frontiers in
NEUROROBOTICS

ORIGINAL RESEARCH ARTICLE
published: 31 May 2010
doi: 10.3389/fnbot.2010.00007

Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot

Davide Marocco[1]*, Angelo Cangelosi[1], Kerstin Fischer[2] and Tony Belpaeme[1]

[1] *Centre for Robotics and Neural Systems, School of Computing and Mathematics, University of Plymouth, Plymouth, UK*
[2] *University of Southern Denmark, Odense, Denmark*

Edited by:
Frederic Kaplan, Ecole Polytechnique Federale De Lausanne, Switzerland

Reviewed by:
Pierre Andry, University of cergy pontoise, France
Peter F. Dominey, Centre National de la Recherche Scientifique, France

****Correspondence:***
Davide Marocco, School of Computing and Mathematics, University of Plymouth, Portland Square A322, Plymouth PL4 8AA, UK.
e-mail: davide.marocco@plymouth.ac.uk

This paper presents a cognitive robotics model for the study of the embodied representation of action words. The present research will present how an iCub humanoid robot can learn the meaning of action words (i.e. words that represent dynamical events that happen in time) by physically interacting with the environment and linking the effects of its own actions with the behavior observed on the objects before and after the action. The control system of the robot is an artificial neural network trained to manipulate an object through a Back-Propagation-Through-Time algorithm. We will show that in the presented model the grounding of action words relies directly to the way in which an agent interacts with the environment and manipulates it.

Keywords: grounding problem, cognitive robotics, embodiment

INTRODUCTION

Human language is a formidable communication system. It allows us to describe the world around us and exchange our thoughts. Nevertheless, despite many decades of studies and research, a complete description of its functions and operations is still missing. In particular, the fundamental mechanisms that allow humans to associate meanings to words are still a matter of ongoing debate among scientists.

For instance, Siskind (2001) suggests three major language functions allowing humans: (i) to describe what they perceive, (ii) to ask others to perform a certain action and (iii) to engage in conversation. At the core of all three functions there is our ability to understand the meanings that words represent. Especially the first two language functions require that language be grounded in perception and action processes. Especially in the description of dynamic processes and specific relations between objects and object properties, the process of grounding language in perception and actions means that, when we describe a given scene or we ask someone to perform a certain action, the words used must be linked with physical entities in the scene or in actions that can be either observed or desired.

In order to understand the link by means of which words are connected with objects and actions, it may be useful to look into studies on child language acquisition.

Children acquire word meanings in direct interaction with the environment. Before they begin to learn words, they go through a long phase of perceptual (visual, haptical, motor, interactional, etc.) exploration of objects in their environment. Interactions with preverbal infants and young children are furthermore anchored in the immediate context; that is, interactions are highly situated in the here and now and allow the child to make direct connections between perceptually available objects and events and linguistic utterances (e.g. Snow, 1977; Hatch, 1983; Sachs, 1983; Karmiloff and Karmiloff-Smith, 2001; Veneziano, 2001). Word meanings are therefore directly related to the child's experience, and the amount of situationally detached information presented to children by their caregivers only gradually increases over time (Veneziano, 2001).

While word meanings are partly acquired based on salient perceptual properties (cf. Clark, 1973; Smith et al. 1996), other word meanings are rather based on the role of functional affordances of objects in interaction (Nelson, 1973; Mandler, 1992). Nelson (1973), for instance, shows that 3-years-old children use their sensorimotor experiences about the function of a given object for categorization. But also linguistic information is taken into account in word meaning learning (cf. Gelman and Heyman, 1999; Bowerman and Choi, 2001; Bowerman, 2005), That is, children understand objects and events that share a linguistic label to share underlying characteristics as well (cf. Gelman, 2009).

But also for adult speakers, word meanings are grounded in embodied experience to a considerable extent (Bergen, 2005; Glenberg, 2007). For instance, distinctions between verbs of grasping are motivated by different hand postures and subtle differences in motor control involved in the actions denoted by a particular motion verb (Bailey et al., 1997). Different motor patterns associated with different action verbs were also found to be reflected in differences in location in the motor cortex (Pulvermüller et al., 2001). Furthermore, language understanding was found to interfere with motor actions if the meaning of the respective sentence evokes a motion in the opposite direction than necessary to carry out the action (Glenberg and Kaschak, 2002).

Further evidence for the embodiment of word meanings in adult language comes from the study of cognitive metaphor and image schemata (e.g. Lakoff and Johnson, 1999) and lexical semantics (Wierzbicka, 1985). These studies draw attention to meanings that are shaped by an implicit understanding of dimensions and functions of the human body. To address word learning from

a grounded language learning perspective is thus supported by research from both child language acquisition and human language understanding.

Several computational models have been proposed to study communication and language in cognitive systems, such as robots and simulated agents (Cangelosi and Parisi, 2002; Lyon et al., 2007). On the one hand there are models of language focusing on the internal characteristics of the individual agent in which the lexicon is constructed based on 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 hand, there are grounded approaches to modeling language, in which linguistic abilities are developed through the direct interaction between the cognitive agents and the physical world they interact with. In these models, the external world plays 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 (Steels, 2003). As pointed out by Cangelosi and Riga (2006), the grounding of language in autonomous cognitive systems requires a direct grounding of the agent's basic lexicon. This assumes the ability to link perceptual (and internal) representations to symbols.

In this modeling paradigm, artificial agents are usually asked to associate features of objects to words, where this association is self-organized by the agents itself. An agent discovers autonomously certain features that are peculiar to a given object and learns from a model, which is usually another agent's, to associate the feature to an arbitrary word. Some of these models aim to study the emergence of shared lexicons through biological and cultural evolution mechanisms (Cangelosi and Parisi, 2002). In these models, a population of cognitive agents that are able to interact with the physical entities in the environment and to construct a sensorimotor representation of it, is initialized to use random languages. Within this population, agents converge toward the use of a shared lexicon after an iterative process of communication and language games.

The paradigm of language games for language evolution and acquisition has been used extensively by Luc Steels (Steels, 2001). For example, Steels and collaborators (Steels et al., 2002; Steels, 2003) 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. In this experiment it has been demonstrated that 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 using the SONY AIBO robot (Steels and Kaplan, 2000). Steels's approach is characterized by his focus on the naming of perceptual categories and by his emphasis on the importance of social mechanisms in the grounding and emergence of language.

Other models focus on the developmental factors that favor the acquisition of language by investigating the role of internal motivation and active exploratory behavior. Oudeyer and Kaplan (2006) show that an intrinsic motivation toward the experience of novel situations (i.e. situations that increase the chance of an agent to learn new environmental and communicational features) lead the agent to autonomously focus the attention toward vocal communicative and language features (see also Oudeyer et al. (2007), on a related topic, and Kaplan et al. (2008) for a compelling review and discussion of computational models of language acquisition).

From a different perspective, Marocco et al. (2003) use evolutionary robotics for the self-organization of simple lexicons in a group of simulated robots. Agents first acquire an ability to manipulate objects (e.g. to touch spheres or to 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 robotics models of language grounding, robotic agents acquire a lexicon through interaction with human users. For example, Roy et al. (2003) have developed an architecture that provides perceptual, procedural and affordance representations for grounding the meaning of words in conversational robots. Sugita and Tani (2005) use a mobile robot that follows human instructions based on 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. Another experiment on human-robot communication has been carried out by Dominey (2005). This particular study provides insight into a developmental and evolutionary transition from idiom-like holophrases to progressively more abstract grammatical constructions.

All of the models presented before adopt the general and widespread assumption that tends to define *nouns* as words associated to physical (or even abstract) entities, and *verbs* as words that represent actions (or, in general, events that happen in time). This practice reflects findings in cognitive (e.g. Langacker, 2008) and functional (e.g. Halliday, 1985) linguistics that nouns are prototypically associated with objects and verbs prototypically correspond to events and actions. Grounded computational models so far mainly focus on grounding nouns on sensorimotor object representations and verbs on actions that are directly performed by the agent (e.g. Sugita and Tani, 2005). In Marocco et al. (2003), for example, a simulated robotic arm was evolved for the ability to discriminate between a sphere and a cube and then to associate different words (nouns) to the two objects. The discrimination was based on a physical exploration of the characteristics of the two shapes. Therefore, the meaning of the nouns was entirely grounded in the sensorimotor dynamic that allowed the discrimination of the two objects. The same procedure has been applied to evolve two different words associated to two different actions performed by the agents. The actions were "avoid" the cube and "touch" the sphere. In this case, the agents were asked to discriminate the objects and perform one of the two actions with respect to the shape. Given the action-word association, these types of words were defined as verbs.

In a different experiment, Cangelosi and Riga (2006) developed a robot able to imitate the actions of another robot (the teacher). The robot was also able to learn from the model an association between actions and words, such as close or open arms. Actions were related to motor patterns performed by the robot and words

were directly associated to those motor patterns. Also in this case, therefore, the grounding of a verb is strictly related to an action that is entirely under the motor control of the agent.

Actions, however, are not restricted to agents. Actions can also be produced by physical objects in the environment, for example. Only few studies focus on the acquisition of actions words that are connected to properties of objects, such as *rolling* for a ball, or on the acquisition of words that express a dynamic and force-varied interaction with objects, such as *hit* or *move*. The application range of these two words can be extremely complex and may vary considerably depending on the physical properties of the object. In the case of the rolling ball, moving the ball can be "similar" to hitting the ball, because the ball has the property to roll after being hit; therefore it will move by itself. On the contrary, hitting a solid cube can produce a different effect from moving the cube by sliding it on the surface of a desk.

Other research in this area has mostly focused on disembodied models that aim to ground the meaning of action words by the elaboration of a visual scene acquired by a fixed camera that observes that scene. In Siskind (2001) the computational model is a computer program called *LEONARD* that analyzes a visual scene and is able to recognize different events, such as *pick-up, put-down, move* or *assemble*. As pointed out before, these actions are not directly performed by an agent, but it is the computational system that observes a visual scene through a fixed camera that is able to reconstruct the meaning. The visual scene typically includes a human hand that perform actions on objects of different colors. For instance, the pick-up scene is represented by a hand that picks up a red cube originally positioned on top of a green cube. The model is based on the principle of force dynamics (Talmy, 1988) and on a specifically designed event logic system, to which in later work (Fern et al., 2002) a learning system has been added. This enabled the computational model to learn and describe events based on a general temporal logic.

A similar approach to the identification of dynamic events has been taken by Steels and Baillie (2003). In their models, two artificial agents embedded in two movable cameras negotiate a self-organized lexicon based on dynamical events observed through the cameras. However, the ability of the agents to recognize and communicate about dynamic events is provided by the interaction between two different ways of using information: A bottom-up and a top-down direction of information flow. The bottom-up system, based on vision, provides information about the actual visual scene and a set of layered software detectors that allow to detect changes in the scene, such as movements of contacts between objects at a lower level, and series of changes at a upper level in order to identify complex dynamics. The top-down system, on the other hand, provides a sort of internal guidance for the vision system that allows, for example, to focus on particular aspects of the visual scene. Thus, the agent's representation of the world is constituted by the interaction between the two processes, encoded in a kind of lisp-type logic, and by sharing its communicative interaction with the other agent.

Cannon and Cohen (2010) and Cohen et al. (2005) ground the meaning of action words on the physical interactions between two bodies. In their system, verbs like *push, hit* and *chase* are represented as pathways through a metric space, defined as *maps for verbs*, that represent distances between two interacting physical entities, their velocity, and the observed transfer of energy after the interaction. Bodies are represented as circles of different colors that interact in various ways. Also in this case, as before, the system is purely based on the passive elaboration of a visual scene.

Following the same path, the aim of the present research is to study how a humanoid robot can learn to understand the meaning of action words (i.e. words that represent dynamical events that happen in time) by physically acting on the environment and linking the effects of its own actions with the behavior observed on the objects before and after the action. This will allow the agent to give an interpretation of a given scene that develops in time, and is grounded on its own bodily actions and sensorimotor coordination. Object manipulation, therefore, is the central concept behind the research. We believe that an active manipulation of the object is an opportunity to test the reaction of that object. Imagine the robot hits a ball. As an effect of the hit, the ball will move. Therefore, the dynamics of this event can be characterized by the action performed by the robot and by the sequence of the activation of its sensors during the movement and the physical interaction with the ball. The movement of the ball can be viewed as an instantaneous contact between the ball and the hand of the robot followed by a displacement of the same ball in the space, away from the hand. In such a case, the integration of the contact sensors with vision information can easily characterize this situation as different from another situation in which, e.g., the robot moves a cube by sliding it over the surface of the desk. In this case, although there is movement, i.e. the object displaces in space, the event is characterized by a continuous contact of the hand with the cube. On the other hand, the fact that different objects react differently to the same movement can also characterize a particular property of the object itself. Therefore, *rolling* and *sliding* are action words that pertain to objects that can be understood by the agent on the basis of the same sensorimotor information used to characterize its own actions.

Such types of interactions can be easily regarded as affordances of the objects for the robot. In fact, the robot learns the effects of its own movement on a given object. Several studies have already addressed affordances on robotics models in a similar way, where a robot learns a specific type of affordances using information provided by sensory states. These models have been mainly used in relation with imitation tasks. For example, in Fitzpatrick et al. (2003) a robot learns the motion dynamics of different objects after having pushed them. Subsequently, it uses the sensorimotor information to recognize actions performed by others and to replicate the observed motion. Similarly, Kozima et al. (2002) created a system that enables a robot to imitate actions driven by the effects of that actions (a more general solution on learning affordances is presented in Stoytchev, 2005; Fritz et al., 2006; Dogar et al., 2007). Montesano et al. (2008) created a humanoid robot controller that uses a Bayesian network for learning object affordances and showed the benefit of the model in imitation games. The model presented here, although inspired by a similar approach, does not have an explicit interest in imitation, and also the actions repertoire is simplified in comparison to those models. However, we believe that this simplification helps to better highlight and understand the sensorimotor grounding of action words, which is the primarily scientific question behind this work. This consideration, of course,

does not prevent possible extensions of the model towards more applied scenarios that involve imitation tasks, as well as tasks that involve a form of linguistic instruction provided by another agent, which might be a human or another robot.

To approach the research issue related to the grounding of action words in sensorimotor coordination, we present a simulated robotic model equipped with a neural control system. By manipulating the environment, the robot can learn the association between certain objects, located on a desk in front of him, and some physical property of such objects. In the next section, a description of the robot used in the experiment, the environment and the neural control system will be described. Subsequently, the results of the experiment will be present and discussed.

MATERIALS AND METHODS

The robotic model used for the experiments is a simulation of the iCub humanoid robot (Tikhanoff et al., 2008) controlled by a recurrent artificial neural network. The robot can interact with objects located on a desk in front of it, and its neural control system is trained through a supervised learning algorithm, namely the "Back-Propagation-Through-Time" algorithm (Rumelhart and McClelland, 1986). In the following sections we provide details on the robotic platform utilized, the environment and on the robot-object interaction. Moreover, a description of the neural network that acts as a control system and of the training procedure will be presented.

THE SIMULATED HUMANOID ROBOTS AND THE ENVIRONMENT

For the experiments a simulated model of the iCub (**Figure 1**) has been used, a small-size humanoid robot, designed and produced by the European project "Robotcub" (robotcub.org; Metta et al. 2008). The iCub dimensions are similar to that of 2.5-year-old child and

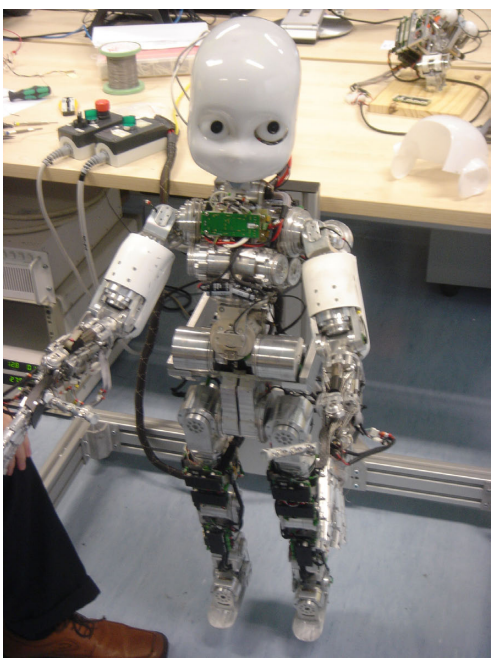

FIGURE 1 | The humanoid robot iCub.

the robot has been specifically designed to act in a cognitive robotics domain, where the robotic platform is a physical entity that allows researchers to test hypothetic cognitive models in the real world. The robot is 90-cm tall and has a weight of 23 kg. iCub has 53 degrees of freedom distributed as follow: seven for each arm, six for each leg, three on the waist, three dedicated to eyes movements and three for the neck. In addition, it has two complex hands with 9 degrees of freedom each. For its size, the iCub is the most complete humanoid robot currently being designed, in terms of kinematic complexity. In contrast to similarly sized humanoid platforms, the eyes can also move. All motors and sensors are accessible through a centralized control system that provides an interface between the robot the and the external world. The interface is implemented on a PC104 board located in the head of the robot. For vision, the robot is equipped with two cameras with VGA resolution and 30-fps speed that provide color images (for additional technical details about the robot body and head see Beira et al., 2006; Tsagarakis et al., 2007). Every communication with the robots uses an Ethernet network protocol. The integrated software platform to control all the sensors and actuators is called YARP (Metta et al., 2006).

In our experiment we used a carefully designed software simulation of the iCub robot that uses ODE (Open Dynamic Engine) to simulate the dynamics of the physical interactions (for details about the simulator see Tikhanoff et al., 2008). The YARP platform is used as the main communication tool for both the simulator and the real robot. The simulator has been designed to test the robot's software application in a safe, yet realistic, environment. In particular, the simulator can be used to safely test potentially dangerous motor commands that might damage the physical structure of the robot. Moreover, for the specific requirement of the model, we had to use tactile sensors in the hand that are currently not implemented on the real robot available to us.

For the present study we used a sub-set of all the degrees of freedom and only one of the two cameras. In particular, for manipulation purpose we only use a single joint on the shoulder that allows the robot to reach and move an object placed on a desk in front of it. The encoder value of this joint is also used as proprioceptive sensory feedback. When the hand gets in contact with an object, a binary tactile sensor placed on the hand is activated. Its activation value provides a coarse tactile sensory feedback. This tactile sensor is activated whatever part of the hand gets in contact with the object. The vision of the robot is provided by a vision system that acts on the left camera of the robot that automatically fixate the object in the environment, regardless of the action currently performed by the robot. The encoder values of two neck joints, that represent the position of the head, express the position of the object in the visual field relative to the robot. The position of the head is then treated as visual input of the system (Yamashita and Tani, 2008). The vision system, in addition to the object relative position in the visual field, also provides coarse information about the shape of the object. A parameter of the shape, which we call *roundness*, is calculated from the image of the object acquired by the robot and its value is added as input to the neural network controller. The robot automatically generates a movement when it receives a target joint angle as input. The movement corresponds to the target angle and is generated

by means a pre-programmed proportional-integral-derivative (PID) controller. The sensorimotor state of the robot is updated every 500 ms.

The environment of the experiment consists in a desk placed in front of the robot. On the desk, one out of three objects is positioned on a given location. These objects are a sphere, a box, and a cylinder placed vertically on the desk. The sphere has a diameter of 12 cm. The three dimensions of the cube are 12 cm on a side and 7 cm on the other two sides. The cylinder has a diameter of 4 cm and 25 cm tall. Roundness values calculated for the three object are ~0.87, ~0.71, and ~0.43 for the sphere, the cube, and the cylinder respectively. Each of these objects has different physical properties associated to the shape and the physical connection to the desk. The sphere, when touched by the robot hand, will roll away on a direction that directly depends on the hand direction and on the applied force. The cube, when touched with the same force and direction as the ball, will slide on the desk while in contact with the robot hand. The cylinder, which was tightly attached to the desk, will not move and will prevent the robot to accomplish its desired movement. Therefore, the three objects represent three different properties, namely, the property to roll, to slide and to resist.

STRUCTURE AND TRAINING OF THE NEURAL CONTROL SYSTEM

The neural system that controls the robot is a fully connected recurrent neural network with 10 hidden units (**Figure 2**), eight input units and eight output units. Activations of input units are divided into five sensory units and three linguistic units. Three of the five sensory units are set to the encoder values of the three corresponding joints (shoulder, pan-neck and tilt-neck), scaled between 0 and 1. Those input units provide information about joints current angles. The fourth sensory unit encodes the value of the binary tactile sensor. This is set to either 0 or 1, depending on the contact of the hand with the object. The fifth sensory unit encodes the value of the roundness. The three linguistic input units represent a local binary encoding of the three objects. The activation value of those units can vary with respect to the experimental phase, that is, training or testing phase, respectively. Activations of hidden and output units y_i are calculated at a discrete time, by passing the net input u_i to the logistic function, as it is described in Eqs 1 and 2:

$$u_i = \sum_j^n y_j \cdot w_{ij} - k_i \tag{1}$$

$$y_i = \frac{1}{1 - e^{-u_i}} \tag{2}$$

where w_{ij} is the synaptic weight that connects unit i with unit j and k_i is the bias of unit i. The output units encode the values of the input at the time step $t + 1$. That is, the output state corresponds to the next input state of the network. The network is trained to predict its own input. As we will see in the next section, during the testing phase, the predicted input state is also used to provide target angles for the actuators.

The structure of the experiment is divided into two phases: in the first phase the network is trained to predict its own subsequent sensorimotor state. In the second phase the network is tested on the robot, in interaction with the environment.

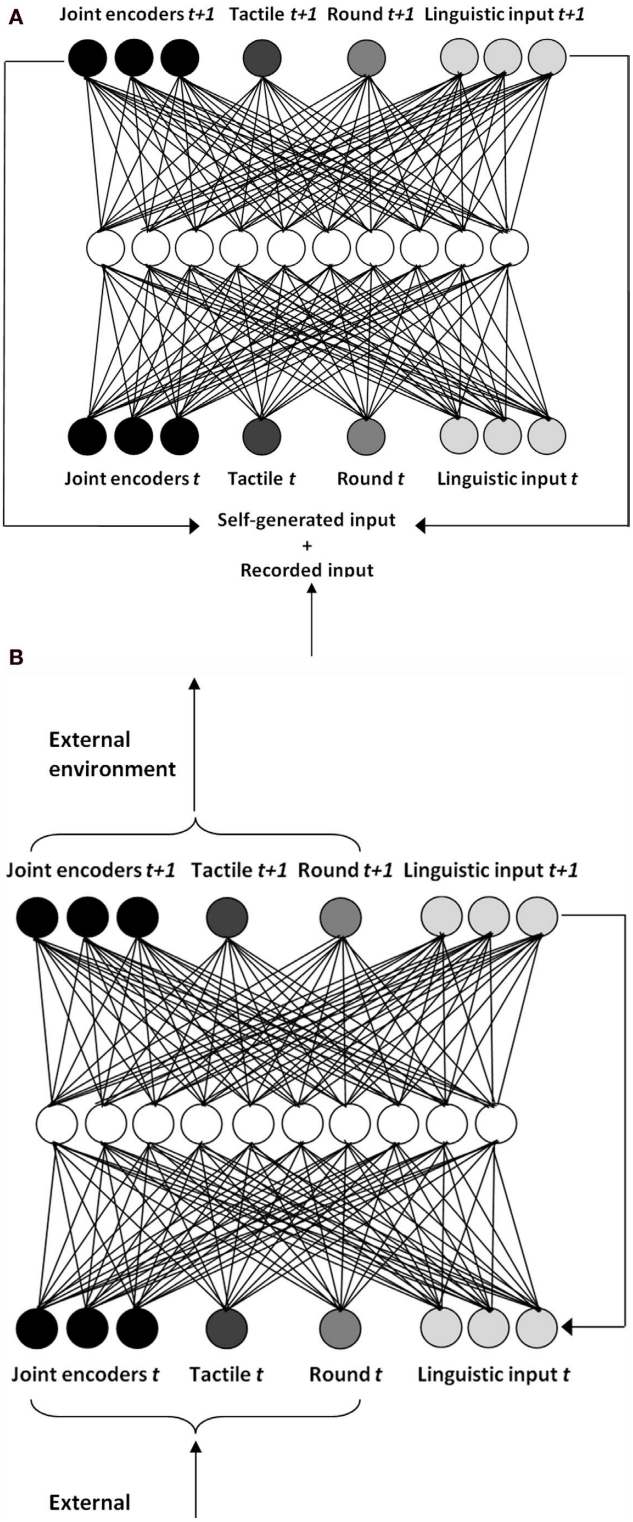

FIGURE 2 | The neural network that acts as a control system for the robot. **(A)** Network activation structure during the training phase (*closed-loop* condition). The output at time t with a small portion of the recorded input is used for setting the input at time $t + 1$. See the text for details. **(B)** Network activation structure during the testing phase (*open-loop* condition). The input is taken by the state of the sensors and the output is used to set the target angle of the actuators.

Training phase

For training the neural network we used the Back-Propagation-Through-Time-algorithm (BPTT), which is typically used to train neural network with recurrent nodes (Rumelhart and McClelland, 1986). This algorithm allows a neural network to learn the dynamical sequences of input-output patterns as they develop in time. Since we are interested in the dynamic and time dependent processes of the robot-object interaction, an algorithm that allows to take into account dynamic events is more suitable than the standard Back-Propagation algorithm (Rumelhart and McClelland, 1986). For a detailed description of the BPTT algorithm, in addition to Rumelhart and McClelland (1986) see also Werbos (1990). The main difference between a standard Back-Propagation algorithm and the BPTT is that, in the latter case the training set consists in a series of input-output sequences, rather than in a single input-output pattern. The BPTT allows the robot to learn sequences of actions. The goal of the learning process is to find optimal values of synaptic weights that minimize the error E, defined as the error between the teaching sequences and the output sequences produced by the network. The error function E is calculated as follows:

$$E = \sum_{S} \sum_{t} \sum_{i \in \text{output}} ((y_{i,t,s*} - y_{i,t,s})(y_{i,t,s} - (1 - y_{i,t,s})))^2 \quad (3)$$

where $y_{i,t,s*}$ is the desired activation value of the output unit i at time t for the sequence s and $y_{i,t,s}$ is the actual activation of the same unit produced by the neural network, calculated using Eqs 1 and 2.

During the training phase, synaptic weights at learning step $n+1$ are updated using the error δ_i (Rumelhart and McClelland, 1986) calculated at the previous learning step (n), that in turn depend on the error E, according to the following equation:

$$\Delta w_{ij}(n+1) = \eta \delta_i y_j + \alpha \Delta w_{ij}(n) \quad (4)$$

where w_{ij} is the synaptic weight that connects unit i with unit j, y_i is the activation of unit j, η is the learning rate and α is the momentum.

The sequences to learn, in our case, are the sensorimotor contingencies produced by the robot's manipulation of the object present in the environment. In order to produce those sequences, the robot is placed in front of the desk together with one of the three objects placed on the desk, at a given position. At this point, the shoulder joint of the robot is activated so as to move the right arm from the side of the robot to the front. By performing this movement, the hand of the robot moves towards the object and gets in contact with it. At the same time, the automatic vision routing turns the head in the direction of the object and keeps the object in the visual field by moving the neck joints (**Figure 3**). During this activity, we recorded the values of shoulder and neck joint encoders, as well as the state of the tactile sensors and the *roundness* value calculated by the image processing system. Each sequence consists of 30 recorded patterns that represent 15 s of activity by the robot. The graphs in **Figure 4** show the activations of the sensory units when the robot is interacting with the three objects. The information provided to the robot, although extremely simple, is sufficient to allow the neural controller to correctly separate the three conditions.

The input pattern of every sequence is completed by adding the linguistic input in the following form: *[1 0 0]* when the robot is interacting with the sphere, *[0 1 0]* when the robot is interacting with the cube, and *[0 0 1]* when the robot is interacting with the cylinder. The linguistic input is explicitly presented only at the beginning of the sequence. For the rest of the 30 patterns that form the training sequence, the linguistic input is self-generated by the network. It should be noted that at this time we deliberately avoid to give a semantic interpretation of the linguistic input and output. So far the "words" chosen as input and, consequently, as output simply correlate with the interaction with different objects.

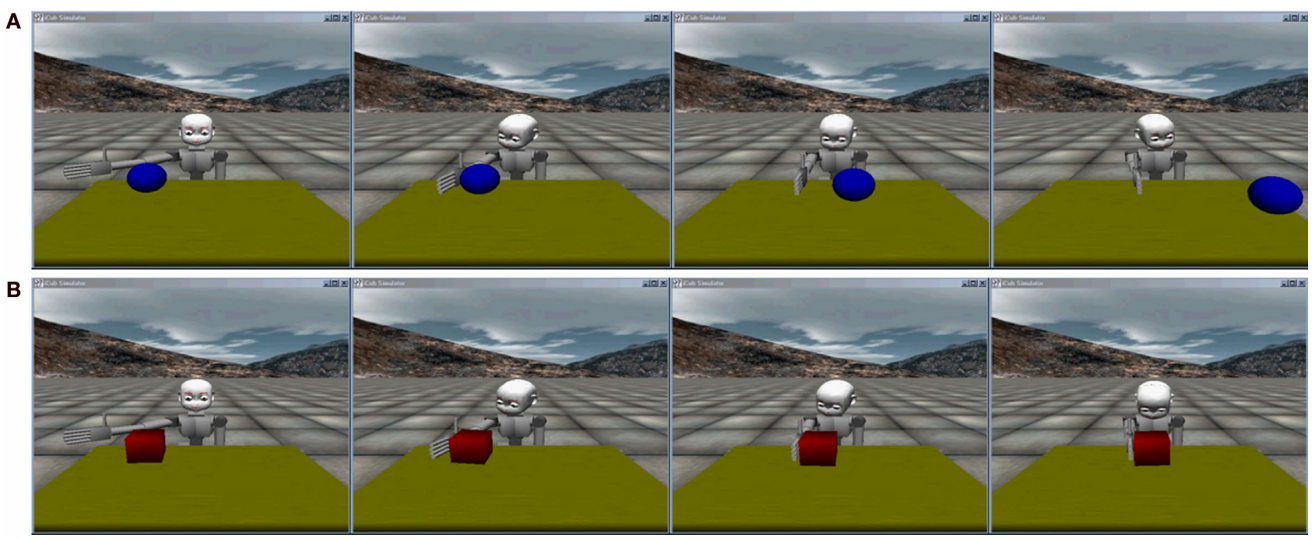

FIGURE 3 | From left to right, a small sample of the 30 step sequences produced for training the network. (A) An example sequence produced by the manipulation of a sphere and **(B)** the same movement towards a cube. The two objects produce different interactions because of their different physical properties. The sequence produced by the interaction with the fixed cylinder is not shown, given the fact that the robot, after the contact with the cylinder does not move anymore. The tracking behavior is due to the automatic visual routine embedded in the control system.

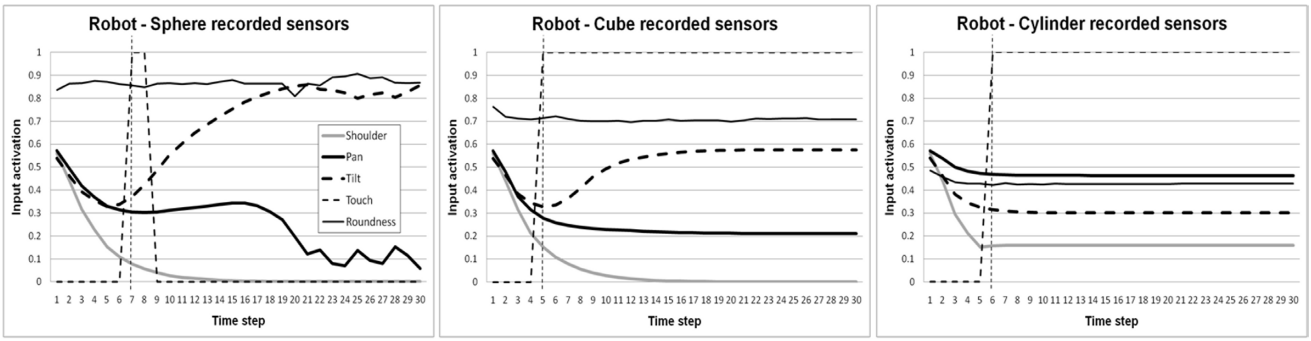

FIGURE 4 | Sensor activations recorded during the 15-s interactions between the robot and the object on the desk. Curves represent activations of the *shoulder* joint, the *pan* and *tilt* movements of the neck, the *touch* sensors and the *roundness* of the object calculated by the vision system over time. Refer to the legend on the graph. The thin dashed vertical line in the three graphs represents the moment in which the robot's arm touches the object.

Their semantic referent, i.e., whether they refer to objects (sphere, cube, cylinder) or to actions associated with the objects (roll, slide, fix), will be discussed later on. For this reason we refer to the linguistic output of the neural network as *linguistic_output_1 [1 0 0]*, *linguistic_output_2 [0 1 0]* and *linguistic_output_3 [0 0 1]*, corresponding to interactions with the sphere, the cube and the cylinder, respectively.

From these values we produce the sequences, one for each object, by setting the sequence element $t + 1$ as target output for the previous element t. In this way, starting from the first pattern, the network has to produce the next pattern. Then, the produced pattern is given as input to the network, which produces the next pattern and so on. This iteration is executed until the end of the sequence is reached.

The complete training set for the present work includes six sequences. Three of these are created in the way just described above, while the other three use the same set of data as before except for the roundness values which is set to 0. The linguistic input is presents in both cases.

During this process, the error produced by the network with respect to the target outputs is accumulated for all three sequences. The synaptic weights are updated according to the global error only after all the sequences have been presented. Therefore, according to the traditional back-propagation notation, the neural network is trained in "batch" mode and not "on-line" (for technical details about the algorithm adopted in this paper and a discussion about computational differences between "batch" and "on-line" training mode in recurrent neural networks see Williams and Zipser, 1995).

To facilitate the training process and to produce a neural network capable of better predicting the sequences, we used a training modality known as *closed-loop training* (Yamashita and Tani, 2008) depicted in **Figure 2A**. In this type of training procedure the input given to the network is the actual output produced by the network itself at the previous time cycle. However, by doing this, the error can accumulate on the input, especially at the beginning of the training process, given that the input is self-generated by a network with random synaptic weights. For this reason, the effect on the learning performance can be heavily affected and can prevent the algorithm to converge to a close to optimal solution. To avoid such a problem, the real input s fed to the network (i.e. the input actually used to calculate the performance), is produced by adding to the self-generated input s+ a small fraction of the recorded input s*, which represents the real input the network should receive. The same is done for the linguistic input m, with the only difference that m* is the linguistic activation fixed by the experimenter:

$$s = 0.1 \cdot s^* + 0.9 \cdot s^+$$

$$m = 0.1 \cdot m^* + 0.9 \cdot m^+$$

The parameters used for training the network used in the following experiments are: Learning rate 0.2; momentum 0.3; initial synaptic weights value between −0.01 and 0.01.

To assure the robustness of the results obtained, 10 replications with different initial random synaptic weights were carried out.

Testing phase

The second part of the experiment is the phase in which the network is tested in *open-loop*. That is, the network is connected to the robot and the input is directly produced by the actual values of its encoders (**Figure 2B**), while the output is used to determine the target angles of the joint for moving the arm. During this phase the robot is placed in front of the desk as before and an object is placed on the desk in front of the robot. By activating the robot, this time the movement of the arm is commanded by the output of the neural network, while the other outputs represent the prediction of the next sensory state.

In this set-up, the only joint that can be directly actuated by the network is the joint on the shoulder, which causes the movement of the arm toward the object. The joints controlling the neck are still commanded by the visual routine that tracks the object in the environment. However, the most interesting part in this experiment is the behavior of the linguistic output. As we will show and discuss in the next session, the interaction between action and language exploited during the training, allows us to better understand what type of sensorimotor contingencies are associated with certain linguistic patterns and how certain categories of action words might be directly grounded in the sensorimotor states of an agent.

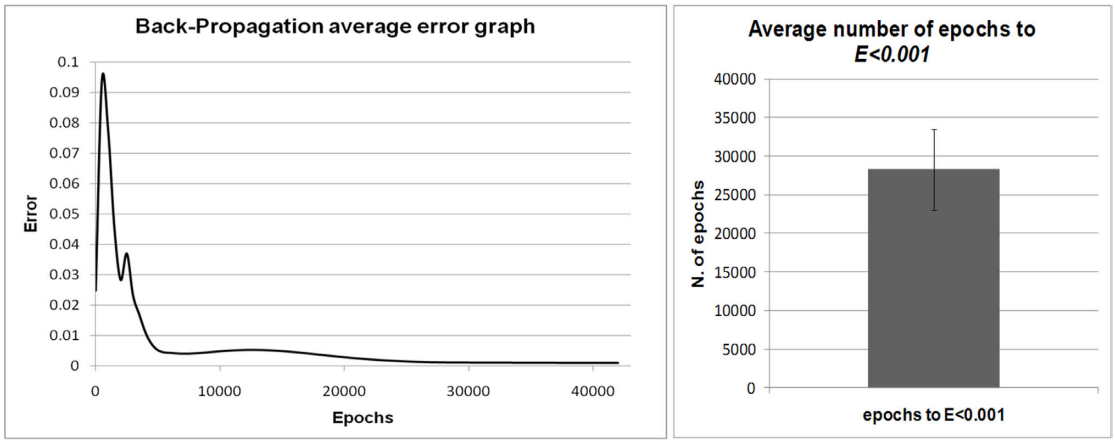

FIGURE 5 | Left: Error graph of the average of 10 replications during the training process. The *x* axis shows the number of epochs and the *y* axis shows the mean error. Right: Average epochs required to reach the error threshold, which was set to 0.001 ($E < 0.001$). Error bar represents Standard deviation.

During the testing phase some parameters of the set-up used for the training have been changed, such as the presence of the linguistic input and the size and position of objects on the desk. In the next section we will describe the tests performed and the results obtained.

RESULTS

After the training of all the 10 neural networks, we obtained 10 controllers that were able to predict the next sensory input state on the basis of the current input state. Previous tests shown that an error E smaller that 0.001 produces neural controllers capable of performing the task with a good degree of generalization. To avoid the overtraining of the network, we decided to set the learning threshold to 0.001. Below this threshold the training process is considered completed. Given this threshold, the 10 replications have been carried out by stopping the training as soon as E was smaller than the threshold. **Figure 5**, left, shows the average error calculated for the 10 replications during the training process. **Figure 5**, right, shows the average epochs that occurred until an error smaller than 0.001 for the 10 replications was reached. From the integration of the two data sets we can see that after about 28.000 epochs the error is already smaller than 0.001 for the majority of the replications, while for some of them additional epochs are required. The fastest replication reached the error threshold in 22695 epochs and the slowest reached the threshold in 42254 epochs.

The trained neural controllers were tested systematically using the *open-loop* procedure connecting the controller with the simulated robot. A test of the robot under the same conditions as experienced during the training process (i.e. with linguistic input and with/without roundness information) showed that the neural controllers were perfectly trained and were able to move the robot and emit the correct outputs in all training conditions (see condition *Ling* in **Figure 6** for a similar test performed with roundness information. Results without roundness are not shown since the networks were able to reproduce the recorded sequence with a negligible error). After this first test, more comprehensive generalization tests were performed in order to estimate the capability of the controller to cope with different conditions.

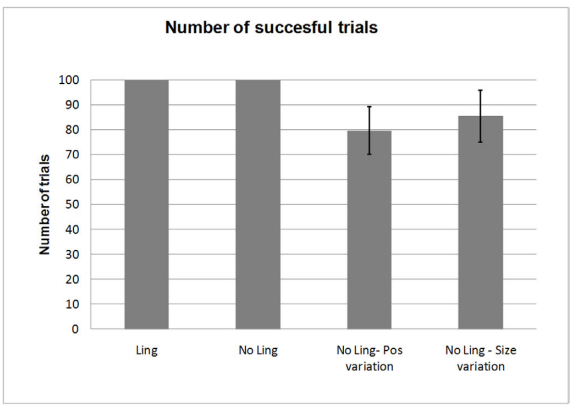

FIGURE 6 | Results of 100 trials from four different testing conditions. Standard deviation is represented as error bar. See text for details.

GENERALIZATION UNDER DIFFERENT CONDITIONS

For this analysis each of the 10 controllers were tested under four different conditions for 100 trials. The number of successful trials was recorded, i.e. the cases in which, at the end of 30 sensorimotor cycles of the neural network controlling the robot, the activation of the output units were the same as the desired output. Given the variation of the initial condition, a deviation of ±0.1 was allowed for every output unit. It should be noted that, given a certain degree of the error, the robot is not able to accomplish the task at all. The four testing conditions are as follows:

Ling. A condition identical to that of the training process, with the linguistic input provided at the beginning of the trial.

No Ling – In this condition the linguistic input is set to zero during the whole duration of the trial. The other parameters are the same of the training process.

No Ling – Pos variation. A condition in which the linguistic input is set to zero during the whole duration of the trial. In addition, at every trial the position of the object randomly varies within a range of ±10 cm.

No Ling – Size variation. A condition in which the linguistic input is set to zero during the whole duration of the trial. In addition, at every trial the global size of the object is randomly scaled within a range of ±20%. For instance, the diameter of the sphere can vary at each trial between a minimum of 9.6 cm and maximum of 14.4 cm.

Information about the roundness was available in all conditions.

Results of the tests are depicted in **Figure 6**. The conditions *Ling* and *No Ling* interestingly are exactly the same for all 10 replications. It should be noted that the neural networks have been trained always with linguistic input. This means that the natural generalization capability of the network is able the reconstruct the input pattern, including the linguistic input, without any loss in terms of performance. In the *No Ling* condition, the robot is placed in front of the object and performs its movement toward it without any linguistic input; still, it is able to produce the correct linguistic pattern after the interaction with the object. This result indicates that the controller can recall and produce the appropriate linguistic output only on the basis of its overall sensory state. *Pos variation* produces slightly worse results and we observe a certain variation among the replications, as indicated by the standard deviation on the graph. It is interesting to note that the worst replication is the one which took more epochs to converge, whilst the best is the one that converged in the fewer number of epochs. Besides the performance decrement, the majority of the replications shows a very high generalization capability, even though in the allowed range of the variation. The same can be observed for the *Size variation* condition, although the results appear slightly better. This can be explained by the fact that the roundness information is, to a certain extent, independent from size variations. Therefore, roundness provides a reliable source of information even in cases of unexpected sensory-motor input in comparison with that experienced during training. This effect has been observed in connection with larger objects.

The tests presented above demonstrate that the neural controller is capable to produce the correct behavior in terms of joint activations and prediction of sensorimotor states, as well as in terms of linguistic activations. Moreover, the correct behavior is performed also without providing a linguistic input. This is also true when the set-up is manipulated to a certain extent.

Nevertheless, this kind of test does not allow us to understand what exactly the information is that is used by the controller to connect an object with its corresponding linguistic label. In order to clarify this issue, additional tests have been performed.

UNDERSTANDING THE MEANING OF WORDS

Further tests and analyzes were carried out in order to better understand the meaning associated to the linguistic labels, which we can imagine as a kind of simplified words, and the relation between the sensorimotor processes triggered by robot-object interactions and these arbitrarily provided words. To analyse the word–meaning mappings that emerge from the current experimental set-up, the dynamics of the activations of the linguistic units and the roundness prediction have been analysed under several conditions in which additional input modifications are explored.

As we stated above, so far we cannot properly link an observed linguistic activation with a particular word. In fact, we still ignore the specific relation between words and meanings created by the controller. Therefore, in the following tests and analyzes we will refer to the linguistic output in very general terms. We will apply a specific word associated to the linguistic output only when the relation between them and their referents will be clarified. The notation for linguistic output identification, already introduced in Section "Training phase", is the following:

linguistic_output_1: correlates with the *robot-sphere* interaction;
linguistic_output_2: correlates with the *robot-cube* interaction;
linguistic_output_3, correlates with the *robot-cylinder* interaction.

Given the differences among the neural controllers in terms of synaptic weights and overall dynamics, the following additional analyzes were carried out using a single controller. The controller of replication 2 was chosen because it demonstrated to be the best one in the previous generalisation tests.

Tests on linguistic outputs

In this test, the robot was placed in front of an object without providing any linguistic input yet with roundness information. During the interaction of the robot with the object, the activations of the linguistic output have been recorded for the usual 30 sensorimotor cycles allowed (15 s). As it is shown in **Figure 7** (left column), linguistic activations vary greatly as the interaction unfolds in time, depending on the object. In the case of the sphere, the roundness provides information that immediately permits *linguistic_output_1* activation, as correctly required by the task. However, after about 3 s, the hand of the robot gets in contact with the object, and for a while the linguistic output changes by activating *linguistic_output_2*, which correlates with robot interacting with the cube. However, as the interaction continues and the sphere rolls away, the robot is then able to produce again the right output pattern after some time. In the graph we can also observe the time dynamics of the roundness prediction, which is affected, like the linguistic output, by the overall sensorimotor state of the robot. The perceived roundness for the sphere is about 0.9, which is correctly predicted by the robot at the beginning. Nevertheless, after the activation of the touch sensor, the prediction switches from "sphere" (~0.87) to "cube" (~0.71), although in input the robot still receive the correct roundness. Finally, the roundness measure returns to the right value after the sphere begins to roll away from the hand.

The dynamics that we observe while the robot is interacting with the cube and the cylinder is very similar. Roundness information, in fact, allows an early recognition of the type of object and the production of the correct linguistic output pattern, i.e., *linguistic_output_2* for the robot-cube interaction and *linguistic_output_3* for the robot-cylinder interaction. In case of the cube, after the contact of the robot with the object, the correct *linguistic_output_2* is triggered and it remains active for the rest of the time. Only a minor activation of the *linguistic_output_3*, related to the cylinder, is observed just after the contact. As for the cylinder, the correct output is emitted by the star and after a brief interference of the *linguistic_output_1*, the correct output, i.e. *linguistic_output_3*, is activated and maintained.

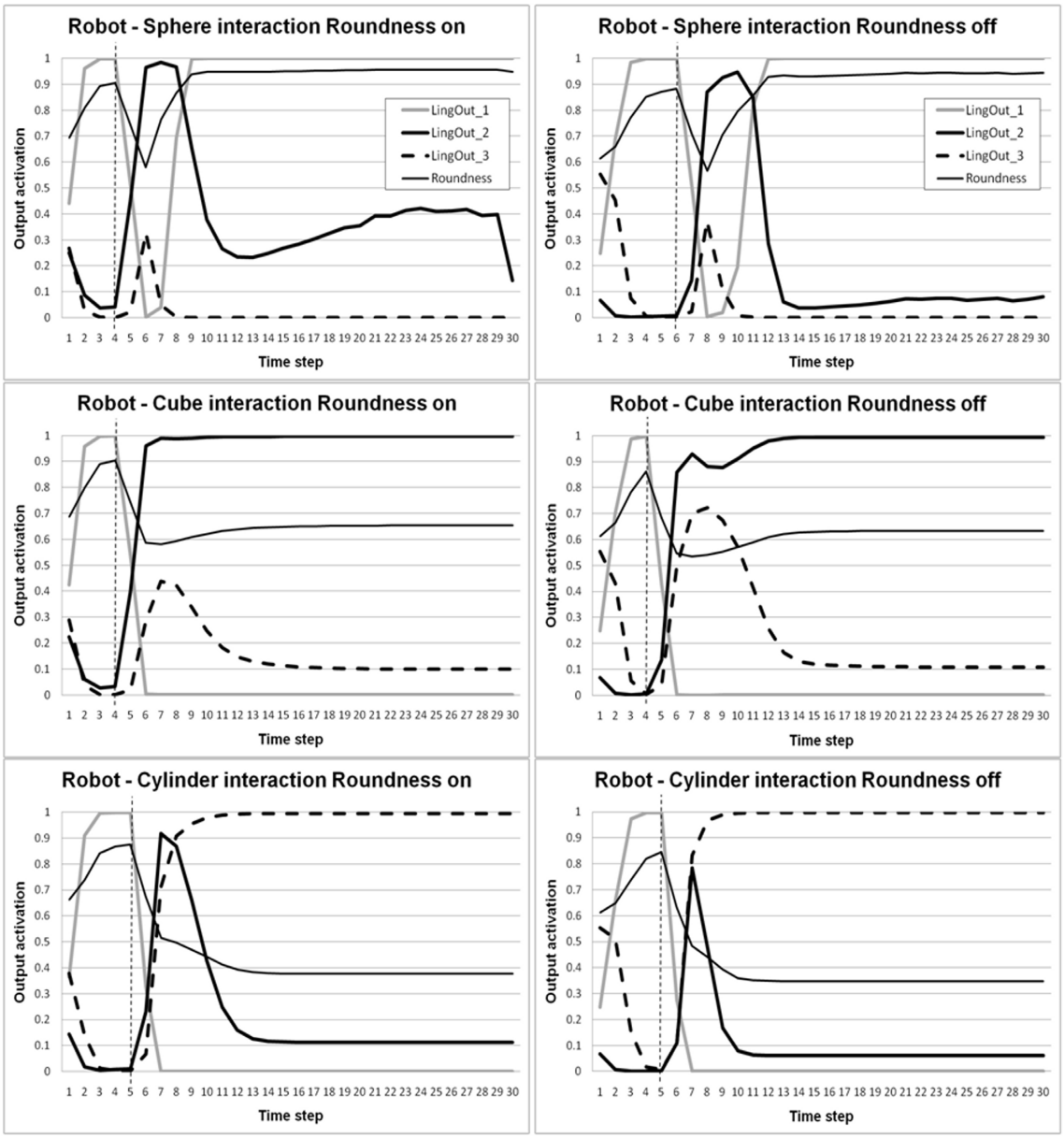

FIGURE 7 | Activations of linguistic output units during the 15-s interaction between the robot and the object on the desk. The graph shows *linguistic_output_1* (interaction with the sphere), *linguistic_output_2* (interaction with the cube), *linguistic_output_3* (interaction with the fixed cylinder), and the roundness prediction. Refer to the legend on the graph. The thin dashed vertical line in the three graphs represents the moment in which the robot's arm touches the object. Left: Condition with roundness information in the input. Right: condition without roundness information in the input.

Figure 7 (right column) shows the same type of analysis as above for a condition in which neither linguistic input nor roundness information are available. This case is much more complex than before because at the beginning of the movement no external information is available. Not surprisingly, the dynamics is also different. When the robot is interacting with the sphere, a presumably stereotypic behavior of the chosen neural controller produces, at the beginning of the movement and without any other information about the object available, a default linguistic output activation, which is the correct one by chance, that is, *linguistic_output_1*. After the robot touches the sphere, similar to what we observed above, the correct output is suppressed and the one corresponding to the cube is activated. During the following interaction, *linguistic_output_2* is active to a smaller extent than in the case with roundness in the input (see corresponding graph on the left column). This effect is probably due to a kind of interaction between the roundness value provided in the input and the linguistic output activation. Roundness values for "sphere" and "cube" are indeed very close. When the robot is interacting with the cube, given the stereotypic behavior of the controller at the beginning, it activates *linguistic_output_1*. However, after the contact the cube slides on the surface and the correct output is produced. For the cylinder,

the same interference as before between *linguistic_output_2* and *linguistic_output_3* is observed, but after few seconds the robot produces the correct output.

Generalization tests on linguistic outputs

Results presented so far clearly show that the linguistic input is tightly connected with the sensorimotor dynamics produced by the interaction with the object. The tests demonstrate the ability of the robot to correctly categorize the objects, also in the absence of direct linguist input, and to produce the corresponding linguistic label only on the bases of its sensorimotor state. From this point of view, the observed interaction between the flow of the sensorimotor states and the activations of the linguistic units leads to the hypothesis that the whole sensorimotor state, rather than a single elements such as, e.g., the roundness, is at the core of the controller's ability to categorize the events correctly. In this section, therefore, we performed an additional test to verify this hypothesis and to investigate what the real meaning is on which the linguistic labels are based.

The test consists of three different conditions in which the robot was tested. Again, no linguistic input is provided. The three conditions are the following:

(a) A cylinder very similar to the one used throughout the training process is placed in front of the robot. This time the cylinder is not attached to the table and is free to move. Its starting orientation is parallel to the starting position of the robot arm. Thus, it can roll away when the robot touches it (**Figure 8A** right). The roundness perceived by the robot is the same as for the cylinder.
(b) The same cylinder is placed on the table and free to move. The starting orientation is perpendicular to the robot arm. That is, it is rotated 90° with respect to the previous condition. In this position it can easily slide but not roll (**Figure 8B** right).
(c) A cube is fixed to the table. The perceived roundness is the same of the cube, as during the training, but the cube cannot move if touched (**Figure 8C** right).

Results are shown in **Figure 8**. **Figure 8A** (left) represents the interaction with rolling cylinder, showing that when the cylinder is touched, it starts to roll away. Given the specific sensorimotor dynamics produced by the cylinder, the related pattern dynamics of the linguistic output are very similar to that already seen for the robot-sphere interaction. We may thus conclude that the robot categorizes and labels the rolling cylinder as it categorizes the sphere by activating *linguistic_input_1*. Similarly, the interaction with the sliding cylinder (**Figure 8B** left), given the fact that it slides and produces the same sensorimotor patterns previously seen for the cube, induces the controller to activate *linguistic_input_2*. Finally, **Figure 8C** (left) depicts the linguistic output activations while the robot interacts with the fixed cube. Even though its dimension and perceived roundness are exactly the same as for the cube used during training, the sensorimotor contingency produced by the interaction is identical to that experienced with the cylinder in the training. Not surprisingly, the linguistic activation is the same observed for the cylinder in the previous test: *linguistic_input_3*.

These additional results suggest that the linguistic label are grounded in complex sensorimotor dynamics instead of in the visual features provided by the roundness parameter, despite the fact that roundness information is provided. Specifically, the grounding of the linguistic output can be identified with the dynamics associated to the physical properties of an object.

DISCUSSION AND CONCLUSION

What has been shown so far indicates that the robot is able to extract the sensorimotor contingency of a particular interaction with an object and to reproduce its dynamics by acting on the environment. Moreover, in the absence of linguistic input, the robot is capable of associating a certain temporal sensorimotor dynamics to the learnt linguistic labels. Thus, in the lights of the results provided by the tests, it is now time to ask whether the linguistic label learnt are associated to the objects themselves or whether the label refers to the physical properties of the objects.

The results presented in Section "Tests on linguistic outputs" on the activation of the linguistic output during robot-object interaction, clearly show that the robot is able to correctly categorize and produce the correct label for a given object both in absence of the corresponding linguistic input and roundness information. Moreover, the generalization tests presented in Section "Generalization tests on linguistic outputs" indicate that, irrespective of the roundness information provided and in absence of linguistic input, the linguistic output correlates with the sensorimotor dynamics produced by the specific physical property of the object. Therefore, such results suggest that the linguistic labels are based on an entire sensorimotor dynamics, and not on the visual features provided by the roundness parameter. Specifically, the grounding is exactly the dynamics associated to that physical property.

This interpretation is corroborated by other works that connect active perception with language emergence. For example, in Marocco et al. (2003) (a study based on the previous work on active perception by Nolfi and Marocco, 2002) the evolved robot showed a stereotypic behavior towards the object, which allowed the robot to discover the physical properties of that object and then to categorize it to apply the correct linguistic label. The case we are presenting here is similar to Marocco et al. (2003) in many respects. The iCub robot interacts with the objects in a very stereotypic way and the stratagem by means of which the typology of the object is discovered is based on an active sensorimotor strategy which, given the same exploratory behavior, produces different outcomes. We can therefore speculate that the grounding of the linguistic label is not the actual object, but rather the physical property that allows the object (the patient in this case), when manipulated in a certain way, to produce a specific sensorimotor contingency in the agent.

At this point of the discussion we can definitely affirm that, given the present experimental set-up, the meaning of the labels are not associated to a static representation of the object, but to its dynamical properties. It seems, therefore, that the label that we called *linguistic_output_1* (*[1 0 0]*), is related to the rolling of an object, or in general, to those objects that, when touched, move away from the agent. A corresponding tentative word for this can be "the rolling one", more than "sphere". Similarly, *linguistic_output_2* (*[0 1 0]*) seems to be connected with the affordance of an object

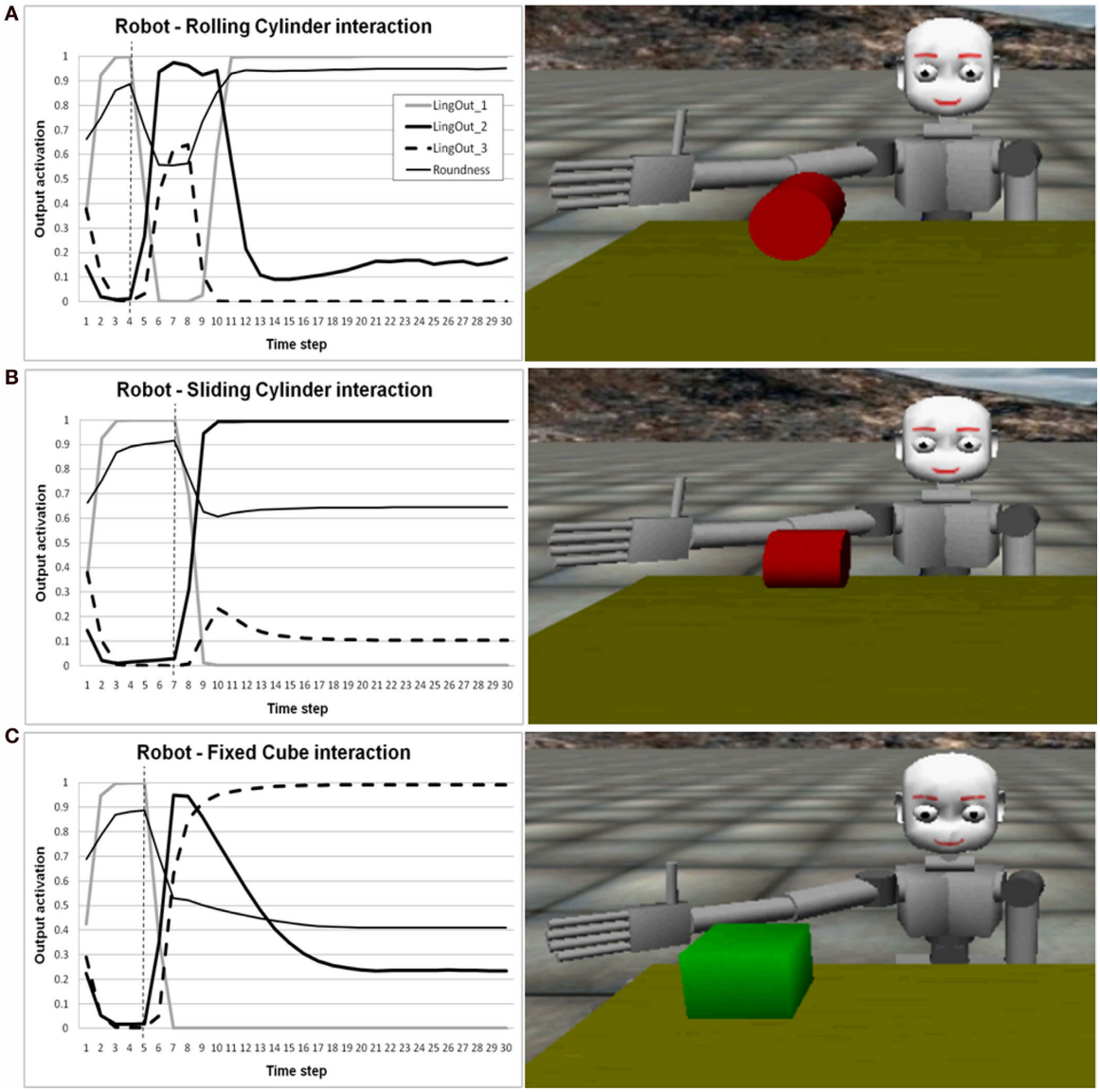

FIGURE 8 | Right column: The three novel conditions used for testing the generalization of the linguistic output units. In the case depicted in **(A)** the cylinder can roll away after being touched by the robot arm. In **(B)** the cylinder tends to slide while in contact with the arm, and in **(C)** the cube is fixed on the table and cannot be moved by the robot. Left column: Activations of linguistic output units during the 15-s interaction between the robot and the object on the desk. The graph shows *linguistic_output_1* (interaction with the rolling cylinder), *linguistic_output_2* (interaction with the sliding cylinder), *linguistic_output_3* (interaction with the fixed cube), and the roundness prediction. Refer to the legend in the graph. The thin dashed vertical line in the three graphs represents the moment in which the robot's arm begins to touch the object.

that can be moved by the agent, for instance, sliding on a surface. Therefore, an appropriate word for this can be "the sliding one" or "the one that slides", rather than "cube". This, indeed, is activated by an object that needs a continuous force applied to it in order to move. *Linguistic_output_3* (*[0 0 1]*), on the other hand, is connected to a fixed object, that is, to an object that does not change its position in space when touched. A word counterpart for this can be "the fixed one". It is interesting to note that "fixed" is not an action. However, it is exactly the property of being fixed (not movable) that, by preventing the robot to accomplish its intended movement, produces the specific sensorimotor contingency that allows the robot controller to identify that particular physical property. All these properties are represented by the control system in terms of the effects produced by the robot itself and dependent on a self-generated movement.

Thus, it appears that in the interpretation of a given dynamic, the robot learns some property related to the force dynamics between objects. This is consistent with Talmy's (1988) cognitive linguistic analysis of the grounding of language in temporal events and, implicitly, of the grounding of action words that describe those events. This concept was explicitly used by Siskind in its software model *LEONARD*, briefly described in the introduction. However, the main difference between our model and Siskind's work is that in Siskind (2001) the force dynamic rules

are explicitly embedded in the perceptual system, as well as the events that the system can recognize, e.g. *PICK-UP* or *MOVE*. In a later work (Fern et al., 2002), Siskind added a learning rule to his system that allows *LEONARD* to learn any kind of dynamic events that are shown to the camera. It should be noted, however, that basic force dynamics rules, called *states* by the authors, such as *CONTACTS*, *SUPPORTS* or *ATTACHED* are still predefined by the experimenter. This new model allows *LEONARD* to learn the temporal sequences of *states* observed in a dynamic event. Therefore, any kind of events can be recognized on the basis of predefined force dynamics states.

A similar consideration can be raised with respect to the work by Baillie and Steels (2003). In their model, events are based on a set of predefined detectors in interaction with a kind of top-down reasoning system that allows an agent to create an internal representation of the external world. This internal world, in turn, is the actual grounding of the utterance produced.

In contrast, in our model the robot is able to capture the essence of certain interactions between objects (i.e. its hand and the object of the desk) and to create an embodied representation of those interactions autonomously. Moreover, the embodied knowledge is implemented in the neural control system as specific dynamical patterns of sensorimotor contingencies and does not require an explicit internal representation of the external word.

Yet the results presented here are in line with the work by Cohen et al. (2005) and Cannon and Cohen (2010) on the grounding of action words. In their work, they refer to the concept of energy transfer between agents, which is to some extent connected to the idea of the force dynamics. In our model, we can analyse the way in which the robot categorizes the events in terms of Cannon and Cohen concepts. However, the main difference is that in their work they only refer to visual stimuli, while in our model the grounding of the action words we discovered is deeply rooted in the integration of many sensor stimuli, both visual and proprioceptive.

From a linguistic perspective, the results obtained are also useful for understanding the acquisition of word meanings by young children. That is, the sensorimotor experience of the robot, rather than visual properties of the objects or the linguistic labels used, constitutes the basis for categorization. The word learning by the robot, therefore, depends on the way in which an object behaves when it is manipulated under certain conditions, rather than on its appearance. The results obtained thus support approaches to word meaning that focus on the role of functional affordances of objects in interaction (Clark, 1973; Nelson, 1973; Mandler, 1992). In particular, the meaning of "ball" is not defined on the basis of its perceptual appearance (its roundness in our case) but on its property to roll. "Sphere" and "cylinder", in contrast, are grouped into the same category because of the action-based sensorimotor categories created by the robot during the training. Thus, what we found the robot to do corresponds to what Mandler (1999, p. 305) suggests for infant word learning:

Infants are attracted by and interested in moving objects from birth. Moving objects are the basis of events, which is what infants attend to, and, according to my theory, it is attended events that get analysed into the first conceptual meanings (Mandler, 1992). Understanding events is absolutely central to conceptual life, and it would be surprising indeed if even infants did not have the capacity to generalize across them. It appears, however, that this kind of generalization is more abstract than seeing the commonalities among cats or chairs. It involves generalizing the common roles that different categories of objects play (for example, animals pick-up objects, artifacts get picked up) and this rather abstract understanding forms the basis on which more detailed, concrete understanding of who does what to whom will develop. (Mandler, 1999, p. 305)

Moreover, the research presented is also in line with the body of theoretical and empirical evidence grown in the past years in support of the role of embodiment and sensorimotor factors in language use (e.g. Barsalou, 1999; Glenberg and Robertson, 2000; Feldman and Narayanan, 2004; Gallese and Lakoff, 2005; Pecher and Zwaan, 2005), yet with different perspective. Barsalou (1999), for example, focuses on modality-specific perceptual and simulation processes within the Perceptual Symbol System hypothesis, based on experiences of sensorimotor, proprioceptive and introspective events, and also dynamic mental representations of object interaction. Glenberg and collaborators (e.g. Glenberg and Kaschak, 2002) focus on the action and embodiment component of language by demonstrating the existence of action-sentence compatibility effects that support an embodied theory of meaning that relates the meaning of sentences to human action and motor affordances. The shared aim of these studies is to demonstrate that language processes cannot be fully understood without taking into account an embodied perspective.

Thus, the principal contribution of the results presented with respect to the current literature concerns the computational feasibility of grounding of action words directly in the way in which an agent interacts with the environment and manipulates it. The dynamical properties of external objects, such as being movable, or being fixed, are embodied and directly represented in the way in which the agent experiences the reactions produced from its own, self-generated, active manipulation of the world on its perceptual system. This mechanism has two related, desired side-effects: (a) a word in the input produces the activation of a whole sensorimotor process and, conversely, (b) the experience of a given sensorimotor contingency recalls in the robot controller the associated word. Therefore, the model shows that meanings related to dynamical properties of external objects, such as *roll* or *fix*, can be fully grounded in the embodied experience of the robot. From this perspective the activity performed by the robot itself is the key that allows to uncover the properties of the objects by means of physical interactions.

These findings confirm and extend the large body of work on computational and robotics models that focuses on the sensorimotor bases of language acquisition. In particular, as we have highlighted in the introduction, this is partly due to shifting the attention from actions that are performed by the agent, or the robot, to actions or properties that relate to external objects.

We conclude by acknowledging that this work, as we have already mentioned in the Section "Materials and Methods", has been carried out in simulation. We do not claim here that all the work done can be easily transferred onto the real robot, as we are aware of the difference between simulation and reality. There are, indeed, many reasons that justify this choice. The most important of them concerns the practical difficulties of carrying out a large number of

experiments by using a real robotic platform and the fact that the iCub available to us is currently not equipped with touch sensors on the body and compliant motors on the arms, which would allow it to cope with rigid and fixed objects, such as the cylinder stuck on the desk, without breaking the robot. Nevertheless, following Ziemke's (2003) consideration, we believe that robot simulations play an important role in cognitive embodied simulations, although results may be less relevant from an engineering point of view (see also Tikhanoff et al., 2008). Therefore, given the modeling purpose of the present work, the fact that experiments have been carried out in simulation should not diminish the scientific relevance of the results achieved.

ACKNOWLEDGMENTS

This work was supported by the European Commission FP7 Project ITALK (ICT-214668) within the Cognitive Systems and Robotics unit (FP7 ICT Challenge 2) and Apple Research and Technology Support (ARTS).

REFERENCES

Bailey, D., Feldman, J., Narayanan, S., and Lakoff, G. (1997). "Modeling embodied lexical development," in *Proceedings of the 19th Cognitive Science Society Conference,* Mahwah, NJ: Erlbaum.

Barsalou, L. (1999). Perceptual symbol systems. *Behav. Brain. Sci.* 22, 577–609.

Beira, R., Lopes, M., Prac, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchiz, F., and Saltar, R. (2006). Design of the robot-cub (iCub) head. *Proc. IEEE Int. Conf. Robot. Autom.* 94–100.

Bergen, B. (2005). "Mental simulation in literal and figurative language," in *The Literal and Non-Literal in Language and Thought*, eds S. Coulson and B. Lewandowska-Tomaszczyk (Frankfurt: Peter Lang), 255–278.

Bowerman, M., and Choi, S. (2001). "Shaping meanings for language: universal and language-specific in the acquisition of semantic categories," in *Language Acquisition and Conceptual Development,* eds M. Bowerman and S. C. Levinson (Cambridge: Cambridge University Press), 475–511.

Bowerman, M. (2005). "Why can't you "open" a nut or "break" a cooked noodle? Learning covert object categories in action word meanings," in *Building Object Categories in Developmental Time*, eds L. Gershkoff-Stowe and D. H. Rakison (Mahwah, NJ: Erlbaum), 209–243.

Cangelosi, A., and Parisi, D. (2002). *Simulating the Evolution of Language.* London: Springer.

Cangelosi, A., and Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cogn. Sci.* 30, 673–689.

Cannon, E. N., and Cohen, P. R. (2010). "Talk about motion: the semantic representation of verbs by motion dynamics," in *The Spatial Foundations of Cognition and Language: Thinking Through Space,* eds K. S. Mix, L. B. Smith and M. Gasser (New York: Oxford University Press), 235–258.

Clark, H. H. (1973). "Space, time, semantics and the child," in *Cognitive Development and the Acquisition of Language*, ed. T. E. Moore (New York: Academic Press), 27–64.

Cohen, P. R., Morrison, C. T., and Cannon, E. (2005). "Maps for verbs: the relation between interaction dynamics and verb use," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence (IJCAI 2005)*.

Dogar, M., Cakmak, M., Ugur, E., and Sahin, E. (2007). "From primitive behaviors to goal-directed behavior using affordances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Dominey, P. (2005). Emergence of grammatical constructions: evidence from simulation and grounded agent experiments. *Connect. Sci.* 17, 289–306.

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

Fern, A. P., Givan, R. L., and Siskind, J. M. (2002). Specific-to-general learning for temporal events with application to learning event definitions from video. *J. Artif. Intell. Res.* 17, 379–449.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). "Learning about objects through action: initial steps towards artificial cognition," in *IEEE International Conference on Robotics and Automation* (Taipei, Taiwan).

Fritz, G., Paletta, L., Breithaupt, R., Rome, E., and Dorffner, G. (2006). "Learning predictive features in affordance based robotic perception systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. (Beijing, China).

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

Gelman, S. A. (2009). Learning from others: children's construction of concepts. *Annu. Rev. Psychol.* 60, 115–140.

Gelman, S. A., and Heyman, G. D. (1999). Carrot-eaters and creature-believers: the effects of lexicalization on children's inferences about social categories. *Psychol. Sci.* 10, 489–493.

Glenberg, A. M. (2007). "Language and action: creating sensible combinations of ideas" in *The Oxford handbook of psycholinguistics*, ed. G. Gaskell (Oxford, UK: Oxford University Press), 361–370.

Glenberg, A. M., and Kaschak, M. P. (2002). Grounding language in action. *Psychon. Bull. Rev.* 9, 558–565.

Glenberg, A. M., and Robertson, D. A. (2000). Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* 43, 379–401.

Halliday, M. A. K. (1985). *An introduction to Functional Grammar.* London: Edward Arnold.

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

Hatch, E. (1983). Psycholinguistics: a second language perspective. Rowley, MA: Newbury House.

Kaplan, F., Oudeyer, P.-Y., and Bergen, B. (2008). Computational models in the debate over language learnability. *Infant Child Dev.* 17, 55–80.

Karmiloff, K., and Karmiloff-Smith, A. (2001). Pathways to language: from fetus to adolescent. Cambridge, MA: Harvard University Press.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Trans. Evol. Comput.* 5, 102–110.

Kozima, H., Nakagawa, G., and Yano, H. (2002). "Emergence of imitation mediated by objects," in *Second International Workshop on Epigenetic Robotics* (Edinburgh, Scotland).

Lakoff, G., and Johnson, M. (1999). *Philosophy In The Flesh: the Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Langacker, R. W. (2008). *Cognitive Grammar*. New York: Oxford University Press.

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

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychol. Rev.* 99, 587–604.

Mandler, J. M. (1999). Seeing is not the same as thinking: commentary on "making sense of infant categorization". *Dev. Rev.* 19, 297–306.

Marocco, D., Cangelosi, A., and Nolfi, S. (2003). The emergence of communication is evolutionary robots. *Philos. Trans. R Soc. Lond. A* 361, 2397–2421.

Metta, G., Fitzpatrick, P., and Natale, L. (2006). YARP: yet another robot platform. *Int. J. Adv. Robot. Sys.* 3, 43–48.

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

Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: from sensory motor maps to imitation. *IEEE Trans. Robot.* 24, 15–26.

Nelson, K. (1973). Some evidence for the cognitive primacy of categorization and its functional basis. *Merrill Palmer Q.* 19, 21–39.

Nolfi, S., and Marocco, D. (2002). "Active perception: a sensorimotor account of object categorization," in *From Animals to Animats 7 – The Seventh International Conference on the Simulation of Adaptive Behavior*, eds B. Hallam, D. Floreano, J. Hallam, G. Hayes and J.-A. Meyer (Cambridge: MIT Press), 266–271.

Oudeyer, P.-Y., and Kaplan, F. (2006). Discovering communication. *Connect. Sci.* 18, 189–206.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286.

Pecher, D., and Zwaan, R. A. (2005). *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*. Cambridge, UK: Cambridge University Press.

Pulvermüller, F., Haerle, M., and Hummel, F. (2001). Walking or talking? Behavioral and neurophysiological

correlates of action verb processing. *Brain Lang.* 78, 134–168.

Roy, D., Hsiao, K., and 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*.

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. I*. Cambridge, MA: MIT Press.

Sachs, J. (1983). Talking about the there and then: the emergence of displaced reference in parent-child discourse. In *Children's language, Vol. 4*, ed. K. E. Nelson (Hillsdale, NJ: Erlbaum).

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15, 31–90.

Smith, L. B., Jones, S. S., and Landau, B. (1996). Naming in young children: a dumb attentional mechanism? *Cognition* 60, 143–171.

Snow, C. E. (1977). "Mothers' speech research: from input to interaction," in *Talking to Children: Language Input and Acquisition,* eds C. E. Snow and C. Ferguson (Cambridge: Cambridge University Press), 31–49.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intell. Syst.* 16, 16–22.

Steels, L. (2003). Evolving grounded communication for robots. *Trends Cogn. Sci. (Regul. Ed.)* 7, 308–312.

Steels, L., and Baillie, J.-C. (2003). Shared grounding of event descriptions by autonomous robots. *Rob. Auton. Syst.* 43, 163–173.

Steels, L., and Kaplan, F. (2000). AIBO's first words: the social learning of language and meaning. *Evol. Commun.* 4, 3–32.

Steels, L., Kaplan, F., McIntyre, A., and Van Looveren, J. (2002). "Crucial factors in the origins of word-meaning," in *The Transition to Language* ed. A. Wray (Oxford: Oxford University Press), 252–271.

Stoytchev, A. (2005). "Behavior-grounded representation of tool affordances," in *International Conference on Robotics and Automation* (Barcelona, Spain).

Sugita, Y., and Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adap. Behav.* 13, 211–225.

Talmy, L. (1988). Force dynamics in language and cognition. *Cogn. Sci.* 12, 49–100.

Tikhanoff, V., Cangelosi, A., Fitzpatrick, P., Metta, G., Natale, L., and Nori, F. (2008). "An open-source simulator for cognitive robotics research: the prototype of the iCub humanoid robot simulator," in *Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08)*, eds R. Madhavan and E. R. Messina (Washington, DC).

Tsagarakis, G., Metta, G., Sandini, D., Vernon, R., Beira, F., Becchi, L., Righetti, J., Santos-Victor, J., Ijspeert, A. J., Carrozza, M. D., and Caldwell, D. G. (2007). iCub: the design and realization of an open humanoid platform for cognitive and neuroscience research. *Adv. Robot.* 21, 1151–1175.

Veneziano, E. (2001). Displacement and informativeness in child-directed talk. *First Lang.* 21, 323–356.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 1550–1560.

Wierzbicka, A. (1985). *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.

Williams, R. J., and Zipser, D. (1995). "Gradient-based learning algorithms for recurrent networks and their computational complexity," in *Backpropagation: Theory, Architectures, and Applications*, eds Y. Chauvin and D. E. Rumelhart (Mahwah, NJ: Lawrence Erlbaum Associates), 433–486.

Yamashita, Y., and Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Comput. Biol.* 4, e1000220. doi:10.1371/journal.pcbi.1000220

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: a computational study. *Connect. Sci.* 17, 381–397.

Ziemke, T. (2003). On the role of robot simulations in embodied cognitive science. *AISB J.* 1, 389–399.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 December 2009; paper pending published: 24 February 2010; accepted: 30 April 2010; published online: 31 May 2010.

Citation: Marocco D, Cangelosi A, Fischer K and Belpaeme T (2010) Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot. Front. Neurorobot. ***4****:7. doi: 10.3389/fnbot.2010.00007*