Authors:

Angelo Cangelosi, Thomas Riga

Title:

An embodied model for sensorimotor grounding and grounding transfer: experiments with epigenetic robots

in press, *Cognitive Science*

Affiliation

Angelo Cangelosi

Adaptive Behaviour & Cognition Research Group

School of Computing, Communications and Electronics

University of Plymouth (UK)

email: acangelosi@plymouth.ac.uk

http://www.tech.plym.ac.uk/soc/research/abc

Thomas Riga

Aitek SpA: Information and Communication Technologies

Via della Crocetta 15

16122 Genova

email: thomas@aitek.it

Running head:

Embodied robotic model of symbol grounding

Keywords

Symbol grounding, epigenetic robotics, human-robot interaction, embodied cognition, language evolution, imitation, grounding transfer

Abstract

The grounding of symbols in computational models of linguistic abilities is one of the fundamental properties of psychologically-plausible cognitive models. This paper presents an embodied model for the grounding of language in action based on epigenetic robots. Epigenetic robotics is one of the new cognitive modeling approaches to modeling autonomous mental development. The robot model is based on an integrative vision of language, in which linguistic abilities are strictly dependent on, and grounded in, other behaviors and skills. It uses simulated robots that learn through imitation the names of basic actions. Robots also learn higher-order action concepts through the process of grounding transfer. The simulation demonstrates how new, higher-order behavioral abilities can be autonomously built upon previously-grounded basic action categories, following linguistic interaction with human users.

1. Introduction

Various computational modeling approaches have been proposed to study communication and language in cognitive systems, such as robots and simulated agents. On one end there are models of language primarily focused at the internal characteristics of individual agents, where the lexicon is constructed upon a self-referential symbolic system. The cognitive agents only possess a series of abstract symbols used for both communication and for representing meanings (e.g. Kirby, 2001). These models are subject to the symbol grounding problem (Harnad, 1990). That is, symbols are self-referential entities that require the interpretation of an external experimenter to identify the referential meaning of the lexical items.

On the other end, there are grounded approaches to modeling language, where linguistic abilities are developed through the direct interaction between the cognitive agents and the social and physical world they interact with. The external world, and the agent's own internal representation of it, play an essential role in shaping the language used by these cognitive systems. Language is therefore grounded in the cognitive and sensorimotor knowledge of the agents (Cangelosi, Bugmann & Borisyuk, 2005; Steels, 2003). For example, environmental stimuli are perceptually transformed by the agent's own sensorimotor systems and might constitute the topic of conversation. This is the case of categorical perception, where the agent's perceptual abilities constrain the representation of the environment that an agent can build. At the same time, the environment is subject to changes due to the communicating act of the agents themselves, e.g. when the agents' lexicon creates new categorical representation of environmental entities.

The grounding of language in autonomous cognitive systems requires two mechanisms. The first is the direct grounding of the agent's basic lexicon. This assumes the ability to link perceptual (and internal) representations to symbols through supervised feedback. For example, an agent can learn that the symbol "horse" is grounded in its direct experience with this animal. The second mechanism implies the ability to transfer the grounding from the basic symbols to new symbols obtained by logical (e.g. syntactic) combinations of the elementary lexicon. The same agent can learn, without direct experience, that there is a hypothetical animal, the "unicorn", that is perceptually grounded in the linguistic description of "horse with a horn".

Direct grounding has been widely studied in embodied autonomous agents (see Cangelosi 2005 for a review), whilst grounding transfer has only been demonstrated in connectionist simulations (Cangelosi, Greco & Harnad, 2000; Riga, Cangelosi & Greco, 2004). This paper reports a new study on grounding transfer in cognitive robots for the acquisition of higher-order action categories via linguistic instructions. This uses an epigenetic robotic approach (Weng et al., 2001; Prince & Demiris, 2003; McClelland, Plunkett & Weng, in press) where a simulated robot initially learns, via imitation, a series of basic actions and their corresponding names. An artificial neural network controls the robot's motor and linguistic behavior. The robot then acquires the names of new high-order action categories following linguistic interaction with human users. The hypothesis is that the combination of direct grounding of basic words and their use to express new categories will results in the actual acquisition of new sensorimotor capabilities. After training, the agent will be tested to establish whether it can actually produce the new composite actions when their hearing their names. This would demonstrate that grounding from the basic action names has been transferred to the new composite categories.

The motivation for developing such a model of language embodiment and grounding is two-fold. First, there is a need for psychologically-plausible computational models of language embodiment to further support the growing theoretical and experimental evidence on sensorimotor grounding (Pecher & Zwaan, 2005). Computational models can help investigating the detailed mechanisms involved in the process of grounding. The proposed approach is based on the combination of neural network and robotic methodologies, which we call "Embodied Connectionism" (Cangelosi, in prep.). This will provide a modeling platform for the development of grounded language systems that overcome the known shortcoming of featured-based connectionist models of language (Glenberg, 2005 – see also discussion in section 4) and of the symbolic-only models (Burgess & Lund, 1997; Landauer & Dumais, 1997; Kirby, 2001).

The second motivation regards the scientific and technological advances in the design of interactive cognitive systems able to communicate with humans and other robots. In artificial intelligence and robotics, the issue of instruction-based learning and linguistic interaction has become of the priority areas for future research. Some of the most promising results have come from grounded robotic approaches based on the acquisition of language through direct sensorimotor interaction with the environment (Cangelosi et al., 2005).

In the following sub-sections, we will look at the state of the art in both experimental and modeling studies of the grounding of language. We will then present the epigenetic robot modeling setup (section 2) and the results of the simulated robotic experiments on symbol grounding and the autonomous transfer of sensorimotor grounding (section 3). The final section will discuss the advantages of such an approach to model embodied cognition and its potential application in further experimental and computational investigations of language grounding.

1.1 Grounding language in action and perception: theoretical and experimental studies

In the past few years there has been a growing body of theoretical and empirical evidence in support of the role of embodiment and sensorimotor factors in language use (e.g. Barsalou, 1999; Coventry & Garrod, 2004; Feldman & Narayanan, 2004; Gallese & Lakoff, 2005; Glenberg & Robertson, 2000; Pulvermuller, 1999; Zwaan, 2004 – see also Pecher & Zwaan, 2004 for a recent review). Overall, language grounding theories support the view that language use involves modality-specific simulations of the referents and the actions described in the sentences.

Simulation theories vary for the focus they put on some of the mechanisms involved in these simulations and the detailed function of the simulation process. For example, Barsalou (1999) focuses on modality-specific perceptual and simulation processes within the Perceptual Symbol System hypothesis. He suggests that the brain association area partially reactivate sensorimotor areas to implement perceptual symbols. This includes memories of sensorimotor, proprioceptive and introspective events, and also dynamic mental representations of object interaction (e.g. Zwaan, Madden, Yaxley & Aveyard, 2004). Such memories are organized around a common frame, which constitute the structure of a simulator. The coordinate activity of simulators implement a basic conceptual system that represents types, supports categorization, and produces categorical inferences. Barsalou also shows how abstract concepts are grounded in complex simulations of combined physical and introspective events.

Glenberg and collaborators (e.g. Glenberg & Kaschak, 2002; Borghi, Glenberg & Kaschak, 2004; Kaschak et al., 2005) focus on the action and embodiment component of language. They demonstrate the existence of Action-sentence Compatibility Effects. In sentence comprehension tasks participants are faster to judge the sensibility of sentences

implying motion toward the body (e.g. “Courtney gave you the notebook”) when the response requires moving toward the body. When the sentence implied movement away from the body, participants were faster to respond by literally moving away from their bodies. The data support an embodied theory of meaning that relates the meaning of sentences to human action and motor affordances. This view, called Indexical Hypothesis (Glenberg, 1997; Glenberg & Robertson, 2000), suggest that in the reading of a sentence, the first process is to index words and phrases to objects in the environment or to analogical perceptual symbols. The second process is deriving affordances from the object or perceptual symbol. Finally, the third process is to mesh the affordances into a coherent set of actions. The mesh process is guided by syntax of the sentence being processed.

Gallese and Lakoff (2005) use neurophysiology evidence to show that language makes direct use of the same brain structures used in perception and action. They suggest that brain structures in the sensorimotor regions are exploited to characterize abstract symbolic concepts that constitute the meanings of grammatical constructions and general inference patterns. The semantics of grammar is constituted by cogs, i.e. structuring circuits used in the sensory-motor system.

Such theories of the sensorimotor grounding of language propose an alternative account to classical symbolic theories of meaning and cognition (e.g. Fodor, 1975). According to this account, the meaning of words comes from the contexts in which these words are used (Burgess & Lund, 1997; Landauer & Dumais, 1997) and there is no need for direct correspondence between the symbolic system and perceptual states. Proponents of symbolic theories acknowledge the role of perceptual and sensorimotor factors in the acquisition of language, but after assume the autonomy of language and symbolic processes in cognitive tasks. Embodiment theories of language, instead, stress the on-line effects of sensorimotor processes in language use.

1.2 Grounding language in action and perception: Computational models

Grounded approaches to modeling language are based on the principles of autonomous and embodied communication. Cognitive agents can autonomously acquire communication capabilities through interaction with each other and with humans. An important characteristic of this approach is the fact that the properties of the robots' own body and their physical environment influence, and contribute to, the acquisition of a lexicon directly grounded in world they live in (Steels, 2003; Cangelosi et al., 2005).

Some of these models focus on the emergence of shared lexicons through biological and/or cultural evolution mechanisms (Cangelosi & Parisi, 2002). In these models, a population of cognitive agents is initialized that use random languages. Agents are able to interact with the physical entities in the environment and to construct a sensorimotor representation of it. Through an iterative process of communication and language games, agents converge toward a shared lexicon. For example, Luc Steels and collaborators (Steels, 2003; Steels, Kaplan, McIntyre & Van Looveren, 2002) use hybrid population of robots, internet agents and humans engaged in language games. Agents are in turn embodied into two "talking head" robots to play language games. A shared lexicon gradually emerges to describe a world made of colored shapes. This model has been also extended to study the emergence of communication between humans and robots, suing the SONY AIBO interactive toy robot (Steels & Kaplan, 2000). Steels's approach is characterized for the focus on the naming of perceptual categories and for the stress of social mechanisms in the grounding and emergence of language. Marocco, Cangelosi and Nolfi (2003) use evolutionary robotics for the self-organization of simple lexicons in a group of simulated robots. Agents are first evolved for their ability to manipulate objects (e.g. touché spheres, avoid cubes).

Subsequently, they are allowed to communicate with each other. Populations of agents are able to evolve a shared lexicon to name the objects and the actions being performed on them.

In other models of language grounding, robotic agents acquire a lexicon through interaction with human users. For example, Roy, Hsiao and Mavridis (2003) have developed an architecture that provides perceptual, procedural and affordance representations for grounding the meaning of words in conversational robots. Sugita & Tani (2004) use a mobile robot that follows human instructions based on the combinations of five basic commands. Yu (2005) focuses on the combination of word learning and category acquisition to show improvements in both word-to-world mapping and perceptual categorization. This suggests a unified view of lexical and category learning in an integrative framework.

The above models clearly support the view that language is intrinsically linked to the constraints imposed by the human perceptual, cognitive and embodiment system. However, they have seldom been used to address specific issues and findings in language embodiment research. For example, Coventry, Cangelosi and collaborators (Joyce, Richards, Cangelosi and Coventry, 2003; Coventry et al. 2004); have developed a neural network model of spatial language that directly simulates the perceptual symbol system hypothesis of Barsalou. They use simple recurrent network (Elman, 1990), within a hybrid connectionist/vision architecture, to simulate and integrate perceptual factors in the production of spatial quantifiers. Dominey (2005) carried out some human-robot communication experiments on the emergence of grammar. This study provides insight into a developmental and evolutionary passage from unitary idiom-like holophrases to progressively more abstract grammatical constructions. Finally, in a computational model based on population of agents, Cangelosi and Parisi (2004) use synthetic brain imaging methods to analyze the activity of the agents' neural networks. Results show that different linguistic categories, such as nouns and verbs, share the neural substrate of different sensorimotor processes. Results show that

nouns (names of objects) produce more neural activity in the hidden layer dedicated to sensory processing of visual stimuli, while verbs (names of actions) produce enhanced synaptic activity in the layer where sensory information is integrated with proprioceptive input. Such findings are qualitatively compared with human brain imaging data that indicate that nouns activate more the posterior areas of the brain related to sensory and associative processing while verbs activate more the anterior motor areas (Cappa & Perani, 2003).

2. An epigenetic robotic model for grounding transfer

The model is based on an on-line imitation learning algorithm for the acquisition of behavioral and linguistic knowledge in a group of robots. The combination of imitation and language learning is mainly motivated by the fact that imitation has been consistently considered as one of the fundamental mechanisms for the acquisition of language (Tomasello, 2002). This model will be based on a simple on-line supervised neural network algorithm. It uses error backpropagation to continuously correct the motor response of an imitator robot so that it closely matches the behavior of a demonstrator robot. The backpropagation algorithm is also used to teach the robot the names of actions.

2.1 Robot body

The model consists of a computer simulation of two robotic agents embedded in a virtual environment. The simulation program accurately models the physical constraints and object-

object interactions using the physics engine Open Dynamics Engine[1] (ODE). ODE is an open source library for simulating rigid body dynamics, advanced joint types and integrated collision detection with friction. It can be used for simulating vehicles, objects in virtual reality environments and virtual creatures. Although the ODE robotic model cannot fully take into account all the complex embodiment properties of real robots, it permits a good inclusion and consideration of physical systems.

The robot's body consists of two 3-segment arms (rotating shoulder, upperarm, forearm) attached to a torso and a base with 4 wheels (Fig. 1). The details of the robots body are as follows (in ODE length points):

- wheels (4): width 0.2, ray 0.25
- base: width 0.75, length 0.75, height 0.25
- torso: width 0.25, length 0.25, height 0.75
- neck: width 0.25, length 0.25, height 0.25
- head: width 0.35, length 0.5, height 0.25
- shoulder (2): width 0.25, length 0.25, height 0.25
- upperarm (2): width 0.25, length 0.25, height 0.75
- forearm (2): width 0.25, length 0.25, height 0.75

The robot has 12 degrees of freedom. The constraints of the degrees of freedom of the joints are as follows:

- wheels-base (4): no limit
- torso-shoulder (2): 180 degrees (vertical plane)
- shoulder-upperarm (2): 90 degrees (horizontal plane)
- upperarm-forearm (2): 90 degrees
- torso-neck: 90 degrees (vertical plane)

[1] http://opende.sourceforge.net

- neck-head: 180 degrees (horizontal plane)

The first agent, called “demonstrator”, has the role of showing the correct performance of some basic motor actions. This robot is manually programmed to perform actions on objects, i.e. with pre-specified forces to apply to the motor joints at every time-step. The second agent, an “imitator”, learns the actions by imitating the demonstrator’s behavior. This agent is equipped with an artificial neural network controller. The imitator learns to perform basic actions by predicting the demonstrator’s movement trajectories using an "imitation algorithm" which supplies teacher input to a neural network. The resulting motion dynamics are elaborated by the neural network that is able first to repeat the actions during imitation, and successively is able to execute them autonomously in absence of the imitator input and feedback. The robot’s neural controller also learns the words associated to the actions, so that when the imitator “hears” a word, it can perform the corresponding action.

2.2 Neural network controller

The neural network controller of the imitator consists of a fully connected feed-forward network with bipolar sigmoid units. Twenty-six input units encode the names of all possible actions. The hidden layer contains eight units. These are modularly connected to the eight output motor units (see Fig. 1). The output value of each motor neuron corresponds to the force applied to the corresponding motorized joint. The modularity of the hidden layer is realized by separately connecting four groups of two hidden nodes to four pairs of output nodes. These pairs encode the following motorized joint groups: left upperarm-forearm; right upperarm-forearm; shoulder-upperarm; wheels (same for all 4). The modular organization of the hidden-output connections has been designed to allow the robotic agents to learn combinations of the action words. Such a modular, connectionist architecture has been

demonstrated to be necessary for action/language tasks requiring the acquisition of higher-order categories via combinations of their names (Greco, Riga & Cangelosi, 2003).

---------- Insert Figure 1 about here ----------

The diagram in Fig. 1 gives an overview of the imitator agent's functional modules, its neural controller, and a view of the 3D robots and environment. When the demonstrator agent performs an action and utters the corresponding word, the imitator agent activates the following procedure: the symbolic parser filters the linguistic input and converts it to a format suitable for the network (localist encoding of one word per linguistic input unit). In parallel, the imitation algorithm computes an estimation of the motor output necessary to perform the same action. The neural network then computes the actual motor output at the current time-step. This output is sent to the actuators to produce the action. Successively, an on-line error backpropagation is applied to the imitator's neural controller, using the motor output estimated by the imitation algorithm as teaching input. All weights and biases are subject to change. The backpropagation algorithm is applied at each time-step.

The imitation algorithm, based on a hyperbolic tangent function, is defined by the following functions:

$$f(t+1) = f(t) + g(x(t), y(t))$$

$$g(x(t), y(t)) = \alpha\left(\frac{2}{1+\exp(-2\beta(x(t)-y(t))}\right)-1 \qquad (\alpha = scale\ \beta = gain)$$

The first function computes an estimation of the necessary force *f(t + 1)* to apply to each motorized joint in the next time-step, so that it approximates the posture currently exhibited by the demonstrator. It takes as input the joint angles *x(t)* of the demonstrator agent and the joint angles *y(t)* and motor forces *f(t)* of the imitator agent for all joints in the current time-step. Experimental evidence has demonstrated that joint angles are used for postural control in imitation (Desmurget & Prablanc, 1997). The scale α and gain β are constant values, set to 0.5 in the present simulation. The scale parameter α is similar to the learning rate in the error backpropagation algorithm, where higher values produce bigger weight changes and faster learning. The gain parameter β changes the hyperbolic function (lower values correspond to flatter sigmoids).

For simplicity of implementation, the input regarding the posture of the teacher is assumed to have been preprocessed in order to identify and compute the demonstrator's posture angles. Thus the imitator agent directly receives the joint angle values, instead of having to analyze the scene of a moving arm and generate the values of the joint angles. The implementation based on direct imitation is justified by the need to have a process of grounding based on the pre-acquisition of action categories from a teacher or parent agent. The choice of direct overt imitation for action and language learning is also motivated by the central role of imitation in the evolution and acquisition of language and cognition (e.g. Tomasello, 2002; Charman, Baron-Cohen, Swettenham, Baird, Cox & Drew, 2000). The overt imitation setup has also been chosen for the potential is has of allowing the imitator agent to learn to imitate actions directly performed by a human participant, e.g. through motion capture software. However, at this stage we did not want to deal with the complexity of robotics and motion capture systems (Schaal, 1999; Dautenhahn & Nehaniv, 2002), because of the focus on language learning and grounding transfer.

2.3 Robot training

The simulation consists of three training stages and a testing phase. Training is incremental and follows these three stages: (i) Basic Grounding (BG), (ii) Higher-order Grounding 1 (HG1) and (iii) Higher-order Grounding 2 (HG2). The testing stage, at the end of the training, consists of the autonomous execution of all basic and higher-order actions, following the input of the corresponding action names.

2.2.1 Basic Grounding BG

During the BG training stage, the imitator learns to execute eight basic actions by observing the demonstrator and mimicking its movement. Words corresponding to the action names are presented in input to the learner's neural controller. The imitator simultaneously learns the actions and their names, thus directly grounding the word in the perception of the imitator's action and the production of its own motor response. This constitutes the basic grounding of action words. The following eight basic actions/words are taught during each BG training epoch: CLOSE_LEFT_ARM; CLOSE_RIGHT_ARM; OPEN_LEFT_ARM; OPEN_RIGHT_ARM; LIFT_LEFT_ARM; LIFT_RIGHT_ARM; MOVE_FORWARD; MOVE_BACKWARD. To perform each of these basic actions, the robot always starts from a default position with an angle of 45^{o} for both the joints upperarm-lowearm and shoulder-upperarm. Each action lasts for 100 time-steps. The error backpropagation is applied on-line at every time-step. The basic grounding learning lasts for 50 training epochs.

2.2.1 Grounding transfer during higher-order learning HG1 and HG2

During the next two higher-order grounding stages (HG1 and HG2), the imitator robots learn the names of combined actions by receiving linguistic descriptions through a natural language interface, or directly from the teacher agent. The higher-order learning has the role of acquiring the names (and concepts) of new actions. This is possible through the process of symbol grounding transfer by which the sensorimotor grounding of basic action names is indirectly transferred to that of new words.

A human operator can communicate with the agent using a keyboard to write simple instructions using an ad-hoc pidgin English language. Two types of utterances are possible: Higher order descriptions and commands. Higher-order descriptions consist of three words respectively naming a new higher-order action word and two basic/lower-order actions. These instructions serve to learn the new word and its associated action pattern. Commands consist of the name for an action. They cause the agent to execute the appropriate action by activating the corresponding input node in the network and producing the motor action. Higher-order descriptions are used during learning, while commands are used for testing. For speed of execution, the higher-order linguistic descriptions and commands are recorded before the simulation so that the teacher agent can send them in input to the learner during the higher-order training stages.

A higher-order action based on the combination of two basic actions is called 1st level higher-order behavior (HG1). For example, one of such behaviors is object grabbing and has the following description[2]: "GRAB [is] CLOSE_LEFT_ARM [and] CLOSE_RIGHT_ARM" (see top row of Fig. 2). Grounding transfer takes place from the directly grounded "CLOSE_LEFT_ARM" and "CLOSE_RIGHT_ARM" words to the new "GRAB" word. This enables the agent to correctly execute the command "GRAB" by combining the actions of pushing both arms towards the object and grabbing it.

[2] The words between bracket are filtered out by the parsing and ignored during the training.

A higher-order behavior consisting of the combination of one basic action and one 1st order action is called 2nd order action (HG2). For example, the description "CARRY [is] GRAB [and] MOVE_FORWARD" is a 2nd order action (see bottom row of Fig. 2).

After the last BG epoch, the imitator robot receives HG1 linguistic descriptions (i.e. a new word and two known words referring to basic actions). Each HG1 training epoch contains 13 learning trials, i.e. five 1st order actions (GRAB, PUSH_LEFT, PUSH_RIGHT, OPEN_ARMS, ARMS_UP) and eight BG actions. HG1 training stage lasts for 100 epochs.

The imitator agent starts HG2 training at the 151st epoch. Three 2nd level higher order actions (CARRY, PULL, CHEER) are taught during HG2 stage for 150 additional epochs. The current implementation with 50 epochs for the BG stage, 100 epochs for HG1 and 150 epochs for HG2 training reflects the increasing difficulty of the incremental learning task the imitator agent needs to master.

---------- Insert Figure 2 about here ----------

To achieve grounding transfer, the imitator agent learns to use some of the neural representations acquired during BG to those of stages HG1 and HG2. This process grounds new words in the neural controller by adaptively linking the hidden units' activations of the words contained in the description, as previously demonstrated in Cangelosi *et al.* (2000). In the present model, this is achieved first by separately providing each defining (i.e. basic/lower-order) word in input to the network and temporarily recording the motor response (without applying error backpropagation). Successively, the network receives as input only

the newly defined (i.e. higher-order) word so that the resulting output is corrected through backpropagation by using as teaching input the output previously recorded.

---------- Insert Figure 3 about here ----------

The backpropagation weight-correction procedure consists of two training cycles, respectively for each of the two basic words used in the description (Fig. 3). For example, to learn the novel behavior of grabbing from the description "GRAB [is] CLOSE_LEFT_ARM [and] CLOSE_RIGHT_ARM", the agent's controller first produces the output corresponding to the input of the first word "CLOSE_LEFT_ARM". This force is not applied to the joint motors, but is temporarily stored to be used as teaching input in the next activation cycle. The joints values are generated and recorded for all the 100 time-steps of action execution. Subsequently, the input node corresponding to the "GRAB" action is activated and the network produces a motor response in the output nodes. The previous teaching input is now used to compute the error and apply the backpropagation algorithm for 100 time-steps. During the second phase, the same procedure is repeated for the generation of the teaching input signal from the activation of the input node "CLOSE_RIGHT_ARM" and the subsequent weight correction from the input of the word "Grab". These two steps are repeated for each combined action description in training stages HG1 and HG2.

This grounding transfer mechanism enables an agent to learn new actions not only through direct experience and trial-and-error learning, as during BG, but also indirectly through the exchange of linguistic utterances with other agents. New actions are learned without the need of direct observation and imitation of the demonstrator agent.

3. Simulation results

Each simulation experiment consisted of 300 training epochs (50 BG, 100 HG1, 150 HG2). Each action lasts for 100 time-steps, so each simulation lasts for 30,000 cycles. Ten replications of such an experiment were performed, using neural networks with different initial random weights. Weights were initialized in the range ±1.0 at the first epoch. The learning rate was 0.05 during BG learning, and 0.01 during the grounding transfer process of HG1 and HG2.

The final posture errors and the average posture errors were registered for the BG, HG1, HG2 and testing stages of every epoch. The final posture error measures the difference in posture between the two agents only at the last (100^{th}) time-step of each action. This error does not consider the movement trajectories, but only the final posture. The average posture error records the difference in posture (i.e. 8 joint angles) between the imitator and the demonstrator averaged over all 100 time-steps, thus taking into account the movement trajectories. All error values are computed as Root Mean Square values (RMS), using as correct value the joint angles of the demonstrator. Note that although the imitator's joints are compared with those of the demonstrator for visualizing the RMS errors during training, during the HG stages these errors are never used by the imitator during backpropagation learning.

The imitation learning of the 8 basic actions was successful. All actions are correctly acquired, with a final average posture error of 0.08 after the last epoch (average error over the 10 replications). All five 1^{st} order actions/names were also successfully learned with a final posture error of 0.05 after the last epoch. The three 2^{nd} order actions/names were successfully

acquired with a final posture RMS error of 0.09 after the last epoch. Thus agents correctly execute all basic, 1st and 2nd order actions in response to the input of their names. For example, after hearing the 2nd order action name "PULL", agents pushed both arms against the object and moved backward, effectively exhibiting the behavior of dragging the object backward as defined in the current experiment.

Overall the average posture error remains higher than the final posture error. This means that the imitator agent gradually approximates the movement trajectory towards the target posture, but finishes in the desired position with great accuracy. This effect is present in the basic behaviors, but becomes more evident when executing 1st level composite actions and is very clear in the 2nd level behaviors (Fig. 4). The level of grounding transfer of a word has a clear effect on the behavior it generates, as the ideal trajectory to a target position is not followed accurately, though always leading to the correct final posture.

This pattern of results, i.e. the learning of all basic and HG actions up to an final posture RMS error of between 0.05 and 0.09, is found in all replications and there are no major qualitative or quantitative differences between the 10 simulations.

---------- Insert Figure 4 about here ----------

4. Discussion and conclusions

The simulation presented here provides a clear demonstration of the grounding transfer mechanism for simulated linguistic robots. New actions are acquired through the process of symbol grounding transfer from basic, directly-grounded action categories to higher-order, indirectly-grounded behaviors. The grounding transfer is a very important aspect of research on autonomous cognitive systems. For a system to be fully autonomous it is important that it is able to use its own linguistic and cognitive abilities to further expand its knowledge of the environment. The design of a linguistic agent able to acquire and ground language only through direct perception and experience of the external world is not enough (Harnad, 1990). One of the most important aspects of human language is productivity, by which new concepts can be expressed through combinations of the words. Although the robotic agents studied in this simulation do not have full linguistic and compositional abilities (e.g. the use of a syntactic lexicon), they can rely on simple compositional mechanisms to enrich their lexicon. The grounding transfer makes sure that new concepts are grounded into the agents' own sensorimotor repertoire. In addition, the agents do not necessarily need to rely on the external input of the demonstrator robot (or a human experimenter) to acquire new concepts, since they can autonomously combined the basic words to construct new composite action categories.

The procedure used for the autonomous acquisition (production) of high-order action categories (see Fig. 3) can be considered an implementation of Barsalou's (1999: section 3.1) symbol productivity mechanism in the perceptual symbol system framework. The agent plays some kind of internal "mental" simulation when they produce and record the output values corresponding to the input activation of the two basic action names (e.g. CLOSE_LEFT_ARM; and CLOSE_RIGHT_ARM). These mental records are then used by the agent to merge the results of the two motor simulations and auto-teach the output values corresponding to the name of the new action (GRAB). In addition, the type of higher-order composite actions

described here also related to research on conceptual combination, such as in the categories based on noun-noun combinations (Wisnieski, 1997).

The design and test of this first robotic model of symbol grounding transfer required some simplifications, both in the repertoire of behavior/lexicon and the imitation algorithm. However, ongoing research is focusing on the scaling up of this model. For example, in Hourkadis and Cangelosi (2005; Cangelosi, Hourdakis & Tikhanoff, 2006), we have expanded the neural network controller of the robot to include both language production and comprehension capabilities. The neural network receives in input both visual information and language so that the agent can produce linguistic descriptions (vision input to language output) as well as being able to understand language (from language input to motor output). In Massera, Nolfi and Cangelosi (2005), new simulations have focuses on the autonomous acquisition of arm control capabilities without the need for direct imitation. This advances model of the robotic arm model uses evolutionary algorithms. Other simulations are explicitly addressing the scaling up of the lexicon to hundreds of words and the use of more structured lexicons. This is based on the gradual introducing of syntactic structures. For example, the first step will consist in the ability to use arguments for the learned actions. For example, through the introduction of three types of objects (e.g. round spheres for balls, flat objects for books, long cylinders for sticks) it is possible to train robots to apply the same action to different objects, such as “Grab(Ball)”, “Grab(Book)”. At the same time, the use of objects with different shapes will permit the construction of a variety of linguistic categories whose representation might vary depending on the interaction between the robot’s own embodiment properties and the object motor affordances.

The potential extensions discussed above will permit the use of this model as an embodied simulation platform for new computational investigations that replicate the well known grounding effects. For example, a model able to learn actions in response to objects

requiring different motor affordances could be used to replicate the action-sentence compatibility effects (ACE) found by Glenberg and collaborators (e.g. Glenberg & Kaschak, 2002). One could train a robot to perform various sets of action all following specific spatial directions (e.g. pull/push, open/close) and to learn linguistic descriptions of scenes involving the manipulation of objects with front/backward movements. The analyses of the activity of the neural networks during the successful replication of ACE effects could permit a detailed investigation of the interaction and sharing between sensorimotor and linguistic representations. Embodied simulation agents have been already used to study embodiment effects, although not linked to language. Tsiotas, Borghi and Parisi (2005) have built an evolutionary agent model of the action compatibility effects. Tucker and Ellis (2001) have demonstrated action compatibility effect between the type of grasp (precision vs. power grip in the response to micro-affordances for a pen vs. apple) and a task-irrelevant dimension (e.g. color). Tsiotas *et al.* first trained agents (consisting of an arm with by two fingers) to grasp objects according to their size – e.g. precision grip for small objects and power grip for large objects (compatible condition). These corresponded to the default object affordances. Then they also trained agents to grasp objects according to their color, ignoring their size (incompatible condition). Agents produced the same compatibility effects in terms of shorter training cycles for the compatibility condition versus the incompatible. In addition, analyses of the agents' neural networks showed that in the hidden units the visual input of an object automatically activates information on how to grasp them, also when this information is not relevant to the task. This study demonstrates the potential of computational agent-based models for studying embodiment effects.

This study also has important potential implications for robotics research, in particular in cognitive robotics. In this area, epigenetic robotics is one of the most promising approaches for the design of autonomous robots (Weng *et al.*, 2001; McClelland *et al.*, in press; Smith &

Gasser, 2005). This approach takes inspiration from research in developmental psychology and neuroscience and focuses on the emergence of complex cognitive and perceptual structures as a result of the interaction of an embodied system with a physical and social environment. The present simulation mostly focuses on the grounding of linguistic abilities and the acquisition of early words. As a consequence, the other cognitive capabilities of the robotic agent are based on simplified assumptions. For example, the model is based on the technical assumption that the imitator can "read" the demonstrator's joints angle and use them as teaching input. A variety of models of imitation have been proposed, some of which are based on more psychologically-plausible mechanisms. Demiris and Johnson (2003) have recently focused on the fact that the robot must infer and predict the actions being demonstrated. The future integration of various imitation, cognitive and linguistic abilities in one integrated cognitive system can better help the epigenetic design of autonomous robotic systems.

Finally, this research also has a general practical and technological bearing. In robotics and artificial intelligence, language grounding models can provide novel algorithms and methodologies for the development of effective interaction between humans and autonomous computer and robotic systems. If robots are to be introduced into everyday life, they will need to be "programmable" by users that don't necessarily have formal computer programming skills. Humans acquire language through a rich combination of learning strategies, including imitation, attentional cues, feedback cues, gestures and verbal instructions. These modalities could be combined in a linguistic robotic model to achieve a natural, intuitive, way of programming robots.

References

Barsalou, L. (1999), Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577-609.

Borghi, A.M., Glenberg, A.M., & Kaschak, M.P. (2004). Putting words in perspective. *Memory & Cognition*, *32* (6), 863-873.

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*, 177-210.

Cangelosi, A. (in preparation). Embodied Connectionism: From feature-based neural network simulations to embodied neural network agents.

Cangelosi, A. (2005). Approaches to grounding symbols in perceptual and sensorimotor categories. In: H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 719-737), Elsevier.

Cangelosi A., Bugmann G., & Borisyuk R. (Eds.) (2005). *Modeling language, cognition and action: Proceedings of the 9th neural computation and psychology workshop*. Singapore: World Scientific.

Cangelosi A., Greco A., & Harnad, S. (2000). From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Connection Science*, *12* (2), 143-162.

Cangelosi A., & Parisi, D. (Eds.) (2002). *Simulating the evolution of language*. London: Springer.

Cangelosi, A., & Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, *89* (2), 401-408.

Cangelosi, A., Hourdakis, E., & Tikhanoff, V. (2006). Language acquisition and symbol grounding transfer with neural networks and cognitive robots. In: *Proceedings of IEEE Conference on Computational Intelligence*, Vancouver, July 2006.

Cappa, S.F., & Perani, D. (2003). The neural correlates of noun and verb processing. *Journal of Neurolinguistics*, *16*, 183-189

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Cox, A., & Drew, A. (2000). Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development*, *15* (4), 481-498.

Coventry, K.R., & Garrod, S.C. (2004). *Saying, seeing and acting: The Psychological Semantics of Spatial Prepositions*. Psychology Press: Hove.

Coventry, K.R., Cangelosi, A., Rajapakse, R., Bacon, A., Newstead, S., Joyce, D., & Richards, L.V. (2005). Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In: C. Freksa, B. Knauff, B. Krieg-Bruckner & B. Nebel (Eds.), *Spatial Cognition, Volume IV. Reasoning, Action and Interaction (Lecture notes in Computer Science)* (pp. 98-110). Springer-Verlag.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211

Fodor, J.A. (1975). *The Language of Thought*, Cambridge, MA: Harvard University Press.

Hourkadis E., & Cangelosi A. (2005). Grounding transfer in autonomous robots. *22nd Annual Workshop of the European Society for the Study of Cognitive Systems*, London.

Dautenhahn, K., & Nehaniv C. (Eds.) (2002). *Imitation in animals and artifacts*. MIT Press

Demiris Y., & Johnson M. (2003). Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning. *Connection Science*, *15* (4), 231-243.

Desmurget M., & Prablanc, C. (1997). Postural control of three-dimensional prehension movements. *Journal of Neurophysiology*, *77*, 452-464.

Dominey, P. (2005). Emergence of grammatical constructions: Evidence from simulation and grounded agent experiments. *Connection Science*, *17* (3-4), 289-306 (Special issue on The Emergence of Language)

Feldman, J., & Narayanan S. (2004). Embodied meaning in a neural theory of language. *Brain and Language*, *89*, 385-392.

Gallese, V., & Lakoff G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology*, *22*, 455-479.

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

Glenberg, A. M. (2005). Lessons from the embodiment of language: Why simulating human language comprehension is hard. In Cangelosi et al. 2005, pp. 17-30.

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

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language*, *43*, 379-401.

Greco, A., Riga, T., & Cangelosi, A. (2003). The acquisition of new categories through grounded symbols: An extended connectionist model. In: O. Kaynak, E. Alpaydin, E. Oja & L. Xu (Eds.). *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003. (Lecture Notes in Computer Science 2714)*, Berlin: Springer, pp. 773-770

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

Joyce D., Richards L., Cangelosi, A., & Coventry K.R. (2003). On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In: F. Detje, D. Dörner, H. Schaub (Eds.), *The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 147-152), Universitätsverlag Bamberg

Kaschak, M.P., Madden, C.J., Therriault, D.J., Yaxley, R.H., Aveyard, M.E., Blanchard, A.A., & Zwaan, R.A. (2005). Perception of motion affects language processing. *Cognition*, *94*, B79-B89.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation and Cognitive Science*, *5* (2), 102-110.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Marocco, D., Cangelosi, A., & Nolfi, S. (2003). The emergence of communication in evolutionary robots. *Philosophical Transactions of the Royal Society of London – A, 361*, 2397-2421.

Massera, G., Nolfi S., & Cangelosi, A. (2005). Evolving a simulated robotic arm able to grasp objects. In A. Cangelosi et al. 2005, pp. 203-208.

McClelland, J., Plunkett, K., & Weng, J. (in press). Autonomous Mental Development (special issue). *IEEE Transactions on Evolutionary Computation*

Pecher, D., & Zwaan, R.A., (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*. Cambridge, UK: Cambridge University Press.

Prince, G.P., & Demiris, Y. (2003). Editorial: Introduction to the Special Issue on Epigenetic Robotics. *Adaptive Behavior*, *11* (2), 75-77

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, *22*, 253-270.

Riga, T., Cangelosi A., & Greco A. (2004). Symbol grounding transfer with hybrid self-organizing/supervised neural networks. In: *IEEE International Joint Conference on Neural Networks - Proceedings, v. 4*, pp. 2865-2869.

Roy, D., Hsiao, K., & Mavridis, N. (2003). Conversational robots: Building blocks for grounding word meanings. In: *Proceedings of the HLT-NAACL03 workshop on learning word meaning from non-linguistic data*.

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, *3* (6), 233–242

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life*, *11* (1-2), 13-30. Special issue on Embodied and Situated Cognition

Steels, L. (2003) Evolving grounded communication for robots. *Trends in Cognitive Sciences*, *7* (7), 308-312.

Steels, L., & Kaplan, F. (2000). AIBO's first words: The social learning of language and meaning. *Evolution of Communication, 4* (1), 3-32.

Steels, L., Kaplan, F., McIntyre, A., & Van Looveren J. (2002). Crucial factors in the origins of word-meaning. In: A. Wray (Ed.), *The transition to language* (pp. 252-271), Oxford: Oxford University Press.

Sugita, Y., & Tani, J. (2004). A connectionist approach to learn association between sentences and behavioral patterns of a robot. In: S. Schaal, A.J. Jan Ijspeert, A. Billard, S. Vijayakumar, J. Hallam and J.-A. Meyer. (Eds.), *Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior: From Animals to Animats 8,* Cambridge, MA: MIT Press.

Tsiotas, G., Borghi, A., & Parisi, D. (2005). Objects and affordances: An Artificial Life simulation. In: *Proceedings of the XXVII Annual meeting of the Cognitive Science Society*, Stresa.

Tomasello, M. (2002). Some facts about primate (including human) communication and social learning. In A. Cangelosi, & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 327-340), Springer: London.

Tucker, M., & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition*, *8*, 769-800.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., & Thelen E. (2001). Autonomous mental development by robots and animals. *Science*, *291*, 599-600.

Wisniewski, E.J. (1997). When concepts combine. *Psychonomic Bulletin and Review*, *4* (2), 167-183.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, *17* (3-4), 381-397 (Special issue on The Emergence of Language)

Zwaan, R.A. (2004). The immersed experience: toward an embodied theory of language comprehension. In: B.H. Ross (Ed.), *The psychology of language and motivation*, vol. 44. New York: Academic Press.

Zwaan, R.A., Madden, C.J., Yaxley, R.H., & Aveyard, M.E. (2004). Moving words: Dynamic mental representations in language comprehension. *Cognitive Science*, *28*, 611-619.

FIGURE CAPTIONS

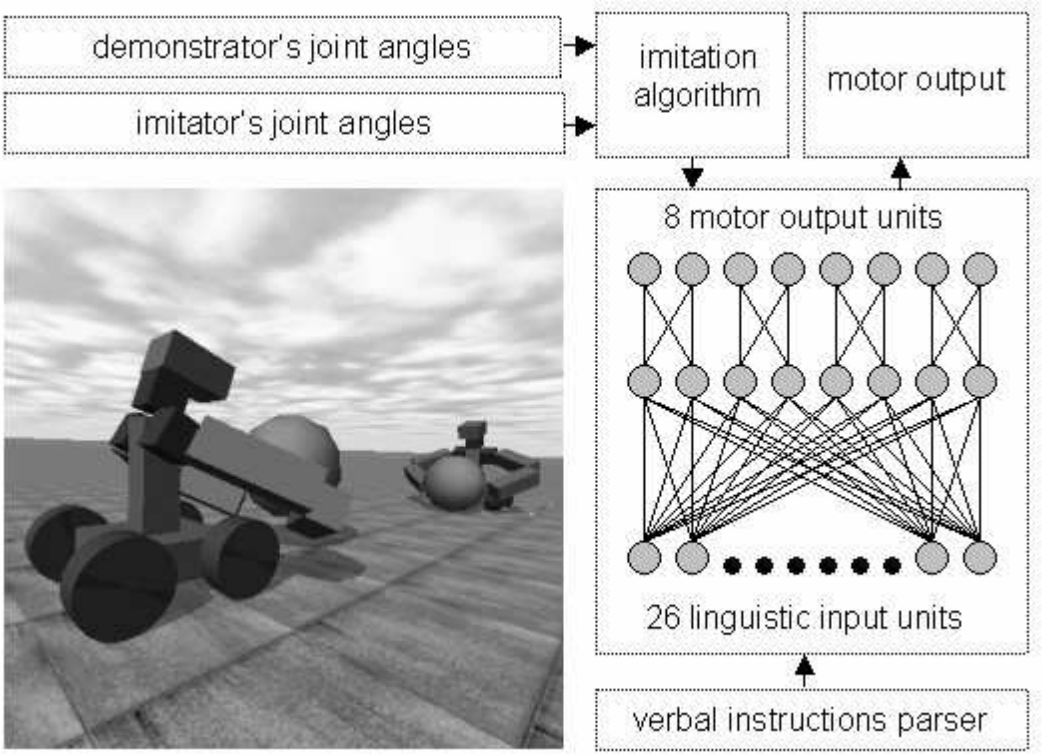

Fig. 1. Functional organization of the robotic model. The picture (bottom left) shows the 3D simulation environment with the demonstrator and imitator robots. The diagram on the right describes the linguistic input from the parser to the neural controller, and the corresponding motor output. The imitation algorithm compares the demonstrator's joint angles with those of the imitator.

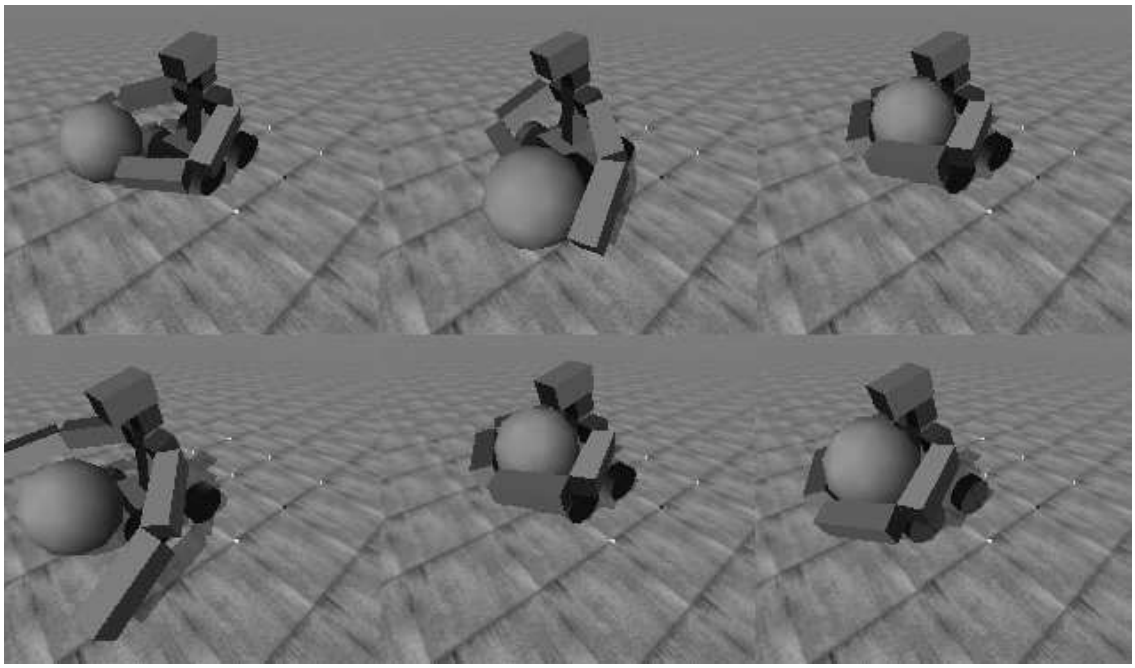

Fig. 2 - Two example sequences for the acquisition of higher-level behaviors HG1 (top row) and HG2 (bottom row). Top row (from left to right): "GRAB is CLOSE_LEFT_ARM and CLOSE_RIGHT_ARM" (BG & BG = HG1). Bottom row: "CARRY is MOVE_FORWARD and GRAB" (BG + HG1 = HG2).

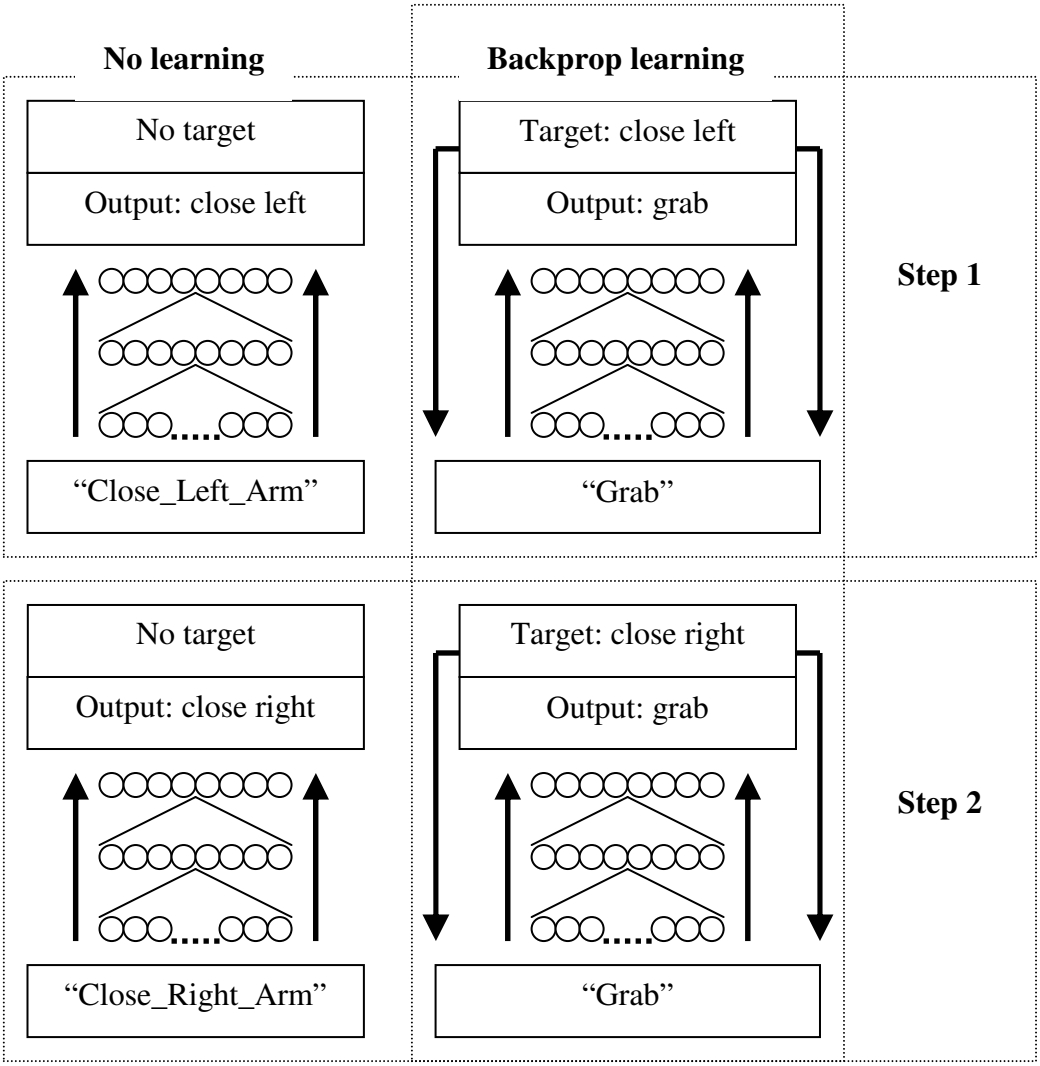

Fig. 3. The procedure that implements the grounding transfer from two basic action words to a combined action word consists of multiple steps, one for each basic word involved. Each of these steps is composed of a feed-forward phase, during which a desirable output is computed, and a learning phase, during which this output is used as a target input for backpropagation learning. Input patterns are binary representations of words, while output patterns are forces applied to each motorized joint.

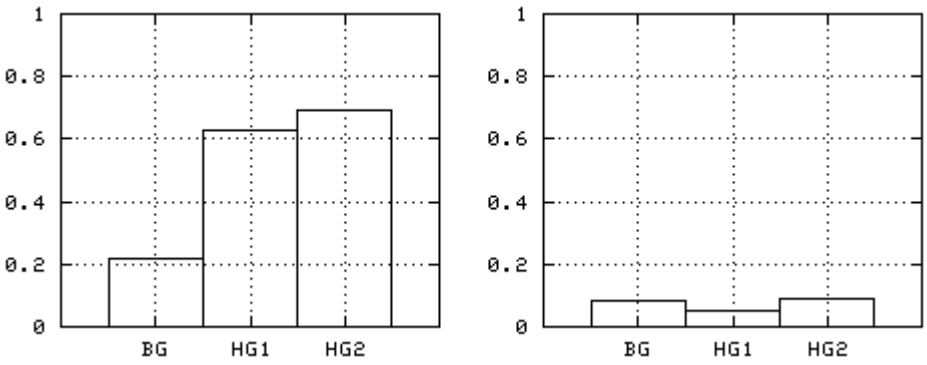

Fig. 4 - Left graph: average posture errors after training for the basic, 1st and 2nd level of word groups. Right graph: final posture errors after training for each level. Data are averaged over the 10 replications.