advantage to use language, we wanted to test how much agents gain by communication. Therefore, we modified the simulation in such a way that agents could no longer select verbal actions to fulfill their desires. We 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 12). Using the values of the 10 runs, we performed a paired t-test, which revealed that this difference is very significant ($p \ll 0.001$). The values obtained in this comparison also reflect the fact that the benefit of receiving objects from other agents and the advantage of communication are coupled in this simulations. The agents who are not allowed to use language have no other possibility of getting objects from each other - they cannot just take them or apply a form of non-verbal communication.

DISCUSSION

Summary of Simulation Results

We have demonstrated that in a game environment where only certain accomplishments are rewarded agents equipped with a value function (trained with reinforcement learning), a pre-programmed 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. We have also demonstrated that in such an environment an agent equipped with an optimal value function and a pre-programmed forward 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.

We have also compared our model with one that does not learn by observation. The slight difference between value function learning with and without observation can be explained by the extreme simplicity of the language used in this study. We predict that the more complex the language is, the greater 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 are 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. The forward model always uses complete world states as utterance contexts and maps pairs consisting of utterances and their context world states onto other world states. A more realistic model would include some form of conceptual knowledge as well as mechanisms for focusing, jointly with one's communication partner, on some small part of the total current state. This might allow a more condensed representation of currently relevant contexts and for more efficient computations which use these representations as input. In this way learning the effects of utterances in context would be much simpler and thus, presumably, also faster. However, optimization of learning time has not been a goal of this study. This is also the reason why we have not tested systematically how the learning curve is affected by the different ways of setting the values of parameters (such as the learning rate α or the discount factor γ). Because of this a systematic comparison between the performance of the model and that of children who acquire the command of new words would have been premature.

Our decision to make it a general policy of the addressee to cooperate is based on the fact that neither communication nor language acquisition would be possible if the members of a language community would not have a cooperative disposition towards each other, and especially towards a language learning child. This disposition appears to be in some part genetical and in some part acquired. How it emerges is not an issue of this article. Studies dealing with those questions are, for instance, Grim and Kokalis (2004) or Mirolli and Parisi (2004).

Basal Ganglia and the Value Function

The account of language use introduced in this paper proposes two major components: a value function to predict the values 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.

Evidence as to which brain areas are involved in those computations which the present account attributes to the value function comes from experiments that show which brain regions are activated in *reward prediction*. Reward prediction in those experiments generally involves the computations of the benefits of alternative actions. A very similar process can be attributed to the value function of our model, as it this function which is responsible for computing the benefits of the states which alternative actions can bring about. 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).

Cerebellum and the Forward model

Traditionally, the cerebellum is regarded as a brain structure concerned with motor control. This view derives mainly from studies of cerebellar patients. Connected with this is the prevalent opinion that the only role of the cerebellum in linguistic processing relates to speech motor control. However, functional imaging and other studies have shown 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). The role of the cerebellum in linguistic processing beyond the level of motor control has not yet been properly understood. We conjecture that the cerebellum is an important component in the neural implementation of the complex internal model that predicts context-dependent utterance effects (which is indispensable for appropriate language use) and that these cerebellar activations represent the usage aspects of the linguistic skills tested in those experiments.

Based on neurophysiological data Ito (1970) already 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 regard the cerebellum as the neural site of an internal model used in speech motor control (Guenther, 2001). Moreover, there is evidence that the domain of the internal models acquired and represented by the cerebellum is not limited to (speech) motor control, but that the cerebellum is generally the neural structure acquiring and representing internal models; and that internal models required for higher linguistic functions, such as the one used in our account, are also (at least to a significant extent) represented in the cerebellum. The evidence, summarized by Doya (1999), consists of anatomical, physiological and computational data which suggest that major brain structures cannot be distinguished in terms of the cognitive tasks they perform (motor control, language, memory, attention etc.), but in terms of the methods by which they compute and learn. 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).

Furthermore, 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, as well as to compute the goal expressed by an utterance of another speaker. A patient without such an internal model (or with an impaired one) could still have certain basic linguistic abilities (e.g. he might be able to name objects). However, he would not be able to predict the effects of his utterances and therefore would not be able to use language appropriately. He would also not be able to understand the intention behind an utterance. This pragmatic aspect of language is known to be impaired in patients suffering from a *high functioning*

variant of autism known as Asperger syndrome[3]. Inability, or reduced ability, to use language effectively in communication is a feature common to all autistic patients. But while *low functioning* Autists are marked by a severe impairment of the command of language generally, Asperger patients tend to have a normal command of the structured properties (phonological, syntactical, semantic) of language. Yet they too fail to *use* this language capacity to engage in interactive communication and they too fail to understand intentions behind utterances (Tager-Flusberg, 1999).

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 the frontal lobes, the parietal lobes, the amygdale and the brain stem. Furthermore, 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, 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 on the one hand to agrammatism and on the other hand to disorders comparable to those of autism (e.g. absence of spontaneous language for the purpose of communication).

[3]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).

Involvement of Other Brain Areas

The cerebellum alone appears to be a short term prediction device, i.e. by itself it can only predict the consequences of actions (motor commands) if these effects are not delayed for more than approximately 200 ms. Hence it seems that the cerebellum is not able by itself to acquire and predict the effects of utterances (which definitely do not fall into a 200 ms time frame). However, loops between cortical areas (such as the cortical language areas or the hippocampus) and the cerebellum could extend this limited time frame to a potentially unlimited time frame (using cortical memory devices) and, thus, handle these long term effects. Furthermore, an internal model predicting context-dependent utterance effects necessarily involve many other linguistic skills which are likely to be represented in cortical areas, 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). Functional connectivity between cortical areas (such as Broca's Area) and the cerebellum that was demonstrated in functional imaging studies (e.g. Tamada et al. (1999b)) supports this view.

Alternative Models of Language Use

In our approach agents are capable of generating their own goals on the basis of what they have learned about rewards given in certain states. Unlike any other approach known to us, the modeled system has to learn when it is useful to speak rather than perform a non-verbal action or do nothing at all. This is an important feature of humans and human communication. We humans are not simply reactive systems. We can initiate conversations, ask questions, when we need information, or talk just for social reasons. Models of language use which are not goal-directed cannot account for this important capability. Most current question answering systems, for example, use questions to trigger the search for some information, which is then possibly assembled into an utterance that the system produces as reply to the given question. In general, such systems cannot understand speaker's intentions, nor can they form appropriate intentions themselves. The ability to handle the goal-directedness of language in production and perception appears to be one of the most essential features of the human language faculty. Purely reactive systems (as most systems have been until now) are not therefore plausible models of the cognitive architecture of the human communication system.

Goal-directedness can also be modeled by simple reinforcement learning without a predictive component (such as an internal model). However, it has been argued by Chomsky and his colleagues that language cannot be learned by simple reinforcement learning (see e.g. Chomsky (1959)[4]). Indeed it seems plausible that while language learning involves much observation of language related behavior of others and of the effect that one's own utterances have on others, at least some of this is independent of ulterior motives such as the satisfaction of one's own desires or needs. To model this capacity a device is needed which can learn context dependent utterance effects independently of the reward these effects produce. In other words, the language faculty as it is implemented in the brain must be independent of knowledge or beliefs about the desirability of possible states of the world. Our theoretical framework, which - in contrast to simple reinforcement learning - uses the value function and the internal model as two interacting but separate components, vouchsafes this independence.

With simple reinforcement learning, utterances could be selected by a single function which predicts the reward of possible utterances in the current context. In contrast, the value function used in our account is, by itself, not sufficient for utterance selection; an internal model of the environmental dynamics is a necessary second component. Although the internal model used in the simulations reported in this article was a forward model, *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[5], i.e. they would calculate the necessary action 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. Even for an environment of a complexity comparable to the one in our study (which is still quite simple in comparison to the real world), such computations are very expensive.

The alternative we have chosen in our approach - the combination of value function and forward model - selects utterances by predicting the effect of all possible actions or

[4]These points are still controversial.

[5]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.

utterances and selecting the utterance or action which brings about the effect the agent desires the most. Of course, for actual natural languages it is an impossible task to compute the expected effects of all possible utterances. Since there are infinitely many of them, such a computation could not be carried out in finite time. If, however, 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 that the usage of language is learned in terms of relations between contexts, utterances and effects is also supported by research in language acquisition (Tomasello, 2000, 2003).

Predictions by the Model and Experiments to Test Them

In motor control, forward models appear to be necessary, since most 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 cannot be based on sensory feedback, because the consequences of utterances cannot be perceived until a reaction of the environment has taken place. Therefore, it is likely that utterance selection involves predictions of how the environment will react and that these predictions are made by the speaker's internal model.

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 in which the dynamic characteristic of the arm was changed by a force field (Shadmehr and Mussa-Ivaldi, 1994, Lackner and Dizio, 1994). As action selection is done using a internal model, motor commands continued to be selected on the basis of the relation between motor commands and effects that was stored in the internal model (which was no longer appropriate for the changed dynamics), and so it took some time until the model was adapted to the new dynamics. Thus, movement was distorted. When the force field was removed, the movement was temporarily distorted once more.

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.

Such an experiment could also be executed in a PET or an fMRI scanner. In comparison to studies investigating neural activation connected with the operation of an internal model in motor control (Tamada et al., 1999a, 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 with processing the difference between predicted and perceived effects will be found in the cerebellum.

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 (as in simple reinforcement learning) 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. In the light of this it would be interesting to see whether the activation in the striatum could be reproduced in settings where the reward is given for states independently of the actions chosen to bring them about. This could be done by creating experimental situations in which the subject would have to predict the reward for an action a in a context state s_1, given that (i) she has already learned that action a performed in s_1 brings about state s_2 (e.g. by observing another agent bringing about this effect), (ii) she has not yet learned which reward is given for action a selected in context s_1, and (iii) she has learned that in state s_2 a certain reward r is given. If in such a situation the subject would be able to predict the reward correctly, a value function is necessarily involved. In a similar study, verbal and non-verbal actions could be used to bring about certain states. This would allow us to see whether the activation correlating with reward prediction is the same for verbal and non-verbal actions, as is predicted by our model.

CONCLUSION

In this article we 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. We 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 at the present time it appears that alternative approaches fail to explain some essential properties of language use and language acquisition. 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.

ACKNOWLEDGMENT

This work was supported by the German Academic Exchange Service (DAAD), the German Research Foundation (DFG), and the National Institute of Information and Communications Technology of Japan (NICT). We would like to thank Helmut Schmidt for comments on the manuscript.

References

Arbib, M. A. (2000). The mirror system, imitation, and the evolution of language. In Nehaniv, C. and Dautenhahn, K., editors, *Imitation in Animals and Artefacts*. MIT Press.

Austin, J. L. (1961). *Philosophical Papers*. Oxford University Press.

Bailey, A., Luthert, P., Dean, A., Harding, B., Janota, I., Montgomery, M., Rutter, M., and Lantos, P. (1998). A clinicopathological study of autism. *Brain*, 121:889–905.

Baron-Cohen, S. (2004). The cognitive neuroscience of autism. *Journal Neurol. Neurosurg. Psychiatry*, 75:945–948.

Bauman, M. L. and Kemper, T. L. (1994). Neuroanatomic observations of the brain in autism. In Bauman, M. L. and Kemper, T. L., editors, *The neurobiology of autism*, pages 119–145. John Hopkins University Press.

Bauman, M. L. and Kempner, T. (1985). Histoanatomic observation of the brain in early infantile autism. *Neurology*, 35:866–874.

Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., and Shizgal, P. (2001). Function imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, 30:619–639.

Chomsky, N. (1959). A review of b. f. skinner's verbal behavior. *Language*, 35(1):26–58.

Courchesne, E., Townsend, J., Alkshoomoff, N. A., Saitoh, O., Yeung-Courchesne, R., Lincoln, A. J., James, H. E., Haas, R. H., Schreibman, L., and Lau, L. (1994). Impairment in shifting attention in autistic and cerebellar patients. *Behavioral Neuroscience*, 108:848–865.

Desmond, J. E. and Fiez, J. A. (1998). Neuroimaging studies of the cerebellum: Language, learning and memory. *Trends in Cognitive Sciences*, 2:355–362.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, 12:961–974.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10(6):732–9.

Gordon (2001). *Brain Development*, pages 83–7.

Grice, H. P. (1957). Meaning. *Philosophical Review*, pages 377–88.

Grim, P. and Kokalis, T. (2004). Boom and bust: Environmental varialbility favors the emergence of communication. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T., and Watson, R. A., 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. MIT press.

Guenther, F. H. (2001). Neural modeling of speech production. In *speech motor control in normal and disordered speech - 4ht international speech motor conference*, pages 12–15.

Haruno, M., Kuroda, T., Doya, K., Toyama, K., Kimura, M., Samejima, K., Imamizu, H., and Kawato, M. (2004). 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.

Hill, E. L. and Frith, U. (2003). Understanding autism: insights from mind and brain. *Phil. Trans. Royal Society London*, 385:281–289.

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Puetz, B., Yoshioka, T., and Kawato, M. (2000). Human cerebellar activity reflecting and acuired internal model of a new tool. *nature*, 403(6777):192–196.

Ito, M. (1970). Neurophysiological aspects of the cerebellar motor control. *Int J Neurol*, 7:162–176.

Jordan, M. and Rummelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354.

Justus, T. (2004). The cerebellum and english grammatical morphology: evidence from production, comprehension, and grammaticality judgments. *Journal of Cognitive Neuroscience*, 16(7):1115–30.

Kawagoe, R., Takikawa, Y., and Hikosaka, O. (2004). 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.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology*, (9):718–727.

Lackner, J. R. and Dizio, P. (1994). Rapid adaption to coriolis force pertubations of arm trajectory. *Journal of Neurophysiology*, 72:299–313.

Levelt, W. J. M. (1989). *speaking - from intention to articulation.* MIT - Press.

Marien, P., Engelborghs, S., Fabbroc, F., and Deyn, P. P. D. (2001). The lateralized linguistic cerebellum: A review and a new hypothesis. *Brain and Language*, 79(3):580–600.

Mirolli, M. and Parisi, D. (2004). Language, altruism, and docility: How cultural learning can favour language evolution. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T., and Watson, R. A., 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. MIT press.

Noppeney, U. and Price, C. J. (2002). A pet study of stimulus- and task-induced semantic processing. *Neuroimage*, 15(4):927–935.

Papathanassiou, D., Etard, O., Mellet, E., Zago, L., Mazoyer, B., and Tzourio-Mazoyer, N. (2000). A common language network for comprehension and production: a contribution to the definition of language epicenters with pet. *Neuroimage*, 11(4):347–57.

Ritvo, E. R., Freeman, B. J., and et al., A. B. S. (1986). 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.

Riva, D. and Giorgi, C. (2000). The cerebellum contributes to higher functions during development. *Brain*, 123:1051–1061.

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

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., and Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–7.

Shadmehr, R. and Mussa-Ivaldi, F. A. (1994). Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*, 14:3208–3224.

Stowe, L. A., Paansb, A. M. J., Wijersc, A. A., and Zwartsd, F. (2004). Activations of "motor" and other non-language structures during sentence comprehension. *Brain and Language*.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning - An Introduction*. MIT Press.

Tager-Flusberg, H. (1999). Language and understanding minds: connections in autism. In Baron-Cohen, S., Tager-Flusberg, H., and Cohen, D. J., editors, *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, chapter 6. Oxford University Press, second edition.

Tamada, T., Miyauchi, S., Imamizu, H., Yoshioka, T., and Kawato, M. (1999a). Activation of the cerebellum in grip force load force coordination: an fmri study. *Neuroimage*, 6:492.

Tamada, T., Miyauchi, S., Imamizu, H., Yoshioka, T., and Kawato, M. (1999b). Cerebro-cerebellar functional connectivity revealed by the laterality index in tool-use learning. *Neuroreport*, 10:325–331.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7(8):887–893.

Tomasello, M. (2000). First steps towards a usage-based theory of language acquisition. *Cognitive Linguistics*, 11:61–82.

Tomasello, M. (2003). *Constructing a Language - A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Turgut, M. (1998). Transient "cerebellar" mutism. *Child's Nervous System*, 14:161–166.

Watanabe, K., Lauwereyns, J., and Hikosaka, O. (2003). Neural correlates of rewarded and unrewarded eye movements in the primate caudate nucleus. *Journal of Neuroscience*, 23(31):10052–7.

Williams, R. S., Hauser, S. L., Purpura, D. P., and et al. (1980). Autism and mental retardation: neuropathologic studies performed in four retarded persons with autistic behaviour. *Arch Neurol*, 37:749–753.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Phil. Trans. R. Soc. Lond.*, 358:593–602.

Xiang, H., Lin, C., Ma, X., Zhang, Z., Bower, J. M., Weng, X., and Gao, J.-H. (2003). Involvement of the cerebellum in semantic discrimination: An fmri study. *Human Brain Mapping*, 18(3):208 – 214.

APPENDIX

Algorithms

This section explains the most important algorithms used in our simulations. Note that the value functions of all agents are implemented as a neural-network and that all agents

use a rule-based forward model to predict the effect of non-verbal actions. With respect to verbal actions, our simulations distinguish between two types of agents. The first type (*PP*-agents) employ a pre-programmed forward model and some additional algorithms for understanding and producing utterances. These are the agents who are already competent users of the language when a simulation starts. The second type (*NN*-agents) process language through the neural network-based architecture described above. They use this architecture to acquire linguistic competence as the simulations proceed.

Non-verbal actions are represented as action parameter sets that specify which object type τ is moved from which source location σ (either agent or tree) to which target agent Θ. The pre-programmed forward model for non-verbal actions computes the subsequent (effect) state s_{t+1} from the current state s_t and the given action parameter set.

To select non-verbal actions, agents predict the outcome of all non-verbal action parameter sets which are applicable in the current situation with their pre-programmed forward model and test which of the predicted effect states has the highest value.

To minimize computational and coding effort we did not program another forward model for the verbal actions of the PP-agents. Instead, we used the pre-programmed forward model for non-verbal actions to *simulate* the effect verbal actions would have by predicting the effect of so called *virtual take-actions*. These virtual actions would consist of an PP-agent taking a food item of a certain kind from one of the other agents. Agents are prevented from performing take-actions. So when such an action is evaluated as the one which best serves the agent's interest, the agent tries to obtain the effect of the action *indirectly* by producing an utterance which will induce the addressee to give him a food item of the right kind[6], i.e. the take-action parameter set is transformed into a verbal action parameter set (including a parameter Λ for the addressee and a parameter ω for the word). Given the simplicity of the language, the computation of the utterance of a given agent A_i which matches the prohibited take-action (characterized by a parameter set consisting of the food type τ of which an item is to be taken and agent A_j $(j \neq i)$ from whom the item is to be taken) is straightforward. The matching utterance will involve

[6]Because the learner agent has no linguistic competence when the simulations start, he will in the early stages of the game often fail to give another agent the food item he has requested. However, the linguistically competent PP-agents have been programmed in such a way that they always request a food item from another agent, including the learner, in case obtaining this item serves their purpose best.

the word ω for the type of food τ and the addressee Λ will be A_j.

In contrast to PP-agents, NN-agents do not predict the effects of virtual take-actions. As described above, they use their neural network-based forward model to predict the effect of all possible verbal actions directly.

If an agents A_j addresses another agent A_i, then A_i is required to react. If A_i is an PP-agent, he will transform the verbal action parameter set (Λ, ω, etc.) which defines the utterance into the non-verbal action which complies with the request that has been made of him. This action consists in giving A_j an item of the food type τ which matches the parameter ω.

If the addressee A_i of A_j's utterance is an NN-agent, then the (neural representation of the) verbal action parameter set and the current state of the game are given as input to A_i's neural network-based forward model; the output of this network is a possible game state which A_i takes to be the one that A_j is trying to bring about. A_i now has to compute the action which will realize this state[7]. Since the neural network-based forward model of the NN-agent is, for a considerable number of time steps, insufficiently trained to compute the correct intended state of A_j, the algorithm which computes A_i's reaction must be able to cope with representations of intended states which can differ in various degrees from the state A_j really intends. To deal with these, NN-agents are programmed to perform any action which will bring the current state closer to the presumed intended state of A_j, so long as this action does not violate the rules of the game.

Parameters of Simulation

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 (representing the trees) and the matrices (representing the agents) of s. The output of the network is the linear combination of the weighted binary inputs. As exploration mechanism we used a *soft-max* method. We 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

[7]This computation is not acquired in our simulations but implemented as another pre-programmed algorithm - a decision we made because we were not interested in agents learning to transform intentions into non-verbal actions, but in agents learning to compute the intentions behind utterances.

TD-error and the higher the γ-value, the faster was the increase in performance (with respect to score) and the better the agent performed in the game in total (see also figs. 4 and 5). Therefore, we 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 numbers were selected between a maximal value v and 0. Until the end of the simulation v was linearly decreased to 0. As initial value of v we chose 4.

The neural network implementation of the forward model is a single-layer network with two types of input: the first 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 one word sentence of which the effect is to be predicted. The output of the model is the predicted subsequent state of the game. 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 number m of past game states an agent remembers was set to 5.

The simulation parameters 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 it takes the agents to digest one of their food items ($d_o = 5$).

Further parameters 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