

The Grounding of Higher Order Concepts in Action and Language: a Cognitive Robotics Model

Francesca Stramandinoli*, Davide Marocco, Angelo Cangelosi

*Centre for Robotics and Neural Systems, University of Plymouth, Devon, PL48AA,
United Kingdom*

Abstract

In this paper we present a neuro-robotic model that uses artificial neural networks for investigating the relations between the development of symbol manipulation capabilities and of sensorimotor knowledge in the humanoid robot iCub. We describe a cognitive robotics model in which the linguistic input provided by the experimenter guides the autonomous organization of the robot's knowledge. In this model, sequences of linguistic inputs lead to the development of higher-order concepts grounded on basic concepts and actions. In particular, we show that higher-order symbolic representations can be indirectly grounded in action primitives directly grounded in sensorimotor experiences. The use of recurrent neural network also permits the learning of higher-order concepts based on temporal sequences of action primitives. Hence, the meaning of a higher-order concept is obtained through the combination of basic sensorimotor knowledge. We argue that such a hierarchical organization of concepts can be a possible account for the acquisition of abstract words in cognitive robots.

Keywords: Symbol grounding, abstract language, computational modeling, recurrent neural network, developmental cognitive robotics

*Corresponding author. Tel.: +44 1752584908

Email addresses: francesca.stramandinoli@plymouth.ac.uk (Francesca Stramandinoli), davide.marocco@plymouth.ac.uk (Davide Marocco), acangelosi@plymouth.ac.uk (Angelo Cangelosi)

1. Introduction

Cognitive robotics, taking inspiration from developmental mechanisms studied in children by psychologists and neuroscientists, aims to overcome the current limitations in robots design (Cangelosi et al., 2010). The main goal of this research area is to design artificial systems that, differently from industrial robots, can act in an unstructured environment and that are capable of reasoning, decision making and autonomous learning. Amongst the various cognitive capabilities (e.g. memory, attention, perception, intuition, mental imagery, etc.) linguistic skills are one of the most powerful tools of an agent for understanding situations and interacting with the environment. Until recently, research studies about concepts formation and naming have mainly focused on the acquisition of concrete words; that is, of words that refer to real and tangible entities. Hence, very little is known about the concepts involved on other aspects of language, such as those representations that refer to combinations of other concepts. In contrast to basic concepts (primitive), that can be directly grounded on the sensorimotor experience of an agent, we refer to this type of language as “higher-order” concepts that are typically based on combinatorial aspects of language. Language, indeed, permits to combine simpler and basic concepts together, in order to create the semantic reference of words that do not have a direct and tangible relation with the perceptual world. This is the case, for example, of abstract words. Whilst for concrete words the semantic reference can be perceived through the senses, and these can be directly linked to physical experiences, abstract words refer to things that are mostly intangible, like for instance “*truth*”, “*democracy*”, or “*justice*” (Barsalou, 2008; Wiemer-Hastings et al., 2001). According to a general definition (Barsalou & Wiemer-Hastings, 2005), abstract words represent everything which is not physically defined nor spatially constrained (e.g. mental states) and refer to things that can be perceived not through the senses but by the mind. The kinds of concepts that lay behind abstract words cannot be directly linked to sensorimotor experience because they cannot be seen or touched and it is not possible to directly interact with them. Some theories (Barsalou, 1999; Glenberg & Kaschak, 2002) suggest that concepts manipulation provided by language is a key element to understand the semantic representation of words that refer to abstract entities. Abstract concepts themselves are formed through language and entail a form of higher-order concepts based upon the combination of simpler representations ((Borghini et al., 2011; Barsalou et al., 2008; Barsalou & Wiemer-

Hastings, 2005)). On the other hand, the distinction between concrete and abstract words cannot be easily regarded as a simple dichotomy (Wiemer-Hastings et al., 2001). Indeed, there is not a “clear cut” between words that can be classified as referring to concrete or abstract entities. There is instead a continuum according to which all entities can be varied in abstractness (e.g. “*scientist*” might be more abstract than “*book*” but less abstract than “*odd number*”) (Borghi et al., 2011). A number of scholars have proposed that when concepts become more detached from the physical entities and more associated with “mental events”, they become increasingly abstract (Barsalou, 1999; Paivio et al., 1968; Wiemer-Hastings et al., 2001). Furthermore, to testify the complexity of the problem, (Altarriba et al., 1999) have proposed to classify terms that refer to emotions in a distinct group from concrete and abstract words. Given the current debate in the field and the complexity of the matter, nowadays the task of creating higher-order concepts through language has been proved an extremely complex task on cognitive robotics.

In this work we aim to address the formation of higher-order concepts and, partially, addressing the dichotomy between concreteness and abstractness, through the implementation of a neuro-robotic model. Through this model we aim at linking symbolic information provided in a form of simplified linguistic token (e.g. *language* component) and the sensorimotor knowledge (e.g. *action*) in order to better clarify the underlying mechanisms involved during the acquisition of the meaning of words characterized by different level of combinatorial complexity. In particular, we argue that higher-order concepts can be indirectly grounded in action primitives, which are directly grounded in sensorimotor experiences (Barsalou, 2008; Glenberg & Kaschak, 2002) through the combinatorial power of language. We also aim to address the relevant problem of the grounding of those higher level concepts and, in turn, the grounding of words that refers to such representations. The experiments performed have been developed on a software environment for the iCub robot (Metta et al., 2008; Tikhonoff et al., 2008, 2011) that has been adopted as a study platform for the research on the grounding of higher-order concepts in humanoid robots.

The paper is organized as follow: section 2 contains an overview of the most relevant embodied theories of language learning and it describes a series of grounded models proposed in literature for modeling language in cognitive systems. Sections 3 describes the feed-forward neural controller of the humanoid robots and the mechanism implemented for the training of the neural network. In section 4 we present an extended version of the previous

model that permits the learning of higher-order concepts based on temporal sequences of action primitives. In section 5 we present the experiments developed for the simulated robot iCub and the achieved results. Conclusions and final remarks on section 6 close this work.

2. Language learning in humans and cognitive robots

One central topic in cognitive science is the acquisition of the meaning of words; it has often seemed as a very complex task because it involves different cognitive capabilities working together (Bloom, 2002). However, the most significant events in language development are concentrated during the first years of life of a child when the brain is maturing and developing all its functionalities. In the research field of language development, different stances have been adopted by scholars (Barrett, 1999). In particular, it is possible to distinguish between “nativism” and “developmentalism”; while for nativist (Chomsky, 1979) some aspects of the language are innate in humans, according to constructivists (Tomasello, 2003), linguistic capabilities are gradually acquired during the course of development. Another distinction can be traced between “domain specific” and “domain general” theories. The former theories assume that cognitive processes are specialized for specific knowledge domains, while for the latter language development is dependent on processes that can handle different knowledge domains. At the core of this debate there is the unsolved problem of explaining conceptual representations in cognitive systems, which becomes central in a debate over the acquisition of word meanings, especially from a computational cognitive science perspective. Two main views have been proposed in the literature on conceptual representations. According to the so called *symbolic theory* (Fodor, 1998) concepts are generated by abstract, arbitrary and amodal symbols for which their internal structures are unrelated to the perceptual states that produce them. Symbolic theory assumes that conceptual representations are non perceptual and cognition and perception are separate systems that work according to different principles (Johnson-Laird, 1983; Pylyshyn, 1973; Posner, 1995). Following this approach, the mind is viewed as a symbol system and cognition is therefore symbol manipulation (Harnad, 1990). A well known criticism on this approach was formulated by Searle with the famous “Chinese Room” argument (Searle, 1980) according to which in traditional computational models that deal with language learning tasks, symbols are self-referential entities that require the interpretation of an external ex-

perimeter to identify the referential meaning of the lexical items. The same problem has been highlighted by (Harnad, 1990) with the name of “symbol grounding problem”.

A different approach is taken by the *embodied theory* of language learning, by which concepts are generated by modal symbols grounded in perception and action systems (Barsalou, 1999; Borghi et al., 2011). In particular, the internal structure of these symbols is modal and they are analogically related to the perceptual states that produce them. On the contrary of symbolic theories, perception and cognition are not independent systems but there is an underlying common representational system (Barsalou, 1999).

A number of different “grounded approaches” for modeling language, in which linguistic abilities are developed through the direct interaction between the cognitive agents and the physical world, have been proposed. 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; Cangelosi, 2010). As pointed out by (Cangelosi & 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 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, 2001; Cangelosi & 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 acquisition has been used extensively by Luc Steels (Steels, 2001). For example, Steels and collaborators (Steels, 2003; Steels et al., 2002) use hybrid population of robots, internet agents and humans engaged in language games. Agents are embodied, one by one, 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 & Kaplan, 2002). Other models focus on the developmental factors that favour the acquisition of language by investigating the role of internal motivation and active exploratory behaviour. (Oudeyer & 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) leads the agent to autonomously focus the attention toward vocal communication 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).

Recently, a growing number of robotics models present connectionist architectures as control systems. That is, the “artificial brain” of the robot is an artificial neural network that typically takes different kinds of sensory information as inputs and activates the robot’s motor joints according to the elaborated output. 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 same procedure was then applied to evolve two different words associated to two different actions performed by the agents. (Sugita & Tani, 2005) describes experiments on a real mobile robot that learns a set of behaviours by interacting with objects that are associated with two-words sentences consisting of a noun and a verb. In (Cangelosi & Riga, 2006) a simulated robot learns to perform a set of basic actions through imitation and is able to recall those actions by their names. Subsequently, the robot learns higher-order concepts by combining words. In (Marocco et al., 2010) a simulated iCub humanoid robot learns an embodied representation of action words through the interaction with the environment and by linking the effects of its own actions with the behaviour observed on the object before and after the action.

Those models show that cognitive robots have already been successfully used for learning words that refer to concrete objects and actions. However, building intelligent systems that can understand the meaning of more abstract words is still a challenging task for cognitive developmental robotics. In the neuro-robotic model we present in this paper we argue that the meaning of higher-order concepts is obtained through the combination of basic sensorimotor concepts and such hierarchical organization of concepts can be a possible account for the acquisition of abstract words in cognitive robots.

3. The acquisition of higher-order concepts in cognitive robots

In this paper we propose a neuro-robotics model for the development of higher-order concepts grounded on basic concepts (primitive actions). The model we present is an extension of the architecture proposed in (Stramandinoli et al., 2010) in which experiments on the grounding of language have been developed on a software environment for the iCub robot. Given the relevance of this architecture a brief account of the model and its main results is provided in the following section together with a description of the robot used for conducting the experiments. Two different neural architectures will be experimented with for the learning of higher-order concepts. The first model uses a feed-forward neural network, as in (Cangelosi & Riga, 2006), to test the learning of composite actions via simple combination of primitives. The second model uses recurrent neural networks as it permits the implementation of the learning of temporal sequences of actions. We will discuss in more details the benefits of using a recurrent architecture in section 4.

3.1. The iCub robotic platform

The humanoid iCub has been adopted as a study platform for the research studies on the grounding of abstract words in cognitive robots. The iCub is an open source robotic platform for research in the field of embodied cognition (Metta et al., 2008). The robot has the dimensions similar to that of a three years old child (104 cm tall) and it is equipped with 53 Degrees Of Freedom (DOF) allocated as follow: six for the head, three for the torso, twelve for the legs and thirty two for each upper limb. In particular, the upper limbs have sixteen DOF each, seven of which are allocated in the arm and nine in the hand (three for the thumb, two for the index, two for the middle finger, one for the coupled ring and little finger and for the adduction/abduction). The iCub is also equipped with different types of sensors, such as digital cameras and microphones, gyroscopes, accelerometers, and force/torque sensors on the motorized joints. A distributed sensorized skin is currently under development. Together with the real robot, a simulation software for the iCub has been developed (Tikhanoff et al., 2008, 2011). This simulator is mainly used for testing the developed algorithms before using the real robotic platform. The simulator has been developed using open source libraries and by collecting data directly from the robot design specifications. The simulated iCub aims at replicate the physics and the dynamics of the real robot and they both have the same software interface (Fig.1). Through

the YARP middleware (Metta et al., 2006), that is an open source framework for decoupling devices from software architecture, it is possible to exchange information between the user code and the robot with its environment.



Figure 1: The iCub: real robotic architecture and software simulator.

In the present experiments, in order to teach the iCub to perform a set of actions primitives, the Action Primitives library of the iCub repository has been used (Pattacini et al., 2010). The library relies on the YARP Cartesian Interface, which allows the user to control the upper limbs of the robot by defining a specific pose (position and orientation in axis-angle representation) for the end-effector. In order to determine the joints configuration to move the robot arms to a desired position, a nonlinear optimization technique is used (Wächter & Biegler, 2006). The Action Primitives library provides a set of functions to perform basic action primitives and to combine them in order to obtain more complex behaviours. With the functions contained in the library it is possible to move the arm of the robot to a specific position, to execute a predefined fingers sequence and to wait for a specific time interval and it is also possible to define complex action by inserting different elementary actions in a sequence.

3.2. Feed-forward neural network for the grounding of higher-order concepts

In (Stramandinoli et al., 2010) the neural controller of a simulated iCub robot has been trained to learn a set of words that express general actions and

characterized by an evident sensorimotor component. Subsequently, combining basic words grounded in sensorimotor experience the robot learns what we call “higher-order” concepts. This experiment takes inspiration from the model proposed in (Cangelosi & Riga, 2006). The training of the robot consists of three incremental steps: (i) the Basic Grounding (BG), (ii) the Higher-order Grounding 1 (HG1) and (iii) the Higher-order Grounding 2 (HG2). During the BG training stage, the simulated robot learns to perform a set of action primitives through direct sensorimotor experience, while the HG1 and HG2 training phases implement the grounding transfer process when the grounding of basic terms is transferred to higher-order words. The training algorithm is a standard back-propagation. The grounding transfer mechanism from basic concepts to higher level concepts consists of multiple steps, depending on the number of action primitives that are combined to obtain higher-order representations. The combination of concept is achieved by providing the robot a linguistic description of the meaning of novel words. For example, in order to transfer the grounding from the basic concept *GRASP* and *STOP* to the higher-order concept *KEEP* we provide to the robot, as input, the linguistic description of the concept we want to teach (Fig. 2). In this specific case is: *KEEP [is] GRASP [and] STOP*.

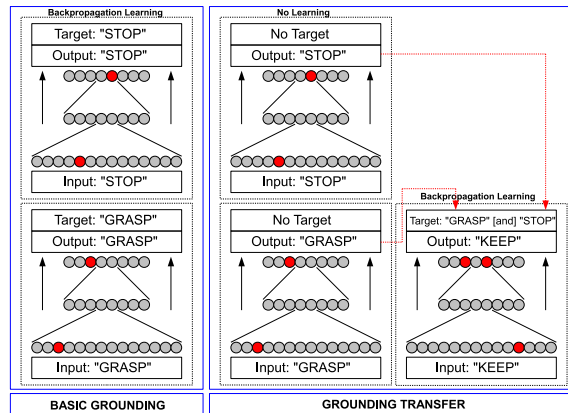


Figure 2: Representation of the procedure that implements the grounding transfer mechanism.

According to the linguistic input provided, the following steps are executed: (1) the neural network receives as input the action primitive words contained in the linguistic description of the higher-order word (*GRASP* and

STOP) separately, it computes the corresponding output and the obtained output are stored. (2) the network receives as input the name of the higher-order word (*KEEP*) that has to be learnt and as target output the combined output calculated during the previous activation phase. (3) a back-propagation training phase is then performed. This procedure is adopted for both HG1 and HG2 training stages. The architecture is a three layers fully connected feed-forward neural network (McClelland et al., 1986; Rumelhart & the PDP Research Group, 1986) with 13 input units, 8 output units and 8 hidden units. Hidden and output nodes are activate according to a sigmoid activation function. That is, the activation of the hidden and outputs nodes y is calculated according to the following (Eq.1)

$$y_j = \varphi(net_j) = 1/(1 + \exp(-net_j)) \quad (1)$$

where net is the net input of a given node and it is defined by (Eq.2)

$$net_j = \sum_{i=1}^n (w_{ij} \cdot x_i) \quad (2)$$

where w_{ij} are the synaptic weights of the neural network and x_i the value of the input units, with $\hat{X} = [\hat{x}_1, \hat{x}_2, \dots, \hat{x}_N] \in \mathbb{R}^N$.

The input units of the network encode the name of the actions that the robot has to learn. The output of the neural network is then used to activate a combination of action primitives that ground the meaning of higher-order concepts on the primitive actions. The actual output of the network is the input for an iCub module that implements the execution of the action primitives on the simulated robot. After this process, whenever a linguistic input, a word, known to the robot is provided, the network selects the appropriate combination of output to be executed.

The network is trained by using a standard back-propagation algorithm ((McClelland et al., 1986; Rumelhart & the PDP Research Group, 1986)), as we already mentioned. The error is calculated according to the following (Eq.3)

$$E = 1/2 \sum_{i=1}^N \sum_{j=1}^M (\hat{Y}_j^i - Y_j^i)^2 \quad (3)$$

where $\hat{Y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M] \in \mathbb{R}^M$ being the target output and $Y = [y_1, y_2, \dots, y_M] \in \mathbb{R}^M$ being the calculated output of the network.

Weights changes are calculated according to the following

$$\Delta w_{ij} = \eta \cdot \delta_i \cdot y_j \quad (4)$$

where η is the learning rate and δ_i is calculated by using (Eq.5) for the output units:

$$\delta_i = (\hat{y}_j - y_j) \cdot y_j \cdot (1 - y_j) \quad (5)$$

and (Eq.6) for the hidden units:

$$\delta_i = \sum_{j=1}^M (\delta_j \cdot w_{ij}) \cdot y_j \cdot (1 - y_j) \quad (6)$$

At the end of the training of the neural network all the actions primitives and higher-order concepts were successfully learnt and the network performed correctly the mapping between inputs and outputs. In particular, in these experiments we have shown a methodology for teaching a robot the meaning of words that express general actions and characterized by a sensorimotor component. The grounding of these words has been subsequently used for teaching the robot the meaning of more abstract words (more details on the model can be found in (Stramandinoli et al., 2010)).

4. Recurrent network for the grounding of higher-order concepts

The methodology used on the previous model was proven to be effective for teaching to a humanoid robot the meaning of words that lack of a concrete referent, although some limitations of the model were also evident. In particular, the activation of the action primitives might not be temporally specified. A temporal specification for action executions is not only important for the control of the robot; it also allows us to increase the combinations of actions that can contribute to the generation of complex movements. This fact, in turn, directly reflects on the number of meaning we can specify for different words. For example, in the previous model there can be no difference between the sequences *KEEP [is] GRASP [and] STOP* and *KEEP [is] STOP [and] GRASP*, as it would be impossible to distinguish the two sentences on the basis of the output activations. Indeed, in both cases the output units corresponding to *GRASP* and *STOP* would be activated simultaneously. On the other hand, a number of studies in neuroscience have

demonstrated that during action words and sentence comprehension different areas of the brain are activated depending on the effector involved (e.g., arm, leg) (Pulvermüller et al., 2001). At the same time, behavioural studies have shown that the comprehension of action verbs and action sentences involves the same sensorimotor and emotional brain circuits that are also activated during the actual interaction with the objects, situations and events the sentences refer to (e.g. (Barsalou, 2008; Glenberg & Kaschak, 2002)). These studies, together with many others, have led to the formulation of the hypothesis that the motor system is activated in a direct way during the processing of words and sentences (see (Mahon & Caramazza, 2008) for evidence against these studies).

4.1. The neural network model architecture

To overcome the limitations of the feed-forward model in terms of time specification and combinatorial ambiguity we developed a new model based on a recurrent neural network. Recurrent neural networks have been used since the beginning of the connectionist era for addressing language related research (Hinton & Shallice, 1991; Elman, 1990) and offer a useful framework for understanding the underlying mechanism in the process of language acquisition and concepts formation, which is strongly related to the problem of modeling short term memory in artificial systems (see (Botvinick & Plaut, 2006) for a critical discussion).

The proposed neural network takes inspiration from the architecture discussed in (Botvinick & Plaut, 2006). In preliminary tests, this architecture produced more stable and reliable learning results than other network topology based on standard simple recurrent networks. The inputs and output encoding of the network layers, as well as the training methodology, are the same as in the feed-forward method. That is, the input of the network is a localistic encoding of 13 words (13 input units) and the output is a localistic encoding of 7 basic action primitives (7 output units), as defined in the Action Primitives library (Pattacini et al., 2010). The hidden layer is formed by 27 units. The input layer is connected to the hidden layer, and the hidden layer to the output. Recurrent connections link the output units to the hidden layer and from units in the hidden layer to all other units in the same layer (Fig.3).

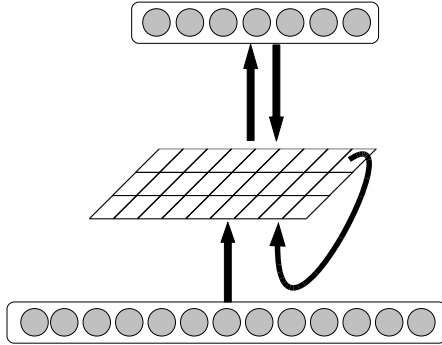


Figure 3: Neural network architecture for the learning of higher-order concepts.

4.2. Training of the model

The training set of this network, differently from the previous architecture, consists of sequences of temporal patterns that encode sequences of actions. The words that directly refer to an action primitive activate a single output pattern, that represents the action to be performed. Whereas, higher-order words activate a sequence of action primitives (Tab.1).

BG	INPUTS												OUTPUTS					
PUSH	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
PULL	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
GRASP	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
RELEASE	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SMILE	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0
FROWN	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
NEUTRAL	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
HG1	INPUTS												OUTPUTS					
GIVE	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
HG2	INPUTS												OUTPUTS					
REJECT	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 1: Training set sample of the network.

As with the feed-forward model, the training of the recurrent neural network is performed by means of the standard back-propagation algorithm described in the previous section. To form higher-order concepts, which refer to words whose meaning is a combination of basic motor primitives, the same methodology of the previous model is adopted. After the neural network

has learnt the associations between basic grounding words and motor action primitives, the various stages that lead to the acquisition of combinatorial meaning are the following:

- the network receives as input the action primitive words that form the linguistic description of the higher-order word we want to teach (e.g., *GIVE [is] GRASP [and] PUSH [and] RELEASE*);
- the motor outputs corresponding to the action primitives words are computed by the network, one by one, and stored one after the other according to the position of the corresponding word within the linguistic description so to form a sequence of primitive actions (note that the sequence, e.g., *GRASP [and] PUSH* is different from the sequence *PUSH [and] GRASP*, since the temporal sequences of motor activations are different).
- the network receives as input the unknown higher-order word and as target output the sequence of motor outputs calculated during the previous activation phase and back-propagation is applied to the new formed training set.

The meaning of a word relies on complex sequences of actions that can be formed iteratively, every time a new linguistic description is provided to the network. Therefore, the activations of the hidden units have to create different temporal patterns according to the different motor actions that define a “meaning” of a given word. Below we show the words and linguistic description we use in the present experiment.

1. Basic Grounding words:

PUSH, PULL, GRASP, RELEASE, SMILE, FROWN, NEUTRAL

2. Higher-order Grounding 1 (HG1):

GIVE [is] GRASP [and] PUSH [and] RELEASE
RECEIVE [is] PUSH [and] GRASP [and] PULL
PICK [is] GRASP [and] PULL [and] RELEASE

3. Higher-order Grounding 2 (HG2):

ACCEPT [is] RECEIVE [and] SMILE
REJECT [is] GIVE [and] FROWN

KEEP [is] PICK [and] NEUTRAL

This methodology is extremely flexible and allows to freely add novel words to the known vocabulary of the robot, or to completely rearrange the word-meaning associations.

5. Simulation results and observations

As described in Section 3.2 and Section 4 the training mechanism of the network consists of three incremental stages. Figures (Fig.4(a),(b),(c)) show the root mean square error calculated at the end of each training stage.

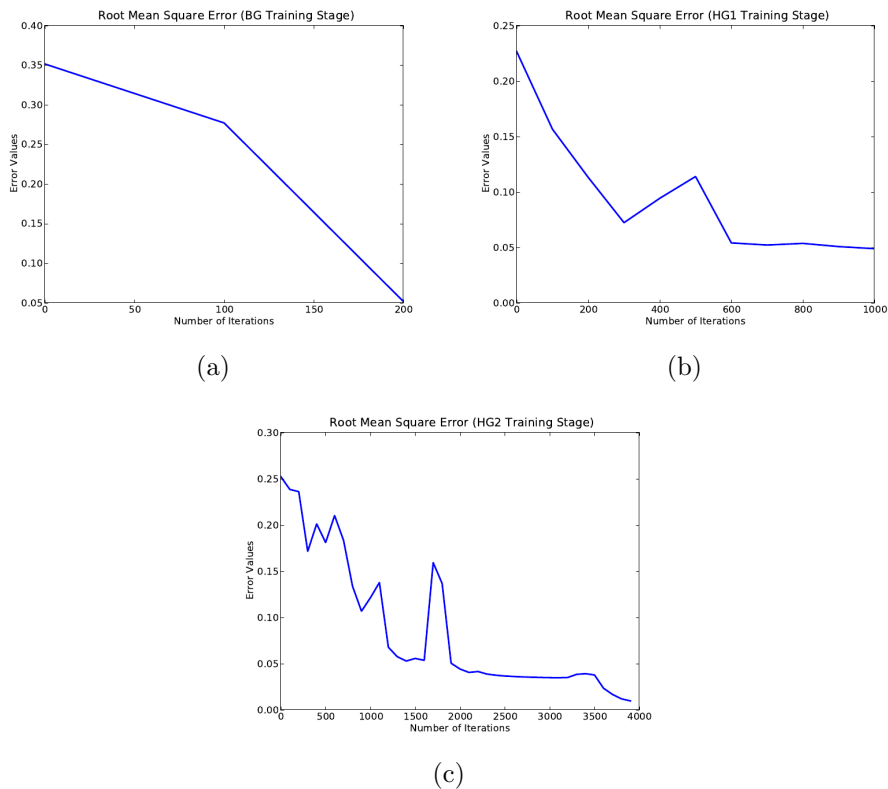


Figure 4: Root Mean Square Error: BG stage (a), HG1 stage (b), HG2 stage(c).

The BG training stage is a simple association between input and output and as it can be observed from (Fig.4(a)) the network is able to learn this task in very few iterations (200). The HG1 and HG2 training stages require

more training cycles, as in these cases the task is much more complex than before, considering that the network has to learn the mapping of single input patterns (the words that the robot has to acquire) with the entire sequences of temporal motor actions, which are arbitrary and, in most cases, of different lengths. The greater complexity of the task is also testified by the irregular shapes of the error curves in (Fig.4(b),(c)).

After the training, tests on the simulated robot show that the neural controller is able to select and activate without errors the correct sequence of motor actions in relation to a word given in input.

It is interesting to note that, in the current model, the implementation of words meaning acquisition takes inspiration from the Perceptual Symbol Systems (PSS) theory proposed in (Barsalou, 1999). During the HG1 and HG2 stages the robot constructs higher-order concepts (e.g. GIVE) by reactivating the internal “mental simulations” of the basic concepts contained in the corresponding linguistic description (*GIVE [is] GRASP [and] PUSH [and] RELEASE*). Moreover, this procedure allows the model to be unaffected by the symbol grounding problem, since the grounding of higher order-concepts is directly grounded on the basic motor actions that constitute the meaning of the basic words (Cangelosi & Riga, 2006).

In order to better understand the internal organisation of the network and its dynamics, the internal units activations in time have been investigated. Since the neural network creates a hierarchical structure of meanings based on the combinations of primitive concepts, our hypothesis was that similar internal representations would be activated whenever a basic concept was recalled.

The analysis of a recurrent network with a sufficiently large number of hidden units poses a number of challenges and it is often difficult to understand and clarify certain dynamics. In our case, in order to show that similar hidden units patterns were activated according to similar primitive actions (a kind of pre-motor activation), a cluster analysis on the internal activations was performed. The results of such analysis were ineffective and showed a complex internal dynamics, with very sparse clusters.

To reduce the dimensionality of the space defined by the 27 hidden units, a Principal Components Analysis (PCA) was performed on the hidden activation values in time. This used the activation patterns of every element of every sequence. Figure 5 shows the trajectories of the various patterns in time within the phase space of the first two principal components (those two components represents the 68% of the data set). The graph clearly shows

that the trajectories of hidden activations are similar according to the meaning of the words (black lines indicate HG1 and grey HG2). For example, *ACCEPT*, represented as a grey dashed line on the graph, shares part of its trajectory with *RECEIVE* (black dashed line), as *ACCEPT* is defined as *RECEIVE [and] SMILE*. Similarly, *REJECT* and *GIVE* (continuous grey and black lines), as well as *KEEP* and *PICK* (dotted grey and black lines) show the same temporal activation patterns. This result indicates, in contrast with our previous expectation, that internal representations for a given action are similar when motor patterns have similar outcomes, but different for different motor sequences.

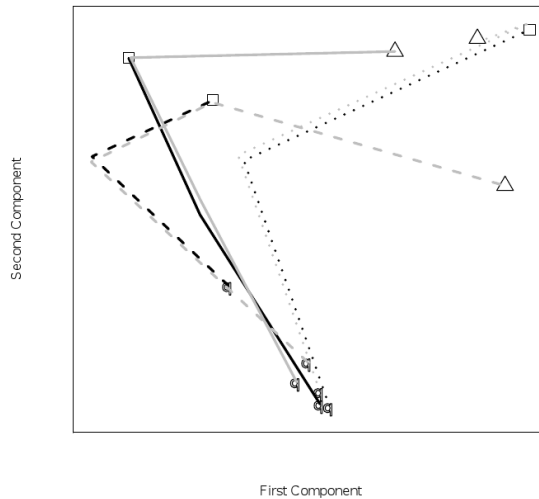


Figure 5: Trajectories of various patterns in time within the phase space of the first two principal components. Circles represent the starting point of a sequence while squares and triangles represent the end point of a time sequence of HG1 and HG2 levels respectively.

Interestingly, such result appears to be consistent with some recent neurophysiological experiments which have shown that motor neurons that encode a specific motor act, like grasping or reaching, present different activation patterns according the final goal of the action sequence in which that particular motor act is embedded. Therefore, a neuron that is highly active during the grasping phase in a “grasping to eat” sequence may show a very little activation during a “grasping to place” sequence (Fogassi et al., 2005). In-

deed, this seems to be particularly consistent with the chain model hypothesis formulated by (Chersi et al., 2006, 2010).

Starting from the evidence that, for sentences that express a motor content, as in our case, the language processing modulates the activity of the motor system, these authors hypothesise that the processing of action-related sentences (and words, we add) involves the activation of the chain - the sequence - of motor neurons directly involved in the sentence. Additional evidences suggest that groups of neurons that represent similar actions are, at least in part, different depending on the overall movement that contains a given action. That is, we should observe that the pool of neurons representing a motor act embedded in several specific movements are only partially similar. Only a fraction of a given pool, specific to a given goal, is activated when the same motor act is embedded on different movements.

In order to investigate whether our model shows the same dynamics, an additional analysis has been conducted. Understanding differences and similarities of the hidden units activation across different patterns on a quantitative basis is not obvious. Therefore, to visually highlight differences and similarities between different patterns, we have plotted the activation values of the hidden units as a matrix of 9×3 elements. Figure 6 shows the result of such visual elaboration, by focusing on the relation between the internal representation observed for the Basic Grounding (e.g., *PULL*) and the internal representation of the same concept embedded in other, high-level words, such as *RECEIVE*, *ACCEPT*, or *PICK*). From this figure we can observe that, by visually comparing the BG with the HGs, the former is very often quite different from the others, and only a small fraction of neurons are activated similarly in all the cases. This is different in the case of words that share part of the meaning, as they share many of the internal representations. This fact was primarily indicated by the previous PCA. Moreover, by comparing the patterns in the other cases, for example the representation of *PUSH* within *RECEIVE* and *PICK*, we can notice that, although some of the activations are in common, the two representations looks quite different. These observations provide some indications that what is hypothesised by (Chersi et al., 2006) can be a general mechanism that explains the way in which recurrent neural networks represents and reuse hierarchical concepts.

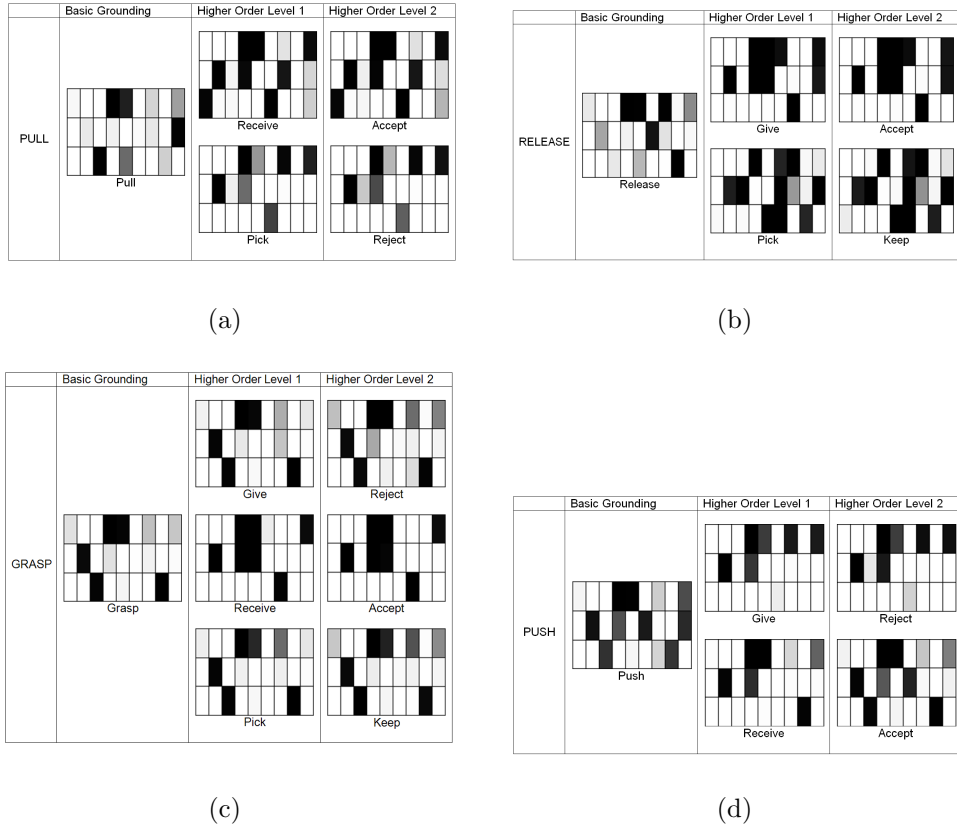


Figure 6: Visual elaboration of activation values of the hidden units as a matrix of 9×3 elements.

6. Conclusion

In this paper we have presented two neural control systems for robot experiments investigating the relations between higher-order symbolic representations and sensorimotor knowledge. The first neural controller is based on feed-forward networks, whilst the second uses recurrent networks to learn higher-order concepts based on sequences of low-level primitives. Our simulation results showed that higher-order symbolic representations can be indirectly grounded in action primitives, which are themselves directly grounded in sensorimotor experiences. Through the analysis of the network dynamics, especially for recurrent architectures, we have observed that motor primitives show different activation patterns according to the action's sequence in

which they are contained. This simulation results are consistent with empirical neuroscience and computational neuroscience studies on action representation that show that the goal of an action changes the substrate of neurons involved in the action processing. Ongoing work is focusing on the extension of the model to include in the network architecture more realistic representations of the perceptual and sensorimotor inputs and output units. For example, instead of using localist representation for actions and words in input and output, the new model directly use output neurons that control individual joints of the robots degrees of freedom. Hence the model can be easily scaled up to handle a large action repertoire, resulting from various combinations of joint activations. In input, more realistic representation of spoken words are being considered. This can be useful for building robots that can be trained through linguistic descriptions provided by users through natural language. The proposed neuro-robotic modelling approach, capable to learn hierarchical higher-order representations based on combination of sensorimotor primitives, can be used to investigate the sensorimotor bases of abstract concepts. This will allow the detailed understanding of the incremental contribution of embodied knowledge in the continuum between concrete words (e.g. push, ball), which are directly grounded in actions and perceptual experience, and the other extreme of abstract words (e.g. truth), where the sensorimotor grounding is based on indirect grounding and metaphorical mechanisms. The same modelling approach can also be extended to investigate the links between action representation and syntax, as in the Action-sentence Compatibility Effect (Glenberg & Kaschak, 2002).

Acknowledgements

This research has been supported by the EU project RobotDoC (235065) from the FP7 Marie Curie Actions ITN and by the EU FP7 Project ITALK (ICT-214668) within the Cognitive Systems and Robotics unit.

References

- Altarriba, J., Bauer, L., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods*, *31*, 578–602.
- Barrett, M. (1999). *The development of language*. Psychology Pr.

- Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*, 577–660.
- Barsalou, L. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617–645.
- Barsalou, L., Santos, A., Simmons, W., & Wilson, C. (2008). Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, (pp. 245–283).
- Barsalou, L., & Wiemer-Hastings, K. (2005). Situating abstract concepts. *Grounding cognition: The role of perception and action in memory, language, and thinking*, (pp. 129–163).
- Bloom, P. (2002). *How children learn the meanings of words: Learning, development and conceptual change*. Cambridge, MA: MIT Press.
- Borghi, A., Flumini, A., Cimatti, F., Marocco, D., & Scorolli, C. (2011). Manipulating objects and telling words: a study on concrete and abstract words acquisition. *Frontiers in Psychology*, *2*.
- Botvinick, M., & Plaut, D. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, *113*, 201–233.
- Cangelosi, A. (2001). Evolution of communication and language using signals, symbols, and words. *IEEE Transactions on Evolutionary Computation*, *5*, 93–101.
- Cangelosi, A. (2010). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of life reviews*, *7*, 139–151.
- Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., Nori, F. et al. (2010). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, *2*, 167–195.
- Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language*. Springer London.
- Cangelosi, A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive science*, *30*, 673–689.

- Chersi, F., Mukovskiy, A., Fogassi, L., Ferrari, P., & Erkhagen, W. (2006). A model of intention understanding based on learned chains of motor acts in the parietal lobe.
- Chersi, F., Thill, S., Ziemke, T., & Borghi, A. (2010). Sentence processing: linking language to motor chains. *Frontiers in Neurorobotics*, 4.
- Chomsky, N. (1979). *Principles and parameters in syntactic theory*.
- Elman, J. (1990). Finding structure in time. *Cognitive science*, 14, 179–211.
- Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. Oxford University Press, USA.
- Fogassi, L., Ferrari, P., Gesierich, B., Rozzi, S., Chersi, F., & Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, 308, 662.
- Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9, 558–565.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42, 335–346.
- Hinton, G., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological review*, 98, 74–95.
- Johnson-Laird, P. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard Univ Pr.
- Kaplan, F., Oudeyer, P., & Bergen, B. (2008). Computational models in the debate over language learnability. *infant and child development*, 17, 55–80.
- Mahon, B., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris*, 102, 59–70.
- Marocco, D., Cangelosi, A., Fischer, K., & Belpaeme, T. (2010). Grounding action words in the sensorimotor interaction with the world: experiments with a simulated icub humanoid robot. *Frontiers in Neurorobotics*, 4.

- Marocco, D., Cangelosi, A., & Nolfi, S. (2003). The emergence of communication in evolutionary robots. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *361*, 2397.
- McClelland, J., Rumelhart, D., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 2: Psychological and Biological Models, Cambridge, MA: MIT PressMass.
- Metta, G., Fitzpatrick, P., & Natale, L. (2006). Yarp: yet another robot platform. *International Journal on Advanced Robotics Systems*, *3*, 43–48.
- Metta, G., Sandini, G., Vernon, D., Natale, L., & Nori, F. (2008). The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems* (pp. 50–56). ACM.
- Oudeyer, P., & Kaplan, F. (2006). Discovering communication. *Connection Science*, *18*, 189–206.
- Oudeyer, P., Kaplan, F., & Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, *11*, 265–286.
- Paivio, A., Yuille, J., & Madigan, S. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, *76*, 1–25.
- Pattacini, U., Nori, F., Natale, L., Metta, G., & Sandini, G. (2010). An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1668–1674).
- Posner, M. (1995). Attention in cognitive neuroscience: an overview. *The cognitive neurosciences*, (pp. 615–624).
- Pulvermüller, F., Härle, M., & Hummel, F. (2001). Walking or talking?: Behavioral and neurophysiological correlates of action verb processing. *Brain and language*, *78*, 143–168.

- Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological bulletin*, 80, 1.
- Rumelhart, D. J. M., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 1: Foundations, Cambridge, MA: MIT Press.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, 3, 417–424.
- Steels, L. (2001). Language games for autonomous robots. *Intelligent Systems, IEEE*, 16, 16–22.
- Steels, L. (2003). Evolving grounded communication for robots. *Trends in cognitive sciences*, 7, 308–312.
- Steels, L., & Kaplan, F. (2002). Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4, 3–32.
- Steels, L., Kaplan, F., McIntyre, A., & Van Looveren, J. (2002). Crucial factors in the origins of word-meaning. *The transition to language*, 2, 4–2.
- Stramandinoli, F., Cangelosi, A., & Marocco, D. (2010). Towards the grounding of abstract words: A neural network model for cognitive robots. In *Proceedings of IJCNN-2011 International Joint Conference on Neural Networks*. IJCNN.
- Sugita, Y., & Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, 13, 33.
- 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 *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems* (pp. 57–61). ACM.
- Tikhanoff, V., Cangelosi, A., & Metta, G. (2011). Integration of speech and action in humanoid robots: icub simulation experiments. *IEEE Transactions on Autonomous Mental Development*, 3, 17–29.

- Tomasello, M. (2003). *Constructing a language*. Harvard University Press.
- Wächter, A., & Biegler, L. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, *106*, 25–57.
- Wiemer-Hastings, K., Krug, J., & Xu, X. (2001). Imagery, context availability, contextual constraint, and abstractness. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 1134–1139).