2010-06

Grounding language in action and perception: From cognitive agents to humanoid robots

Cangelosi, A

10.1016/j.plrev.2010.02.001
Physics of Life Reviews
Elsevier BV

PREPRINT - 37.Cangelosi A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. Physics of Life Reviews, 7(2), 139-151

Grounding Language in Action and Perception: From Cognitive Agents to Humanoid Robots

Angelo Cangelosi
Centre for Robotics and Neural Systems
University of Plymouth (UK)
acangelosi@plymouth.ac.uk

Abstract
In this review we concentrate on a grounded approach to the modeling of cognition through the methodologies of cognitive agents and developmental robotics. This work will focus on the modeling of the evolutionary and developmental acquisition of linguistic capabilities based on the principles of symbol grounding. We review cognitive agent and developmental robotics models of the grounding of language to demonstrate their consistency with the empirical and theoretical evidence on language grounding and embodiment, and to reveal the benefits of such an approach in the design of linguistic capabilities in cognitive robotic agents. In particular, three different models will be discussed, where the complexity of the agent's sensorimotor and cognitive system gradually increases: from a multi-agent simulation of language evolution, to a simulated robotic agent model for symbol grounding transfer, to a model of language comprehension in the humanoid robot iCub. The review also discusses the benefits of the use of humanoid robotic platform, and specifically of the open source iCub platform, for the study of embodied cognition.

1. Introduction

In this review we will concentrate on the proposal of a grounded approach to the modeling of cognition through the methodology of cognitive agents and developmental humanoid robotics. In particular, this work will focus on the modeling of the evolutionary and developmental acquisition of linguistic capabilities based on the principles of symbol grounding (Harnad 1990; Cangelosi & Harnad 2000). We propose a cognitive modeling approach to the grounding of language that is consistent with embodied cognition theory. The computational approach will first be framed within the growing empirical and theoretical literature on embodied cognition. This includes experimental psychology work on the grounding of language in perception and action, neuroscience and brain imaging studies on language embodiment and constructivist developmental psychology theories. We then review some examples of cognitive agent and developmental robotics models of the grounding of language to demonstrate the intrinsic link between language and cognition, and to reveal the benefits of such an approach in the design of linguistic capabilities in humanoid cognitive robots. These models provide a theoretical platform for the development of grounded language systems that overcome the known shortcoming of

symbolic-only theories of language and cognition (e.g. Fodor 1975; Burgess & Lund 1997; Landauer & Dumais 1997; Kirby & Hurford 2002).

1.1 Empirical evidence on the grounding of cognition

There is growing empirical and theoretical evidence on the embodied and grounded approach to cognition in natural and artificial cognitive systems. Recent advances in cognitive psychology, neuroscience, cognitive linguistics and developmental psychology support an embodied view of cognition, i.e. the fact that cognitive functions (e.g. perception, categorization, reasoning, and language in particular) are strictly intertwined with sensorimotor and emotional processes (Wilson 2002). This is particularly evident in numerous experimental psychology studies on the grounding of language, and other cognitive capabilities, in action and perception (Pecher & Zwann 2004).

One of the main theories that explains how symbols are grounded in perception and mental simulations is the Perceptual Symbol Systems hypothesis (Barsalou, 1999). This hypothesis states that perceptual experience captures bottom-up patterns of activation in sensorimotor areas, through association in the brain processing sensory multi-modal information. At the same time, top-down mechanisms use association areas in the brain to partially reactivate sensorimotor areas to implement dynamic simulations of perceptual symbols. Mental simulators, based on distributed and interconnected brain areas, implement a basic conceptual system that supports categorization, produces categorical inferences, and supports productivity, propositions, and abstract concepts. In a similar theoretical line, Coventry and Garrod (2004) propose the on-line activation of situation-specific models for tasks involving spatial cognition and spatial language judgments. For example, they have extensively studied the use of the locative prepositions such as “under” and “below” for describing a visual scene depicting a person holding an umbrella under pouring rain. Experimental and modeling evidence (e.g. Coventry et al. 2001; Cangelosi et al., 2005; Coventry et al. in press) has shown that the human participants’ evaluation of the scene (e.g. rating the relevance of specific spatial prepositions such as “The man is *under* the umbrella” or “The man is *below* the umbrella”) depends on a series of factors grounded on the participants’ previous experience, and on the input stimuli involved in the spatial cognition task. Psycholinguistic experiment results consistently demonstrate that participants prefer to use the preposition “under” to describe scenes in which the protective function of the umbrella is fulfilled (the person under the umbrella stays dry). When instead the umbrella fails to protect the person (who is wet because rain is coming from the side and wets its body), they prefer the use of the preposition “below”. Coventry and Garrod (2004) identify a combination of factors affecting the grounding of preposition in cognition and embodiment. These regard geometric knowledge (such as the relative orientation of an umbrella respect to the direction of the rain and to the position of the human being protected), object-specific knowledge (e.g. the prototypical rain protection function performed by an umbrella) and force dynamics factors (e.g. physics knowledge on the direction of the rain).

The focus on the grounding of language into action and sensorimotor knowledge has been extensively studied by Glenberg and collaborators (Glenberg & Kaschak, 2002; Borghi et al. in press). Glenberg developed an embodied theory of cognition where meaning consists of the set of actions that are a function of the physical situation, of how our bodies work, and of our experiences. Glenberg investigated the Action-sentence Compatibility Effect. This consist in the fact that during sentence comprehension tasks, human 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. Participants are then faster to respond to sentences implying movement away from the body (e.g. “Close the drawer”) when the response requires moving away from the body. This phenomenon is used to support the Indexical Hypothesis (Glenberg & Robertson, 2000). This suggests that when reading 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 to derive affordances from the object or its perceptual symbol. Finally, the third process is to mesh the affordances into a coherent set of actions. The mesh process is guided by the syntax of the sentence being processed, thus explaining the action-sentence compatibility effects.

Cognitive psychology evidence is highly consistent with neuroscientific studies on action and language grounding and related brain mechanisms. One of the main neuroscientific hypotheses supporting grounding is the involvement of the mirror neuron system for action and language learning and evolution. Rizzolatti and Arbib (1998) argue for a role of the mirror neuron system in humans, and its precursor in monkey’s brain (area F5), to explain the origins of language. The mirror neurons match observed events to internally generated actions, thus forming a link between the observer and the actor. Rizzolatti and Arbib suggest that such an observation/execution matching system provides a necessary bridge from action to language. Gallese and Lakoff (2005) further use mirror neuron evidence to show that language makes direct use of the same brain structures used in perception and action. They hypothesize 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.

Brain imaging studies also support the intrinsic link between action and language. For example, Cappa and Perani (2003) have reviewed literature on the neural processing of verbs and nouns. They found a general agreement on the fact that the left temporal neocortex plays a crucial role in lexical-semantic tasks related to the processing of nouns, whereas the processing of words related to actions (verbs) involves additional regions of the left dorsolateral prefrontal cortex. Pulvermuller and collaborators (Pulvermuller 1993; Hauk et al. 2004) consistently show that words activate cortical areas in a somatotopic fashion. For example, their brain imaging experiments show that action words referring to face, arm, or leg actions (e.g., “to lick”, “pick”, or “kick”) differentially activate areas along the motor areas directly adjacent to, or overlapped with, the cortical motor areas activated respectively by actual movement of the tongue, fingers, or feet. These observations support a dynamic view of language grounding, where words are processed by distributed neuronal assemblies, and these are based on cortical topographies that reflect word semantics.

Finally, in cognitive linguistics and developmental psychology, the link between the properties of language and their relationship with cognitive processes has been extensively supported. In linguistics, this link has been formalized in cognitive and constructivist linguistic theories (e.g. Talmy, 1980; Goldberg 2006). In developmental psychology, Tomasello (2003) has proposed a usage-based view of language, where linguistic development is intrinsically related to children's general social, cognitive, and symbolic development. As such, linguistics and grammatical abilities are gradually constructed by the child during development. For example, Tomasello's Verb Island hypothesis (Tomasello 1992) argues that the child's earliest grammatical organization is verb-specific. Initially children use grammatical constructions centered on separated, individual verb items reflecting specific core meanings. Gradually, children acquire a general construct of verb through the merging of verb islands with similar meanings and syntactic constructs. This view is therefore consistent with embodied and grounded approaches where linguistic structure reflects action and perceptual experiences.

1.2 The Grounding of Symbols in Cognitive Agents

This growing empirical evidence on language grounding in action, perception and cognition is highly consistent with computational modeling approaches to language grounding in artificial cognitive agents and robots (Cangelosi et al. 2005; Cangelosi et al. 2008), as well as with other recent modeling approaches that focus on the strict integration of sensorimotor and cognitive capabilities (Perlovsky 2001, 2004, 2009; Feldman & Narayanan 2004; Steels 2003; Cangelosi et al. 2005).

In the modeling approach presented here, cognitive agents, be they simulated agents or humanoid robots, learn symbols (words) that are directly grounded into the agents' own categorical representations, whilst at the same time having logical (e.g. syntactic) relationships with other symbols. This view has some important implications. First, each symbol is directly grounded into internal sensorimotor categorical representations. These representations include perceptual categories (e.g. the concept of blue color, square shape, or female face), sensorimotor categories (e.g. the action concept of grasping, pushing, and carrying), social representations (e.g. individuals, groups and relationships) and other categorizations of the agent's own internal states (e.g. emotions and motivations). These categories are connected to the external world through our perceptual, motor and cognitive interactions with the environment. Secondly, symbols also have logical (e.g. syntactic) relationships with the other symbols of the lexicons used for communication. This allows symbols to be combined, using compositional rules such as grammar, to form new meanings. For example, the combination of the two symbols "stripes" and "horse", which are directly grounded into the agent's own sensorimotor experience of striped objects and horses in its environment, produces the new concept (and word) "zebra". This new symbol becomes indirectly grounded in the agents' experience of the world through the process of "symbol grounding transfer" (Cangelosi & Riga 2006).

This approach satisfactorily addresses the main issues identified in the symbol ground grounding problem (Harnad 1990; Steels 2008; Cangelosi 2009). The symbol grounding problem, as initially stated by Harnad (1990), refers to the capability of natural and artificial cognitive agents to acquire an intrinsic (autonomous) link between internal symbolic representations and some referents in the external word or internal states. In addition, Harnad explicitly proposes a definition of a symbol that requires the existence of logical links (e.g. syntactic) between the symbols themselves. These symbolic links support the phenomena of productivity and generativity in language, and contribute to the grounding of abstract concepts and symbols (Barsalou 1999). Moreover, an important component of the symbol grounding problem is its social and cultural dimension, where social interactions contribute to the sharing of symbols (a.k.a. the external/social symbol grounding problem, as in Cangelosi 2006 and Vogt 1997).

In addition to the focus on the grounding of the linguistic symbols in action and perception, the proposed approach shares the following cognitive modeling principles:

(i) *Embodiment and situatedness*: Cognitive agents and robots act in their environment and develop their own cognitive capabilities through interaction with their physical and social world, mediated by the agent's own sensorimotor structure. Agents acquire categorical representation of their environment (objects, events) and of their own interaction with the world (actions, internal states, motivations, emotions). These will then constitute the internal categorical representation upon which symbols are grounded.

(ii) *Evolutionary and developmental learning*. Population of agents can acquire sensorimotor and cognitive capabilities as a result of the evolutionary adaptation to social and physical environments. During developmental learning, individual agents gradually acquire complex sensorimotor, cognitive and linguistic capabilities via dynamic brain-inspired learning processes, following qualitative developmental stages. The interaction between phylogenetic and ontogenetic learning can lead to bootstrapping mechanisms, such as in the case of the Baldwin effect (Munroe & Cangelosi, 2003).

(iii) *Bottom-up cognitive bootstrapping*: The individual capabilities of the agent's cognitive system closely interact during development and their complex interaction leads to the emergence and bootstrapping of higher-order cognitive capabilities.

(iv) *Bio-inspired agent design*: The design of the agent's control system and behavioral capabilities is constrained, and directly inspired, by psychologically-plausible cognitive development processes and by neuroscientific investigations of behavior control in humans and animals.

The modeling approach to grounding presented here is not the only computational and robotic approach to grounding that has been proposed in the literature (see reviews in Cangelosi & Parisi 2002; Lyon et al. 2005). For example, Steels (1999; 2005) has proposed

various robotic models of the emergence of communication based on the languages games for the Talking Heads experiments and the AIBO and QRIO robots. In Steels's approach relevance is given to the social aspects of the symbol grounding, as well as the perceptual grounding of categories (Steels & Belpaeme, 2005). Vogt (2002) has put forward a similar approach applied to both robots and simulated scenarios. Fontanari and Perlovsky (2007) use language games to study evolution of compositional lexicons. Moreover, the literature on computational modeling of language evolution includes numerous works on multi-agent that do not directly follow a grounded approach, i.e. where the semantic repertoire of the agent is pre-defined by the researcher and based on explicit symbolic representation (e.g. Kirby & Hurford 2002; Brighton et al. 2005).

The distinctiveness of the grounding approach presented here, in comparison with other approaches, is that categorical representation (meanings) emerge as a results of a sensorimotor task that the agents perform to survive in the environment or to imitate a teacher. As the agent's control architecture is always based on an artificial neural network, the meanings consist of distributed internal representation, as with hidden unit activation patterns. The agent's neural network is designed to integrate all sensorimotor, cognitive and linguistic capabilities into one cognitive architecture, and to ground language comprehension and language production into internal categorical representations (Cangelosi 2005). The typical neural architecture used in these models (Figure 1) has two sets of inputs, namely vision/perceptual units and linguistic/speech units. In output there is a set of units to control the agent's motor behavior, and a set of units to produce linguistic utterances. The input and output units are integrated in at least one layer of hidden units. These units, and the connections to/from the hidden layer, allow the grounding of symbols into distributed categorical representations. The organization of the internal units, and the patter of connections from/to these hidden layers, can follow different modular architectures. For example, in evolutionary models on the learning of nouns and verbs, different modular architectures can be designed taking inspiration from the brain organization (Cangelosi & Parisi 2004).

With this neural network approach to language grounding, there is no need to postulate the existence of explicit syntactic rules or syntactic roles in the agent's control architecture. Connectionist neural network architectures are known to be able to learn to process syntactic and recursive structures, where syntactic and semantic structures can emerge from the input training stimuli (Elman, 1990). In the agent grounding method, when agents learn to use compositional lexicons, as in the evolutionary model of proto-nouns and proto-verbs (Cangelosi 2001; Cangelosi & Parisi 2004), the symbol composition capabilities implicitly reside in the network's processing capability, in a distributed manner based on connection weights and activation states.

< Figure 1 here >

In the remaining part of the paper we review some of our recent models of the grounding of language in cognitive agent and developmental robotics to provide detailed example of the proposed approach and to demonstrate its consistency with the above empirical and theoretical evidence. In particular, three different models will be discussed where the complexity of the agent's sensorimotor and cognitive system gradually increases: (i) a multi-agent simulation of language evolution, (ii) a simulated robotic agent model for symbol grounding transfer, (iii) a model of language comprehension in the humanoid robot iCub.

2. Multi-Agent Modeling of Grounding and Language Evolution

The first modeling approach on which we will concentrate is based on simulated multi-agent systems. The model implies a population of agents who interact with each other, and develop sensorimotor and cognitive capabilities through a combination of evolutionary and developmental learning. Agents are modeled as simple simulated organisms which perform foraging tasks (e.g. in a mushroom world metaphor) and communicate with other agents about their interaction with the world (e.g. location of food, their perceptual features, and actions to act on the world). The agents do not possess a realistic representation of their body, as in the humanoid agents and robots described in the next sections, but only have perceptual visual input and motor units controlling simple actuators to navigate the world and perform simple actions on the objects. Even with this simple embodiment system, agents are able to ground communication signals into their own sensorimotor representations of the world and their interaction with it.

Various multi-agent simulations of the evolution of language based on this modeling framework have been developed. In some models, the focus has been on the emergence of a shared lexicon using only evolutionary learning methods (Cangelosi & Parisi 2008) or a combination of evolutionary and ontogenetic learning (Cangelosi 2001; Cangelosi & Harnad 2000; Munroe & Cangelosi 2003). Given the focus of this review on the grounding approach, we describe here in detail one of these evolutionary models on the Symbolic Theft Hypothesis of language origins (Cangelosi and Harnad 2000). This hypothesis considers two competing ways of category learning: (i) by "sensorimotor toil", when new categories are acquired through slow trial and error and feedback experience with the environment; (ii) by "symbolic theft", when new categories are acquired quickly through hearsay from linguistic descriptions provided by language-speaking adults. The significant adaptive advantage of symbolic theft has been hypothesized to produce an adaptive benefit for language and can help explaining the origins of language (Harnad, 1996). Cangelosi and Harnad (2000) developed a computational model based on an evolutionary foraging task using a mushroom world scenario (Cangelosi & Parisi 1998). The agents' survival depends on learning the differences between two categories of foods: mushrooms with visual feature A are to be eaten; mushrooms with feature B are to have their location marked, and mushrooms with joint features A and B are to be eaten, marked and returned to. Mushrooms also have other irrelevant features (C, D and E) that the foragers must learn to ignore. When organisms approach a mushroom, they emit a call associated with their

functionality (“EAT”, “MARK”). Both the correct action pattern (eat, mark) and the correct call (“EAT”, “MARK”) are learned during the foragers’ lifetime through supervised learning (sensorimotor toil). Under some conditions, the foragers also receive the call of another forager as input. This will be used to simulate theft learning of the return behavior.

The behavior of an agent is controlled by an artificial neural network, with a structure similar to that of figure 1. This includes input units encoding the visual features A-E and the linguistic/symbolic descriptions proposed by other agents. The output units encode the agent’s sensorimotor behavior (move to approach mushrooms, and eat/mark actions) and its linguistics productions. The population of foragers is also subject to selection and reproduction through a genetic algorithm, where the fitness reflects the number of mushrooms correctly eaten. The genotype of each agent consists on the connections weight of its neural network controller.

To test the adaptive advantage of symbolic theft versus sensorimotor toil, the foragers’ behavior for the two learning conditions is compared. In one simulation the two strategies are compared in direct competition. In the first generations, all organisms learn through sensorimotor toil to eat mushrooms with feature A and to mark mushrooms with feature B. They also learn the names of the basic categories. The return behavior and its name are not yet taught. After the agents evolve this behavior, the agents are divided into two groups of toilers and thieves. Toil foragers learn to return to AB mushrooms through honest but slow toil. In contrast, Theft foragers learn to return by hearing the vocalization from the other agents (e.g. “EAT” + “MARK” = “RETURN”). Results show that thieves successfully return to more AB mushrooms than Toilers. This means that learning to return from the grounded names “EAT” and “MARK” is more adaptive than learning it through direct toil based on sampling the physical features of the mushrooms. This result is also confirmed when thieves and toiler are in direct competition against one another at the time of the selection. Thieves consistently outnumber toilers in all tests. These results support the original hypothesis that a symbolic theft learning strategy, based on language hearsay, is much more adaptive that a sensorimotor toil strategy. This adaptive advantage could have constituted the basis for the origin of language and its adaptive advantage.

3. Symbol Grounding Transfer in Simulated Robotic Agents

The second approach focuses on action and language learning in cognitive agents, where simulated robotic agents can autonomously evolve grasping and object manipulation skills (Massera, Cangelosi & Nolfi, 2007) and learn to use name of actions to acquire higher-order sensorimotor categories (Cangelosi & Riga, 2006; Cangelosi et al. 2007). This model is a direct follow-up of the previous model, as it focuses on the theft acquisition of categories and their names. In particular, its aim is to study the mechanisms for the transfer of the grounding from the basic categories, directly grounded through a sensorimotor toil strategy, to higher-order compositional categories, acquired only via theft language description.

The model differs from the evolutionary multi-agent system of Cangelosi and Harnad (2000) mainly because it is based on a two-agent scenario (teacher and learner) and it models the ontogenetic learning of language, rather than evolutionary learning. In addition, one important difference is that this new model is based on a richer implementation of the organism's embodiment system and sensorimotor knowledge. Instead of simulating a foraging task in a two-dimensional environment, this new approach is based on the simulation of a 3D robot manipulating objects with its arms and fully embodied in a physics simulation environment. This system is implemented using ODE (Open Dynamics Engine, www.ode.org), a high performance library for simulating rigid body dynamics. This permits a richer investigation of the robotic agent embodied system, such as with the simulation of continuous joints activation, and subject to the principle of physics of object collision dynamics.

The model consists of two simulated robotic agents (teacher and learner) embedded within a virtual simulated environment (Fig. 2). The body configuration is loosely inspired by a humanoid robot, though with simplified arms and no legs. Each robot consists of two 3-segment arms attached to a torso, with 6 degrees of freedom. The torso is connected to a base with four wheels. Through the two arms, the robot can interact with the environment and manipulate objects placed in front of it. Three objects were used in the current simulation: a cube, a flat cuboid horizontally orientated, and a vertical bar. The agent receives in the visual input different retina views of each object and has to learn six basic actions: to lower right shoulder, to lower left shoulder, to close right upperarm, to close left upperarm, to close right elbow, and to close left elbow. They will also learn a name for each of these basic actions. The robot had to learn to systematically associate an action with some of the above objects. For example, the close_left and close_right shoulder actions are associated with different views of the cube.

< Figure 2 here >

The first agent, the teacher, is pre-programmed to demonstrate a variety of basic actions, each associated with a linguistic signal. These are demonstrated to the second robot, the learner, which will learn to reproduce the actions by imitation and name them. First the agent acquires basic actions by observing the teacher, and then it learns the basic action names. This is the "direct grounding" strategy, as in the sensorimotor toil strategy. Subsequently, the agent autonomously uses the linguistic symbols that were grounded in the previous learning stage to acquire new higher-order actions. This is the "symbol grounding transfer", corresponding to the symbolic theft strategy.

The robot is controlled by a neural network, as in the neural network architecture described above (Figure 1). The robot's neural network is trained though the error backpropagation algorithm. The agent undergoes 3 stages of training: (1) Basic action learning, (2) Entry-level naming through direct grounding, and (3) Higher-order learning through grounding

transfer. For basic action learning, the agent learns to imitate all six basic actions in association in response to the presentation of the different objects. No linguistic labels are used at this stage. In the Entry-level naming of basic grounding, the agent learns to associate the previously acquired behaviors to linguistic signals. This process consists of three sequential activation cycles. The first cycle (Linguistic Production), the learner is trained to use the names of the 6 basic actions. In the second Entry Level cycle (Linguistic Comprehension), the learner is taught to respond to the input of the name of the action, without having the ability to perceive the object associated to the action. In the final Entry-level cycle (Imitation), both motor and linguistic inputs were activated in input, and the network learns to reproduce the action in output and activate the corresponding output linguistic unit. This third cycle is necessary to permit the linking of the production and the comprehension tasks in the hidden units activation pattern (Cangelosi et al. 2000).

The final training stage, Higher-level learning, allows the learner to autonomously acquire higher-order compositional actions without the need of a visual demonstration from the teacher. This is achieved only through a neural computational model of the “mental simulation” strategy similar to Barsalou’s (1999) perceptual symbol system hypothesis. The teacher only has to provide new linguistic instructions consisting of the names of two basic actions and the name of a new higher-order action. For example, a sentence consists of “grab [is] close_left_arm [and] close_right_arm”. The learner goes through four higher order learning cycles. First it activates only the input unit of the first basic action name to produce and memorize the corresponding sequence. Second, it activates in input the linguistic units for the first basic action name and the new higher-order action. The resulting output motor activations are compared with the previously stored values to calculate the error and apply the backpropagation weight corrections. The next two cycles are the same as the first two, except that the second basic action name unit is activated as well.

To establish if the agent has learned correctly the new high-order actions and transferred the grounding from the basic action names to the new higher order names, a grounding transfer test is performed. This test aims to evaluate the capability of the learner agent to perform a new compositional action with any of the objects previously associated, in the absence of the linguistic descriptions of the basic actions. Thus the agent is requested to respond solely on the signal of the composite action (e.g. “Grab”) with specific views of the objects.

Analyses of the simulation results indicate that all agents were able to learn successfully the 6 basic actions and the 3 higher-order behaviors. At the end of the stage, the learner was able to execute all actions flawlessly, when presented with an object. In the grounding transfer test, the agent correctly reproduced the higher-order compositional action. This test demonstrates the capacity of the agent to transfer autonomously the grounding form the basic actions to new actions never seen during the demonstration cycles.

To understand the neural network mechanism that permit the higher-order symbol grounding transfer, we have analyzed the contribution of the hidden unit activations and the combination of input/output patterns during the various learning stages. These analyses

were first carried out in a simulation of grounding transfer for perceptual categories (e.g. “zebra” higher order category as a result of the combination of the grounded names of “stripes” and “horses”) (see Cangelosi et al. 2001). What these analyses tell us is that the transfer of grounding from the low level categories to the higher level ones is mediated by the hidden representations. These representations divide the network's semantic space into different regions, due to categorical representation effects. The regions tend to have high inter-categorical distances. The effect of training during entry-level learning is that of creating links between well differentiated categorical representations and discrete symbols. When these symbols are combined together, they also inherit their links to low level categorical representations.

This model has further been extended to scale the number of actions and words that the agents learn. The issue of scaling up and combinatorial complexity in cognitive systems has been recently addressed by Perlovsky (2001). In linguistic systems, combinatorial complexity refers to the hierarchical combinations of bottom-up perceptual and linguistic signals and top-down internal concept-models of objects, scenes and other complex meanings. Perlovsky has proposed the neural Modeling Field Theory (also known as Dynamic Logic) as a new method for overcoming the exponential growth of combinatorial complexity in the computational intelligent techniques traditionally used in cognitive systems design. Modeling Field Theory is based on the principle of associating lower-level signals (e.g., inputs, bottom-up signals) with higher-level concept-models (e.g. internal representations, categories/concepts, top-down signals) avoiding the combinatorial complexity inherent to such a task. This is achieved by using measures of similarity between concept-models and input signals together with a new type of logic, so-called dynamic logic. MFT may be viewed as an unsupervised learning algorithm whereby a series of concept-models adapt to the features of the input stimuli via gradual adjustment dependent on the fuzzy similarity measures.

Perlovsky (2004) has suggested the use of Modeling Field Theory specifically to model linguistic abilities. By using concept-models with multiple sensorimotor modalities, a Modeling Field Theory system can integrate language-specific signals with other internal cognitive representations. To test this hypothesis, the above agent scenario was extended to a repertoire of 112 actions (Cangelosi et al. 2007; Tikhanoff 2009). The learner robot uses MFT to learn to reproduce those actions as well as to learn the actions names. These actions are inspired by the semaphore flag signaling alphabet. For the encoding of the actions, we collected data on the posture of the teacher robots using 6 features, i.e. 3 pairs of angles for the two joints of the shoulder, upper arm and elbow joints. In this simulation, objects are not present. When performing the action, the teacher agent can emit a three-letter word labeling the action (using 6 phonetic features). The learning of actions and their words is based on the Modeling Field Theory algorithm (see Cangelosi et al. 2007 for details on the algorithm parameters). The simulation lasts for 25000 training steps. In the first 12500 cycles, only the 6 action features (joint angles) are provided and the agent is capable to learn to reproduce all actions. In the second part of the training, all 12 feature sets (6 for actions/angles, 6 for phonetic sounds) are considered when computing the MFT

fuzzy similarity functions. Results demonstrate that the robot is able to categorize 95% of actions and learn their unique labels, thus supporting the use of Modeling Field Theory to scale up the action and lexicon repertoire.

4. Language Comprehension in the Humanoid Robot iCub

In this section we briefly introduced a more recent model of symbol grounding and language learning based on developmental (epigenetic) robotics (Lungarella et al. 2003; Weng et al. 2001). One major innovation, in comparison with previous approach, is the use of a realistic model of the iCub humanoid robot, a new robotic platform for cognitive modeling research. Secondly, we implement a scaled up version of the lexicon acquired by the agent, where the robot is trained to acquire a grounded lexicon to understand object manipulation instructions such as “grasp blue cube” or “put red sphere in container” (Tikhanoff 2008; Tikhanoff et al., submitted).

The iCub robot is an open-source humanoid robot platform designed to facilitate developmental robotics research (Metta et al. 2008; robotcub.org). The robot is 1.05 meter tall, and weighs approximately 20.5 kilograms (Figure 3 left). The robot has 53 degrees of freedom (DoF). In particular, for what regards manipulation tasks, it has 7 DoF for each of the two arms and 9 DoF for each of the two hands. A computer simulation model of the iCub has also been developed (Figure 3 right). The simulated iCub (Tikhanoff et al. 2008) has been designed to reproduce, as accurately as possible, the physics and the dynamics of the physical iCub. It consists of multiple rigid bodies connected via joints. The body parts were designed following the open-source robot specifications. It has the same number of DoF (53) of the real robot. In addition, the environment parameters on gravity, objects’ mass, friction and joints are based on known environment conditions.

< Figure 3 here >

The language learning model with the simulated iCub was based on a modular architecture that controls the various cognitive and sensorimotor capabilities of the iCub and integrates them to ground symbols into the agent’s own cognitive system. This consists of two neural network controllers respectively for the reaching of objects in the peripersonal space of the robot and for the grasping of objects with one hand. Visual and speech input processing was based on standard vision algorithms and speech recognition systems. For full details on the motor, vision and speech processing modules (see Tikhanoff 2008; Tikhanoff, Cangelosi & Metta, submitted). Here we will focus on the central module that integrates the various processing capabilities to respond to linguistic instructions. The cognitive integration module was based on a goal selection neural network. This is an architecture similar to the general feedforward network shown in Figure 1, though in this first model only the language comprehension pathway was implemented. The input to the network consists of seven visual features encoding object size and location, and language input units for the speech

signals. The output layer has four units respectively selecting the four actions: idle, reach, grasp, and drop. The active output action unit would then activate the reaching and grasping module, trained separately. The hidden layer has fifteen units. During the training phase, the robot is shown an object along with a speech signal. Examples of sentences for handling a blue object are: “Blue_ball”, “Reach blue_ball”, “Grasp blue_ball”, “Drop blue_ball into basket”

The Goal Selection feed-forward neural network was trained with the above setup using the error backpropagation algorithm. After 5000 training cycles, the network reaches a satisfactory learning error. To test the final language comprehension capabilities, an individual object was first put in front of the robot, such as a blue cube. The robot stays in idle position until a speech signal is presented in input. This activated the goal selection network that selected the appropriate motor sequence response. For example, if the linguistic instruction was “grasp blue_ball”, the reaching and grasping modules are activated appropriately. The iCub first centers the head on the object, to calculate the three dimensional coordinates of the hand and of the position of the object through binocular disparity comparison. It then initiates the reaching action, and subsequently the object is grasped.

This model provides an implementation of the symbol grounded model of language comprehension in the iCub robot. Work is on going to implement the various modules to the physical iCub platform. Although this requires the retraining of some of the modules with new data from the physical robot, some of the training achieved in simulation is utilized in the real robot. For example, the training in simulation of the neural network for reaching permits the sampling of a greater number of positions, thus increasing the generalization and robustness capabilities of the neural controller. In addition, another grasping model (Macura et al. 2009) has recently been developed to train the iCub to grasp object of different sizes and learn to responds selectively to different object microaffordances, such as a power grasp for large objects (apples, tennis ball) and a precision grip for smaller objects (cherry, marble).

5. Conclusions

The cognitive modeling approach presented here provides an important methodology for investigating the processes and mechanisms supporting the grounding of language into sensorimotor experience, during evolution and ontogenetic development. Current models put great emphasis on the direct grounding of symbols into internal categorical representations, i.e. when grounded symbols are acquired by the agents through embodied interactions with the environment. However, with this modeling approach we do not invoke a cognitive system in which all symbols have to be *directly* grounded into perceptual and sensorimotor toil interactions with the environment. Rather, we argue that only the initial lexicon developed by language-speaking agents must be acquired through direct grounding. Once the agent develops a core lexicon that allows it to represent their

fundamental sensorimotor properties of the world, they can expand their own lexicon through the process of language productivity and symbol grounding transfer, as in the symbolic theft strategy demonstrated in Section 3. In addition, the integration of the grounded approach to language learning with current developmental robotics models on cognitive and sensorimotor development (e.g. robotics models of motivation and curiosity, and of theory of mind – e.g. see review in Lungarella et al. 2003) permits the design of complex cognitive skills in interactive robots.

The directly grounded basic lexicon has been called the “grounding kernel” (Harnad 2009), i.e. the essential set of words that cannot be defined by linguistic combinations (descriptions) of other words in the dictionary. Therefore, the grounding kerned must come from direct grounding during early stages of language development. Harnad and collaborators (Blondin-Massé et al. 2008; Chicoisne et al. 2008) have recently carried out linguistic corpora analyses on the Longman’s Dictionary of Contemporary English and Cambridge International Dictionary of English to identify the grounding kernel of English. They analyzed the concreteness, imagery and age of acquisition factors in their grounding kernel words and predicted that the grounding kernel would be more concrete and imaginable, and was acquired earlier than the rest of the vocabulary. Analyses support the hypothesis that dictionary meanings are grounded in a mental lexicon that is in turn grounded in prior sensorimotor learning.

The use of the grounded agent methodology discussed here, in parallel with computational linguistic analyses and other developmental robotics approaches (Weng et al. 2002), can help investigating both the evolutionary and ontogenetic factors affecting the acquisition of the grounding kernel, and its relationship to embodiment factors. For example, one of the open issues in language grounding research regards the scaling up of the agent’s lexicon from few tens of words, as in current models, to lexica with hundreds of words and with complex syntactic constructs. The increase of the agents’ lexicon also permits the investigation of the transition from holistic communication systems to compositional languages where words belong to different syntactic categories. Grounding models can shed light both on evolutionary phenomena, as well as developmental mechanisms leading to the emergence of syntactic languages and the interaction of phylogenetic and ontogenetic mechanisms as in the Baldwin effect (Munroe & Cangelosi 2003). Finally, another challenge in language modeling research regards the investigation of the grounding of abstract words. Computational models can shed light on the specific sensorimotor and cognitive mechanisms leading to the abstract concepts, such as “beauty”, as well as the grounding of function words such as “and”, “not”, “for” (Nehaniv et al. 2007).

This computational modeling approach has the potential of being applied to various types of cognitive agent models and task scenarios. Work reviewed in previous sections varied from foraging tasks in evolutionary multi-agent systems to action learning scenarios in more realistic agent embodiments, up to an application to the realistic simulation model of the humanoid robot iCub. Most of the models are based on computer simulation of increasing complexity, though the iCub simulation experiments are currently being extended to the physical platform (Tikhanoff et al. submitted). Computer simulations play an important role

in cognitive modeling research (Ziemke 2003; Tikhanoff et al. 2008). Despite the fact that the use of a simulation might not provide a faithful model of the complexity present in the real environment, thus not always guaranteeing a fully reliable transferability of the agent's controller from the simulation environment to the real one, robotic agent simulations are of great interest for cognitive scientists for various reasons. For example, one of the main benefits is the fact that the simulation of embodied agents permits the investigation of several types of embodiment configurations (e.g. with varying sensors and actuators setups) without requiring the availability and manipulation of physical artifacts. Simulators also allow researchers to experiment with robots with varying morphological characteristics (Bongard & Pfeifer 2002). This advantage, in turn, permits the discovery of properties of the behavior of an agent that emerges from the interaction between the robot's controller, its body and the environment (Nolfi & Floreano 2002). Moreover, the iCub simulator has the additional advantage of being open source, and as such allows researchers from a variety of labs to share the same benchmarking cognitive agent platform, without the need to own expensive robot platform.

In addition to contributing to cognitive modeling research, the agent methodology described here can also have significant impact on cognitive systems technology and robots design, such as in service robotics and human-robot interaction. For example, in service robotics, future systems will be able to learn language and world understanding from humans, and also to interact with them for entertainment purposes (e.g. Tikhanoff & Miranda, 2005; Steels & Kaplan 2000). A bio-inspired robotic approach, based on the grounding of language and its direct integration with action knowledge, can help overcome current limitations of robot design. In Cangelosi et al. (under review) a roadmap for future research on action and language integration based on a developmental cognitive robotics approach is proposed. This identifies the key research challenges and milestones for action learning, language development and social interaction in cognitive robots.

6. References

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

Blondin-Massé, A; Chicoisne, G; Gargouri, Y; Harnad, S; Picard, O & Marcotte, O (2008). How Is Meaning Grounded in Dictionary Definitions? In *TextGraphs-3 Workshop - 22nd International Conference on Computational Linguistics*

Bongard J.C. & Pfeifer R. (2003). Evolving complete agents using artificial ontogeny. In Hara, F. & R. Pfeifer, Eds., *Morpho-functional Machines: The New Species (Designing Embodied Intelligence)* Springer-Verlag, pp. 237-258

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

Brighton H., Smith K., Kirby S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, 2(3): 177-226

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

Cangelosi A. (2001). Evolution of communication and language using signals, symbols and words. *IEEE Transactions on Evolutionary Computation*. 5(2), 93-101

Cangelosi A. (2005). Approaches to Grounding Symbols in Perceptual and Sensorimotor Categories. In H. Cohen & C. Lefebvre (Eds), *Handbook of Categorization in Cognitive Science*, Elsevier, pp. 719-737

Cangelosi A. (2006). The grounding and sharing of symbols. *Pragmatics and Cognition*, 14(2), 275-285

Cangelosi A. (2009). The symbol grounding problems has been solved: Or maybe not? *AMD Newsletter*, 6(2): 10-12.

Cangelosi A. et al. (under review). Integration of action and language knowledge: A RoadMap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*.

Cangelosi A., Bugmann G. and R. Borisyuk (Eds.), *Modeling Language, Cognition and Action: Proceedings of the 9th Neural Computation and Psychology Workshop*. Singapore: World Scientific, 2005

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

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. & Harnad S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4(1), 117-142

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. & Parisi D. (Eds.) (2002). *Simulating the Evolution of Language*. London: Springer.

Cangelosi A, Riga T (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots, *Cognitive Science*, 30(4), 673-689

Cangelosi A, Tikhanoff V., Fontanari J.F., Hourdakis E. (2007). Integrating language and cognition: A cognitive robotics approach. *IEEE Computational Intelligence Magazine*, 2(3), 65-70

Cappa, S.F., & Perani, D. (2003).The neural correlates of noun and verb processing. *Journal of Neurolinguistics*, 16 (2-3), 183-189

Chicoisne, G., Blondin-Masse, A., Picard, O. and Harnad, S. (2008) Grounding Abstract Word Definitions In Prior Concrete Experience. In: *Sixth Annual Conference on the Mental Lexicon*, University of Alberta, Banff Alberta, 7-10 October 2008.

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. *Spatial Cognition Conference 2004*, Germany, 11-13 October 2004

Coventry, K. R. & Garrod, S. C. (2004). *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*, Essays in Cognitive Psychology Series, Psychology Press, Hove and New York.

Coventry K.R, Lynott L., Cangelosi A., Knight L., Joyce D., Richardson D.C. (in press). Spatial language, visual attention, and perceptual simulation. *Brain and Language*

Coventry, K. R., Prat-Sala, M. & Richards, L. V. (2001). The interplay between geometry and function in the comprehension of over , under , above and below . *Journal of Memory and Language*. 44, 376-398.

Elman J. L. (1990). Finding structure in time. *Cognitive Science*, 14.179-211

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

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

Fontanari, J. F. and Perlovsky, L. I. (2007) Evolving compositionality in evolutionary language games. *IEEE Transactions on Evolutionary Computation*, 11(6):758--769.

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

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(3), pp. 379-401

Goldberg, A.E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

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

Harnad S. (1996). The origin of words: A psychophysical hypothesis. In B. Velichkovsky B. and D. Rumbaugh (Eds.), *Communicating Meaning: Evolution and Development of Language*. NJ: Erlbaum: pp 27-44

Harnad S. (1990). The Symbol Grounding Problem. *Physica D*, vol. 42, pp. 335-346

Harnad S. (2009) From Sensorimotor Categories to Grounded Symbols. *Technical Report ECS*, University of Southampton.

Hauk, O., Johnsrude, i. & Pulvermuller, f. (2004) Somatotopic representation of action words in human motor and premotor cortex, *Neuron*, 41(2), 301-307

Kirby S. & Hurford J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi A. & Parisi D. (Eds.). *Simulating the Evolution of Language*. London: Springer.

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.

Lungarella M., Metta G., Pfeifer R., Sandini G. (2003). Developmental robotics: A survey. *Connection Science*, 15(4): 151-190

Lyon C., Nehaniv C.L., Cangelosi A. (Eds.) (2007). *Emergence of Communication and Language*. London: Springer.

Macura Z., Cangelosi A., Ellis R., Bugmann D., Fischer M.H. & Myachykov A. (2009). A cognitive robotic model of grasping. *Proceedings of the Ninth International Conference on Epigenetic Robotics*, Venice, November 12-14 2009

Massera G., Cangelosi A., Nolfi S. (2007). Evolution of prehension ability in an anthropomorphic neurorobotic arm. *Frontiers in Neurorobotics*, 1(4)

Metta G., Sandini G., Vernon D., Natale L., Nori F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In R. Madhavan & E.R. Messina (Eds.), *Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08)*. Washington, D.C.

Munroe S., Cangelosi A. (2002). Learning and the evolution of language: the role of cultural variation and learning cost in the Baldwin Effect. *Artificial Life,* 8, 311-339

Nehaniv C. L., Lyon C., Cangelosi A. (2007). Current Work and Open Problems: A Roadmap for Research into the Emergence of Communication and LanguageIn

Lyon C., Nehaniv C.L., Cangelosi A. (Eds.) *Emergence of Communication and Language*. London: Springer.

Nolfi S. & Floreano D. (2000). *Evolutionary Robotics: The Biology, Intelligence and Technology of Self-Organizing Machines Cambridge*, MA: MIT Press/Bradford Books

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.

Perlovsky L. (2001). *Neural Networks and Intellect: Using Model-Based Concepts*. Oxford University Press, New York.

Perlovsky L. (2004). Integrating language and cognition. *IEEE Connections*, vol. 2(2), pp. 8-13

Perlovsky L.I. (2009). Language and Cognition. *Neural Networks*, 22(3), 247-257.

Pulvermuller, F. (2003) *The Neuroscience Of Language: On Brain Circuits Of Words and Serial Order*, Cambridge University Press: Cambridge

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

Steels, L. (2003) Evolving Grounded Communication for Robots. *Trends in Cognitive Sciences*, 7(7):308-312.

Steels, L. (2008) The symbol grounding problem has been solved. So what's next? In de Vega, M., (eds.), *Symbols and Embodiment: Debates on Meaning and Cognition*. Oxford: Oxford University Press. pp. 223-244.

Steels, L. & Belpaeme, T. (2005) Coordinating Perceptually Grounded Categories through Language. A Case Study for Colour. *Behavioral and Brain Sciences*, 28(4):469-489

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

Talmy L. (2000). *Toward a Cognitive Semantics*, Vol. I. Cambridge: Cambridge University Press

Tikhanoff V, Cangelosi A. & Metta G. (2009). Language understanding in humanoid robots: Simulation experiments with iCub platform. submitted

Tikhanoff V, Cangelosi A., Fitzpatrick P., Metta G., Natale L., Nori F. (2008). An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator. In R. Madhavan & E.R. Messina (Eds.), *Proceedings*

of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08). Washington, D.C.

Tikhanoff V. (2009). Development of cognitive capabilities in humanoid robots. *PhD Thesis,* School of Computing, Communications & Electronics, University of Plymouth, U.K.

Tikhanoff V., and E.R. Miranda (2005). Musical Composition by an Autonomous Robot: An Approach to AIBO Interaction. In *Proceedings of TAROS 2005 (Towards Autonomous Robotic System)*, Imperial College, London, UK. pp. 181-188.

Tomasello, M. (1992). *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: CUP

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

Vogt P. (2002). The physical symbol grounding problem. *Cognitive Systems Research*, 3(3): 429-457

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.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9, 625-636.

Ziemke T. (2003). On the role of robot simulations in embodied cognitive science, *AISB Journal*, 1(4), 389-99

PREPRINT - 37.Cangelosi A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. Physics of Life Reviews, 7(2), 139-151

Grounding Language in Action and Perception: From Cognitive Agents to Humanoid Robots

Angelo Cangelosi

Figures

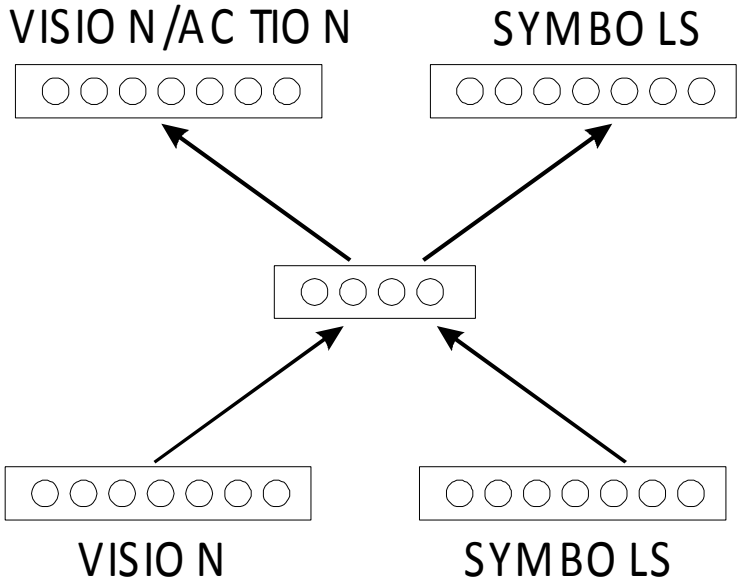

Figure 1. Dual-route architecture for agent's neural network

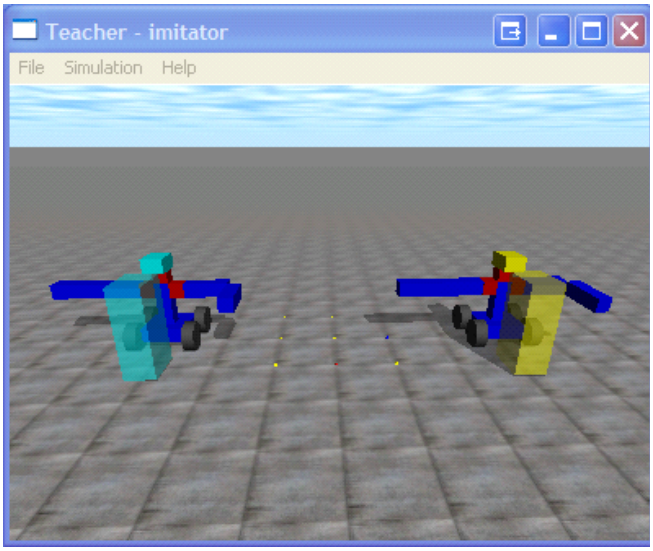

Figure 2: Experiment with simulated robot. Teacher (left) and learner (right) agents.

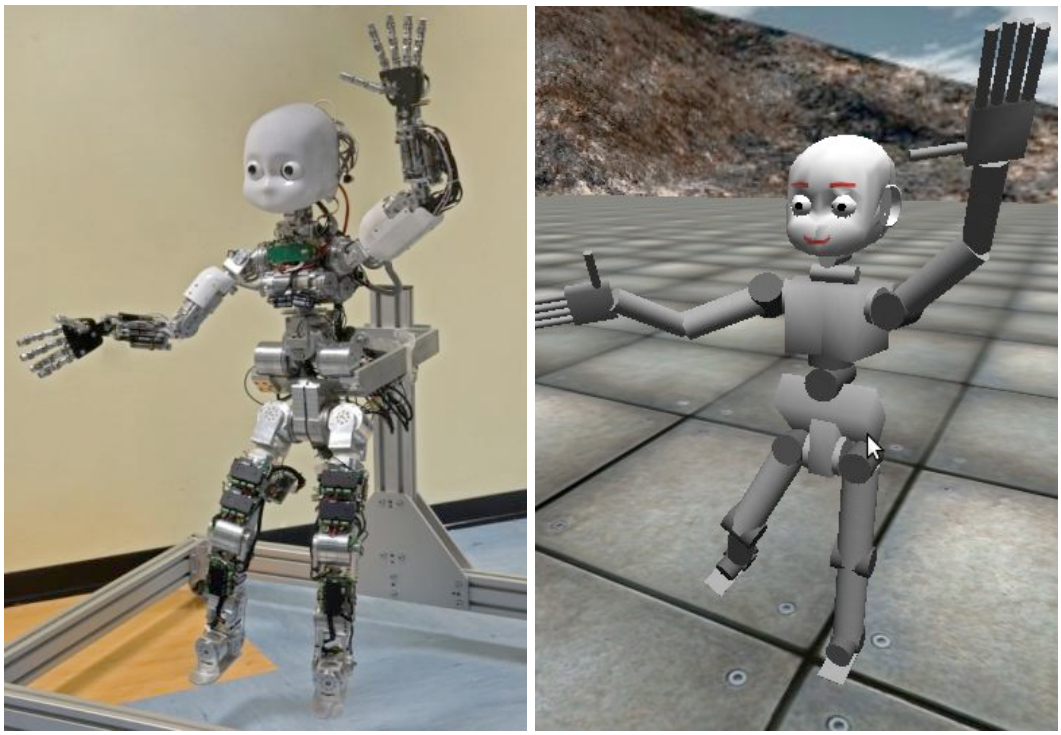

Figure 3: The iCub humanoid robot (left) and it simulation model (right)