2014

TOWARDS THE GROUNDING OF ABSTRACT CATEGORIES IN COGNITIVE ROBOTS

Stramandinoli, Francesca

TOWARDS THE GROUNDING OF ABSTRACT CATEGORIES IN COGNITIVE ROBOTS

by

Francesca Stramandinoli

A thesis submitted to the Plymouth University
in partial fulfilment for the degree of

Doctor of Philosophy

School of Computing and Mathematics
Faculty of Science and Technology

June 2014

Francesca Stramandinoli

Towards the Grounding of Abstract Categories in Cognitive Robots

Abstract

The grounding of language in humanoid robots is a fundamental problem, especially in social scenarios which involve the interaction of robots with human beings. Indeed, natural language represents the most natural interface for humans to interact and exchange information about concrete entities like KNIFE, HAMMER and abstract concepts such as MAKE, USE. This research domain is very important not only for the advances that it can produce in the design of human-robot communication systems, but also for the implication that it can have on cognitive science.

Abstract words are used in daily conversations among people to describe events and situations that occur in the environment. Many scholars have suggested that the distinction between concrete and abstract words is a continuum according to which all entities can be varied in their level of abstractness.

The work presented herein aimed to ground abstract concepts, similarly to concrete ones, in perception and action systems. This permitted to investigate how different behavioural and cognitive capabilities can be integrated in a humanoid robot in order to bootstrap the development of higher-order skills such as the acquisition of abstract words. To this end, three neuro-robotics models were implemented.

The first neuro-robotics experiment consisted in training a humanoid robot to perform a set of motor primitives (e.g. PUSH, PULL, etc.) that hierarchically combined led to the acquisition of higher-order words (e.g. ACCEPT, REJECT). The implementation of this model, based on a feed-forward artificial neural networks, permitted the assessment of the training methodology adopted for the grounding of language in humanoid robots.

In the second experiment, the architecture used for carrying out the first study was reimplemented employing recurrent artificial neural networks that enabled the temporal specification of the action primitives to be executed by the robot. This permitted to increase the combinations of actions that can be taught to the robot for the generation of more complex movements.

For the third experiment, a model based on recurrent neural networks that integrated multi-modal inputs (i.e. language, vision and proprioception) was implemented for the grounding of abstract action words (e.g. USE, MAKE). Abstract representations of actions ("one-hot" encoding) used in the other two experiments, were replaced with the joints values recorded from the iCub robot sensors.

Experimental results showed that motor primitives have different activation patterns according to the action's sequence in which they are embedded. Furthermore, the performed simulations suggested that the acquisition of concepts related to abstract action words requires the reactivation of similar internal representations activated during the acquisition of the basic concepts, directly grounded in perceptual and sensorimotor knowledge, contained in the hierarchical structure of the words used to ground the abstract action words.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Angelo Cangelosi for the support throughout my doctoral studies and research activity. His guidance through his knowledge, experience and continuous feedback was precious during all the stages of my PhD.

I would also like to thank Dr. Davide Marocco and Prof. Tony Belpaeme for the stimulating discussions, comments and feedback regarding my research studies.

My thanks to Dr. Francesco Nori and Prof. Tom Ziemke for the opportunity of going on secondment to their research laboratories and for providing me insightful feedback concerning my research studies.

I would also like to thank Dr. Elena Dell'Aquila for the support in acquiring new key transferable skills that enhanced the development of my scientific career.

I am thankful to my colleague and friend Marek Ruciński for his encouragement, suggestions and stimulating discussions during our daily coffee break. I learned a lot from him through his scholarly and personal interactions.

I would like to acknowledge the RobotDoC European project for providing me numerous opportunities to improve my research skills and develop personal effectiveness and communication skills which had a tremendous impact on the progression of my scientific career. The sharing of knowledge with experts and young researchers in the RobotDoC network fostered my creative thinking and provided opportunities to engage in productive collaborations.

And last but not the least, I would like to thank my parents, my sister and my brother in law, and my brother, for always supporting my decisions and for the encouragement during difficult moments. A special dedication to my nephew Luigi and niece Ludovica.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee.

Work submitted for this research degree at the Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment.

This study was financed by the EU project RobotDoC (235065) from the Seventh Framework Programme (FP7), Marie Curie Actions Initial Training Network.

A programme of advanced PhD training activities was undertaken; it included all the RobotDoC Training Milestones (TMs): Cognitive Robotics Research Methods Workshop (TM1), Project Proposal Workshop (TM2), Interdisciplinary Methods Spring School (TM3), PhD Transfer Workshop & Mid-Project Review Meeting (TM4), Entrepreneurship Workshop (TM5), Postgraduate Conference on Robotics and Development of Cognition (TM6), Spring School of Developmental Robotics and Cognitive Bootstrapping (TM7) and the RobotDoc International Conference on Development of Cognition (TM8). Computational and cognitive modelling knowledge has been acquired by attending all the seminars organized by the Centre for Robotics and Neural Systems at Plymouth University. Programming and machine learning skills have been enhanced by attending the Machine Learning Summer School MLSS 2010 and the iCub Summer Schools vvv10-11. Additionally, research and professional skills have been developed by attending some of the courses organized by the Graduate School at Plymouth University: Introduction to R, Introduction to applying for Research Funding, MatLab User Workshop.

Relevant scientific seminars and conferences, at which work was presented, were regularly attended; external institutions were visited for consultation purposes and several papers prepared for publication.

Publications:

- **2013.** Stramandinoli F., Marocco D., Cangelosi A., *"Grounding Abstract Action Words through the Hierarchical Organization of Motor Primitives"*, Proceedings of IEEE International Conference on Development and Learning and Epigenetic Robotics, IEEE ICDL-Epirob, Osaka, Japan

- **2012.** Stramandinoli F., Cangelosi A., Wermter S., *"Special issue on advances in developmental robotics"*, In Paladyn. Journal of Behavioral Robotics, vol.3, n.3, p.112, DOI: `http://dx.doi.org/10.2478/s13230-013-0112-x`, SP Versita

- **2012.** Stramandinoli F., Marocco D., Cangelosi A., *"The Grounding of Higher Order Concepts in Action and Language: a Cognitive Robotics Model"*, In Neural Networks, vol. 32, pp 165–173, DOI: `http://dx.doi.org/10.1016/j.neunet.2012.02.012`

- **2012.** Bilotta E., Cerasa A., Pantano P., Quattrone A., Staino A., Stramandinoli F., *"Evolving Cellular Neural Networks for the Automated Segmentation of Multiple Sclerosis Lesions"*, In Raymond Chiong, Thomas Weise and Zbigniew Michalewicz (Eds), Variants of Evolutionary Algorithms for Real-World Applications, pp 377–412, DOI: `http://dx.doi.org/10.1007/978-3-642-23424-8_12`

- **2012.** Stramandinoli F., Marocco D., Cangelosi A., *"A Neuro-Robotics Model for the Acquisition of Higher Order Concepts in Action and Language"*, In N. Miyake, D. Peebles & R. P. Cooper (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society, Austin, TX: Cognitive Society

- **2012.** Rucinski M., Stramandinoli F., *"An Embodied View on the Development of Symbolic Capabilities and Abstract Concepts"*, In Proceedings of the Postgraduate Conference on Robotics and Development of Cognition, Lausanne, 10-12 September pp. 62–63, DOI: `http://dx.doi.org/10.2390/biecoll-robotdoc2012-19`

- **2011.** Stramandinoli F., *"Neurorobotics Models for the Grounding of Abstract Words"*, In A. Wisniewska et al. (Eds.), Proceedings of Science - Passion, Mission, Responsibilities - Marie Curie Researchers Symposium, p. 67, Warsaw, 25-27 September

- **2011.** Stramandinoli F., Cangelosi A., Marocco D., *"Towards the Grounding of Abstract Words: A Neural Network Model for Cognitive Robots"*, In Proceedings of the 2011 International Joint Conference on Neural Networks, San Jose, pp. 467–474, DOI: `http://dx.doi.org/10.1109/IJCNN.2011.6033258`

- **2011.** Stramandinoli F., Rucinski M., Znajdek J., Rohlfing K.J., Cangelosi A., *"From Sensorimotor Knowledge to Abstract Symbolic Representations"*, In Procedia Computer Science, FET11 - The European Future Technologies Conference and Exhibition, (**Second Prize, Best Poster**), Budapest, 4-6 May, vol. 7, pp. 269–271, DOI: `http://dx.doi.org/10.1016/j.procs.2011.09.018`

- **2010.** Bilotta E., Cerasa A., Pantano P., Quattrone A., Staino A., Stramandinoli F., *"A CNN Based Algorithm for the Automated Segmentation of Multiple Sclerosis Lesions"*, In Applications of Evolutionary Computation, pp. 211–220, DOI: `http://dx.doi.org/10.1007/978-3-642-12239-2_22`

Presentation and Conferences Attended:

- **2013.** *"Grounding Abstract Action Words through the Hierarchical Organization of Motor Primitives"*, IEEE International Conference on Development and Learning and Epigenetic Robotics, IEEE ICDL-Epirob, Osaka, Japan (POSTER)

- **2013.** *"Towards the Grounding of Abstract Categories in Cognitive Robots"*, International Conference on Development of Cognition, Osaka, Japan (ORAL)

- **2013.** *"Towards the Grounding of Abstract Concepts: An Embodied Approach"*, Internal Adaptive Behaviour and Cognition Laboratory Seminar, Plymouth, UK (ORAL)

- **2012.** *"Towards the Grounding of Abstract Concepts: An Embodied Approach"*, Visiting Researcher at the Italian Institute of Technology, Robotics, Brain and Cognitive Sciences Department Genoa, Italy (ORAL)

- **2012.** *"Towards the Grounding of Abstract Categories"*, Visiting Researcher at the University of Skvde, Cognition and Interaction Laboratory Skvde, Sweden (ORAL)

- **2012.** *"A Neuro-Robotics Model for the Acquisition of Higher Order Concepts in Action and Language"*, Annual Conference of the Cognitive Science Society, CogSci, Sapporo, Japan (POSTER)

- **2012.** *"Why the Grounding of Language in Humanoids Matters"*, Entrepreneurship Workshop, Genoa, Italy (ORAL)

- **2011.** *"Truth and Trust in Communication"*, Invited Speaker at the Innovation Convention, European Commission, Brussels, Belgium (ORAL)

- **2011.** *"Neurorobotics Models for the Grounding of Abstract Words"*, Science, Passion, Mission, Responsibilities - Marie Curie Researchers Symposium, Warsaw, Poland (POSTER)

- **2011.** *"Towards the Grounding of Abstract Categories in Cognitive Robots"*, PhD Transfer Workshop & Mid-Project Review Meeting, Barcelona, Spain (ORAL)

- **2011.** *"Towards the Grounding of Abstract Words: A Neural Network Model for Cognitive Robots"*, International Joint Conference on Neural Networks, San Jose, California (ORAL)

- **2011.** *"Towards the Grounding of Abstract Words in Cognitive Robots"*, Internal Adaptive Behaviour and Cognition Laboratory Seminar, Plymouth, UK (ORAL)

- **2011.** *"From Sensorimotor Knowledge to Abstract Symbolic Representations"*, FET11 - The European Future Technologies Conference and Exhibition (**Second Prize, Best Poster**), Budapest, Hungary (POSTER)

- **2010.** *"Towards the Grounding of Abstract Categories in Cognitive Robots"*, Project Proposal Workshop, Bielefeld, Germany (ORAL)

- **2010.** *"Investigating the Grounding of Abstract Categories in Humanoid Robots through Multidisciplinary Collaborations"*, Machine Learning Summer School MLSS, Canberra, Australia (POSTER)

- **2010.** *"Investigating the Grounding of Abstract Categories in Humanoid Robots through Multidisciplinary Collaborations"*, ESOF 2010 Marie-Curie Conference, Turin, Italy (POSTER)

- **2010.** *"Personal Introduction"*, Cognitive Robotics Research Methods Workshop, Plymouth, UK (ORAL)

External Contacts:

- **FRANCESCO NORI**
  Team Leader and Researcher
  Robotics Brain and Cognitive Science Department, Cognitive Humanoids Lab,
  Istituto Italiano di Tecnologia
  Telephone: +39 (0) 10 71 781 420
  Email: francesco.nori@iit.it
  Mailing Address: Via Morego 30, 16163 Genova, Italy

- **TOM ZIEMKE**
  Professor in Cognitive Science and Cognitive Robotics
  Interaction Lab, Informatics Research Centre
  Telephone: +46 (0) 500 44 83 30
  Email: tom.ziemke@his.se
  Mailing Address: University of Skvde, School of Humanities and Informatics,
  PO Box 408, 54128 Skvde, Sweden

- **STEFAN WERMTER**
  Professor in Computer Science
  Knowledge Technology, Department of Computer Science
  Telephone: +49 (0) 40 428 83 2434
  Email: wermter@informatik.uni-hamburg.de
  Mailing Address: University of Hamburg, Vogt-Klln-Str. 30, 22527, Hamburg,
  Germany

Honors and Awards:

- **2013.** *"Guest Editor of the Advances in Developmental Robotics Special Issue".* In Paladyn. Journal of Behavioral Robotics, co-edited with Prof. Angelo Cangelosi and Prof. Stefan Wermter

- **2012.** *"Conference General Chair".* Postgraduate Conference on Robotics and Development of Cognition (RobotDoC-PhD), Lausanne, Switzerland, 10-12 September

- **2012.** *"Plymouth Special Recognition Award".* Recognition of the achievement in demonstrating a drive to enhance employability, Plymouth University, United Kingdom, 26 April

- **2011.** *"Invited Speaker at the Innovation Convention".* Event organized by the European Commission, Square - Brussels Meeting Centre, Brussels, Belgium, 5-6 December

- **2011.** *"Second Prize, Best Poster".* FET11 - The European Future Technologies Conference and Exhibition, Budapest, Hungary, 4-6 May

- **2010.** *"Scholarship Award".* Machine Learning Summer School 2010, Canberra, Australia, 27 September-6 October

Word count of main body of thesis: 44488

Signed .....

Date .....

Quote

Abstract concepts pose a classic challenge for grounded cognition.
How can theories that focus on modal simulations explain
concepts that do not appear modal?
- L. W. Barsalou

Chapter 1

Introduction

Amongst the various cognitive capabilities (e.g. memory, attention, perception, action, problem solving, intuition, mental imagery, etc.) linguistic skills are one of the most powerful tools available to an agent for understanding situations and interacting with the environment. Until recently, research studies about concepts formation have mainly focused on the acquisition of concrete words; hence, very little is known about the representation of abstract language. On the contrary of concrete words, that can be perceived through the senses and that can be directly linked to the physical experience that occurs with them, abstract words refer to things that are intangible and that are not physically defined nor spatially constrained [Barsalou, 2008, Barsalou and Wiemer-Hastings, 2005, Wiemer-Hastings et al., 2001]. For this reason finding a semantic representation of abstract words has often appeared as a problematic and challenging task within developmental neuro-robotics. Indeed, until recently one of the main focus of developmental neuro-robotics has been the study of sensorimotor skills and the naming of concrete objects, and only very recently few developmental neuro-robotics models have started to investigate the acquisition of abstract words.

This thesis addresses the problem of **Grounding Abstract Categories and Words in Cognitive Robots**; within this framework, the implementation of neuro-robotics models permitted the investigation of the relations between the development

of abstract symbolic representations (e.g. *language*) and sensorimotor knowledge (e.g. *action* and *vision*). Semantic representations of abstract words were obtained through the integration of linguistic, perceptual and sensorimotor experience of a humanoid robotic platform (i.e. iCub). The implementation of cognitive robotics models that link sensorimotor experience (e.g. the action of *pushing* or *lifting*) to abstract symbolic knowledge (e.g. abstract symbols related to the concept of *using a tool*) enabled the acquisition of semantic representations related to abstract words in artificial agents [Barsalou, 2008, Glenberg and Kaschak, 2002].

In this thesis three experimental studies on the grounding of abstract categories in cognitive robots are presented. The first experiment consisted in grounding the meaning of higher-order words like *ACCEPT*, *REJECT* and *PICK* in the iCub sensorimotor experience. In order to achieve the goal of this experiment, the iCub first learned to perform a set of concrete motor primitives (e.g. *PUSH*, *PULL*, *GRASP*, etc.) and then, by combining such primitives together, the robot derived the meaning of higher-order words. This first model, based on a feed-forward artificial neural network controller, permitted the testing of the training methodology adopted for the grounding of language in the iCub robot.

In the second experiment, the architecture used for carrying out the first study was reimplemented employing recurrent artificial neural networks that permitted the temporal specification of the action primitives to be executed by the robot. This permitted to increase the combinations of actions that can be taught to the robot for the generation of more complex movements.

For the third experiment, a model based on recurrent neural networks that integrated multi-modal inputs (i.e. language, vision and proprioception) and that took into account a more realistic representation of the sensorimotor inputs of the iCub robot was implemented. More complex actions (e.g. "CUT", "HIT", "PAINT", etc.) were built by integrating low level motor primitives (e.g. "PUSH - PULL", "LIFT - LOWER", "MOVE LEFT - MOVE RIGHT") iterated for a certain number of time steps. Abstract representations of actions ("one-hot" encoding) used in

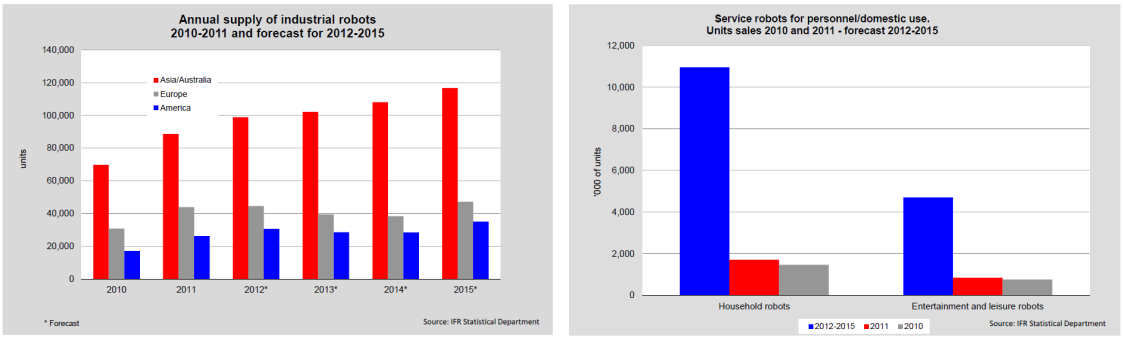
the other two experiments, were replaced with the joints values recorded from the iCub robot sensors. In this model the acquisition of lexical categories was achieved by integrating three different modality inputs: proprioception (joint values), vision (object features) and language (sentences consisting of a verb and a noun). Through the implementation of this model, the hierarchical organization of concepts directly linked to sensorimotor experience permitted the acquisition of higher-level words and categories. The robot learned to generalise the meaning of words like *USE* and *MAKE* by performing a set of iterative actions (e.g. *CUTTING*, *HITTING*, *DRAWING*) for each of which the appropriate tool was employed.

All the experiments were developed using the iCub humanoid platform, a robot that has the same dimension of a three and a half year old child and that is widely used for developmental robotics research [Metta et al., 2008]. In order to verify the validity of the implemented models, experiments were first tested in a simulated environment for the iCub [Tikhanoff et al., 2008, 2011] and subsequently transferred to the real robotic architecture. Since the iCub Simulator and the real robot have the same software interface, the transfer of simulated experiments to the physical robot did not require any particular modification of the implemented models (though extra work was required to handle visual input stream and motor performances).

1.1 Timeliness and Impact of Research

In 2006 Bill Gates compared the current robotics industry to the computers industry of thirty years ago. In thirty years the computer market has become such that nowadays computers are part of our daily life. The same widespread role might be played by robots by thirty years. Indeed, the robotics market is currently widely spreading and many countries are investing on it [IFR, 2012]. Robotics is a highly multidisciplinary discipline which requires knowledge ranging from electronics, mechanics to computer science and so on. Therefore advances on this discipline are a good indicator for the identification of the level of technological progress of a

country. Japan is considered as the country in full expansion for humanoid robotics while the United States of America are leading the military robotics field. The tremendous investments done by the Defense Advanced Research Projects Agency (DARPA) can in part explain the development of military robotics in the United States of America. The growing interest in robotics research in the United States of America is also witnessed by the recent investments made by Google, which has recently acquired the Boston Dynamics engineering company. The Boston Dynamics contracts for the US military and developed the world's fastest-running robot and other animal-based mobile research machines. In Europe, Germany is the country with a prominent role in industrial robotics. According to the 2012 Executive Summary released by the International Federation of Robotics [IFR, 2012], 2011 was the most successful year for **industrial robots** since 1961, considering that robot sales increased by 38% to 166,028 units, by far the highest level ever recorded for one year. The countries that experienced the biggest growth were China, United States and Germany (Fig.1.1(a)).



(a) (b)

Figure 1.1: Annual supply of industrial robots and forecast (a), Sales and forecast for service robots for personal domestic use (b). Source IFR Statistical Department

However, the two biggest markets still remain Japan and the Republic of Korea. Concerning the European market in 2011, about 43,800 **industrial robots** were sold, 43% more than in 2010. About 19,533 new industrial robots were supplied to Germany and following, in Italy the total sales of industrial robots were up by 13% to 5,091 units. Regarding the worldwide annual supply for **service robots**, in

2011 about 2.5 million service robots for personal and domestic use were sold, 15% more than in 2010 and projections for the period 2012-2015 predict that about 15.6 million units of service robots for personal use to be sold (Fig.1.1(b)). This market will increase substantially within the next 20 years.

Hence, research in the field of cognitive systems and human-robot interaction is very timely and of great relevance Nevertheless, before personal domestic robots can be employed in everyday life, developmental neuro-robotics has to face the challenge of building robots capable of working independently, which can autonomously react to dynamic changes that occur in the environment. Providing robots with the capability to comprehend and produce language in a "human-like" manner represents a powerful tool for flexible and intelligent interaction between robots and human beings. Robots endowed with linguistic capabilities could better understand situations and exchange information; through language robots could cooperate and negotiate with human beings in order to accomplish shared plans.

1.2 Objectives and Motivation

Scientists have the extraordinary opportunity to concur to the production of the knowledge that can introduce improvements to people daily life. Scientists working in developmental neuro-robotics can contribute to the achievement of a long-term goal that this research field establishes; that is, the understanding of aspects of human intelligence by building autonomous robots. In the short-term period, robotic platforms provide a useful tool for studying and testing human-robotic interaction based on mutual understanding achievable by reciprocal verbal communication. The grounding of language in robots is a fundamental problem especially in social scenarios in which a robot, for example, can be a co-worker in housekeeping activities or a caregiver for aged people: language can facilitate the human-robot "symbiosis". Natural language represents the most natural interface for people without expertise in formal programming syntax to interact with robots. This research domain is very

important not only for the advances that it can produce in the design of human-robot communication systems, which can lead to a new generation of interactive robotic systems, but also for the implication that it can have on cognitive science. Robots endowed with the capability to understand language and that can adapt their behaviour according to human request could have an important impact on the robotics industry in existing and emerging markets.

Language represents a powerful tool to interact with other agents in order to plan new tasks, make decisions and perform joint activities [Tomasello et al., 2005, Warneken and Tomasello, 2007]. In this context, language represents a collection of shared meanings that enable a common ground with other agents. While for human beings language development is a natural and spontaneous process that occurs over the entire course of their life, for artificial agents (e.g. humanoid robots) one of the major challenges that has still to be faced, involves natural language understanding and processing that can enable agents to derive meaning from natural language inputs.

The development of linguistic skills requires different cognitive capabilities working together; hence, the research field related to the grounding of abstract categories represents a broad domain in which studies ranging from neuroscience to psychology, robotics and computer science can all contribute in order to get a deeper understanding of the integration of linguistic and cognitive skills. The aim of this research studies was to **Ground Abstract Categories in Cognitive Robots** through the implementation of neuro-robotic models that permitted to address the following scientific questions:

- How can cognitive systems (such as robots) use sensorimotor categories to indirectly ground abstract concepts?
- What kind of embodiment and grounding mechanisms are used to combine words?
- How can the symbol grounding mechanism be extended to generate and ground

abstract categories in artificial cognitive systems?

In the novel studies presented in this thesis, the problem of the acquisition of language in robotic platforms is addressed by following the developmental approach to robotics. Differently from classical natural language processing methodologies, developmental robotics considers language embodied in perceptual and sensorimotor knowledge [Asada et al., 2001, Cangelosi and Schlesinger, 2014].

The achievement of the presented research objectives permitted the endowing of robots with basic linguistic skills and further the investigation of the mechanisms underlying language development. Indeed, the analysis of the internal dynamics of such models permitted the investigation of the relations between the development of *abstract symbolic representations* and *sensorimotor knowledge*, in order to understand the underling mechanisms involved during the acquisition of the meaning of abstract words through sensorimotor experience.

1.3 Contribution to Knowledge

The contribution to knowledge of this thesis is summarised herein:

- Presentation of general cognitively inspired design mechanisms for the acquisition of abstract language in the iCub humanoid robot. The studies presented in this thesis represent pioneering work on the grounding of abstract language in cognitive robots. They attempt to fill-in the gap in this research domain by making the first step toward understanding the relation between the development of abstract symbolic representation and sensorimotor knowledge.

- Presentation of a training methodology for humanoid robots that enabled the investigation of the sensorimotor bases of abstract concepts. The application of this methodology permitted to better understand the incremental contribution of embodied knowledge in the continuum between concrete words (e.g. *PUSH*, *KNIFE*), which are directly grounded in actions and perceptual experience,

and abstract words (e.g. *USE*, *MAKE*), for which the sensorimotor grounding is based on indirect experience.

- Presentation of results which provided new knowledge on how to build robots that can interact with other agents in the environment simply processing linguistic descriptions provided by users through language. Such experimental results were presented at the International Joint Conference on Neural Networks in 2011 in a paper titled "Towards the Grounding of Abstract Words: A Neural Network Model for Cognitive Robots" and in a paper titled "The Grounding of Higher Order Concepts in Action and Language: a Cognitive Robotics Model" published on the Neural Networks journal. Results related to the third experiment will be published on a journal paper, which is currently under preparation.

The performed studies have suggested the hypothesis that the acquisition of concepts related to abstract action words requires the reactivation of similar internal representations activated during the acquisition of the basic concepts, directly grounded in perceptual and sensorimotor knowledge and contained in the hierarchical structure of the words used to ground the abstract action words. Therefore, in this study the semantic/conceptual representation of abstract action words consists of reusing sensorimotor and perceptual representational capabilities (embodied understanding of abstract language).

1.4 Structure of the Thesis

The thesis is organized as follows:

- Chapter 2 reviews the literature on the embodied view of cognition applied to language. The chapter presents an overview of the most relevant theories of cognition proposed in literature, such as classical and grounded approaches. The main milestones that occur during language development in

humans are briefly described. Furthermore, the chapter tries to clarify what differs between concrete and abstract words and their representation. The chapter also presents the neural bases of language by describing neurophysiological and behavioural studies about action verb processing.

- Chapter 3 introduces some of the most important approaches proposed in the field of artificial intelligence for knowledge representation and the modelling of language in cognitive systems. A brief overview on symbolic, sub-symbolic and statistical models is provided. Further details are given on the developmental approach to robotics and, grounded and embodied connectionist models that permitted to better assess the current state of the art on the grounding of language in cognitive robotics. The chapter also contains a description of the hardware and software architecture of the iCub robotics platform.

- Chapter 4 provides an introduction on the main methods used in the PhD research. It covers Artificial Neural Networks and it presents some of the main models of artificial neurons. This chapter also presents some of the learning algorithms used in the implementation of the models presented in the experimental part of this thesis. Further, the chapter contains a description of the methods employed for analysing the internal dynamics of the models implemented for carrying out the experimental studies presented in Chapter 5, Chapter 6 and Chapter 7 of this thesis.

- Chapter 5 presents the first robotics experiment on the learning of higher-order concepts for the iCub robotic platform. The chapter, after introducing the theoretical background of the proposed experiment, describes the robot control model, based on feed-forward artificial neural networks, for teaching the robot the meaning of words that lack of a direct concrete referent such as "'ACCEPT" and "REJECT". The training of this model was effective although some limitations of its implementation were evident.

- Chapter 6 describes an extension of the model presented in Chapter 5. This

model uses recurrent artificial neural networks which permitted the introduction of the temporal specification of the input/output patterns used for the training of the robot. After presenting the architecture of the model and the adopted training strategy, simulation results and observations are discussed.

- Chapter 7 presents a model based on recurrent neural networks that enabled the learning of words through the integration of multi-modal inputs (i.e. language, vision and proprioception) and permitted to specify new motor encoding (i.e. action primitives). The chapter describes the implemented architecture, including the input and output coding, and the robotic task together with the related training strategy adopted for the robot. The chapter also describes the evaluation settings used for the model and the related results obtained by running simulations with the iCub robot. In particular, experiments were run in different training and testing conditions. Moreover the ability of the model to generalize new abstract action words was verified. In order to understand how the model responded to the variation of the stimuli in input and further investigating how internal representations of objects were related to action representations, the performance of the model was evaluated in response to "incompatible condition" tests during which the provided linguistic input was either inconsistent with the objects perceived by the robot or with the actions typically associated to the objects.

- In Chapter 8 the main topics addressed in this thesis are recalled in order to evaluate the results obtained through the presented experimental studies. Conclusions, final remarks and the description of future research directions close this thesis.

Chapter 2

Grounded Cognition and Embodied Language

The study of cognition involves interdisciplinary investigations in subjects ranging from philosophy, psychology, neuroscience, linguistics to artificial intelligence and robotics. The first attempts to understand the mechanisms underlying the functioning of the human mind can be traced back to the Ancient Greeks when philosophers such as Plato and Aristotle tried to explain the nature of human knowledge. Many scholars have claimed that Plato was the first philosopher to define the dichotomy between the mind and the body. Plato considered the mind as the "prisoner" of the body and after death, while the body was thought of decomposing into its original elements, the mind, being immaterial, survived the body. On the other hand Aristotle, member of Plato's Academy, disagreed with his teacher and mentor providing a closer relationship between the mind and the body claiming that the mind is the "form" of the body. During the seventeenth century, the most famous philosophical work of René Descartes, "Meditations on First Philosophy", dealt with the mind-body problem, that in philosophy refers to the study of the relation between mental and physical properties. The investigation of the mind-body problem led to the formulation of the Cartesian Dualism, according to which it is possible to distinguish between matter (i.e. things with measurable properties and spatially extended) and

mind (i.e. the non-physical things that can think), the latter assumed to be the centralised control system of human beings. In contrast to the classical view of cognition (e.g. Cartesian dualism and computationalism) in the twentieth century the embodied theory of mind, which has its roots in Immanuel Kant, arouse; indeed, one of the major goals of the "Critique of Pure Reason" of Kant, was to provide a solution to the mind-body problem. Kant proposed that the mind is no longer separate from the body, but it is a manifestation of it, viewed from a specifically human and rational perspective. The modern formulation of the embodied theory of mind considers intelligent behaviours as emergent processes of the interaction between mind, body and environment [Pfeifer et al., 2007]; the mind controls bodily actions, and in turn, the motor system influences our thinking.

Embodied cognition is currently investigated in many disciplines embracing psychology, linguistics, cognitive science, neuroscience, artificial intelligence and robotics. In the framework of cognitive linguistics, George Lakoff proposed conceptual metaphors and image-schemas as a general mechanism to ground abstract knowledge (e.g. mental representations) in concrete domains (e.g. body structure) [Lakoff and Johnson, 1980]. In neuroscience, the relationship between the body structure, some brain areas (e.g. motor and premotor cortex) and the mind (e.g. emotions) has been investigated. For example, the "motor theory of speech perception" suggests that the perception of spoken words is based on the identification of the vocal tract gestures involved for the production of words, rather than on the identification of the sound patterns that speech generates [Liberman et al., 1967]. Moreover, in robotics and artificial intelligence, insights from neurophysiology and psychology have inspired the design of machine which, endowed with at least some of the desirable properties of biological organisms, such as adaptivity, robustness, versatility and agility, can become increasingly capable to interact in non structures scenarios [Pfeifer et al., 2007]. In turn, advances in cognitive robotics and artificial intelligence can represent a crucial tool in the scientific research on cognitive science and in the study of the human behaviour.

It has been proposed that cognitive theories can be placed on a continuum ranging from a purely disembodied account to a purely embodied one [Wilson, 2002]. This chapter focuses on the embodied view of cognition applied to language for which different theories, both disembodied and embodied, such as symbolic theories (based on amodal symbols), statistical approaches (based on statistical representations), connectionism (based on artificial neural networks), grounded theories (based on modal symbols), etc., have been proposed in literature. In the next sections an overview of the most relevant theories of cognition proposed in literature is given; additional details will be provided on grounded theories of cognition. The chapter also identifies the major milestones in the development of language in humans. Further, this chapter contains a section regarding knowledge representation that tries to clarify what differs between concrete and abstract concepts and it presents neurophysiological and behavioural studies about language processing that support the embodiment of language. This overview will set the scene for the modelling of abstract words in humanoid robots.

2.1 The Classical View of Cognition

According to the classical view of cognition, the mind is considered as a symbol system and cognition relates to symbol manipulation capabilities [Harnad, 1990]; cognition and perception are separate and independent systems that work according to different principles [Barsalou, 1999]. Hence, conceptual representations are non perceptual and unrelated to the body. In this framework, concepts are generated by combining and manipulating *abstract*, *arbitrary* and *amodal* symbols for which their internal structures are unrelated to the perceptual states and actions that produced them [Landauer and Dumais, 1997, Fodor, 1998]. Indeed, according to this approach, the link between the internal symbols and the external referents is arbitrary; hence, such symbols must be implemented outside the brain's sensorymotor system [Gallese and Lakoff, 2005]. Therefore, concepts are represented in

terms of lists of properties, features and statements. In this framework, Fodor proposed that concepts are represented in some "language of thought" [Fodor, 1975] made up of symbols and having the properties of productivity and compositionality among others [Gallese and Lakoff, 2005]; in other words, mental representation and thought take place within a mental language which is a representational system that employs both a combinatorial syntax and a compositional semantic.

In the last decades the symbolic approach to cognition, failing to explain how cognition is related to perception and action, has been heavily criticised and challenged. In order to show that the symbolic approach is incorrect, Searle formulated the "Chinese Room Argument" [Searle, 1980]. According to the symbolic theory of mind, if a symbol-processing machine (e.g. a computer) could pass the Turing test [Turing, 1950] in Chinese, then this machine would understands the meaning of Chinese symbols in the same way that an English-speaking person understands the meaning of English symbols [Harnad, 1990]. Searle attempted to show that a symbol-processing machine can never be properly described as "*having a mind*" or "*understanding*", regardless of how intelligently it may seem to behave. Indeed, in the traditional computational models (symbolic approach) that deal with language learning tasks, symbols are self-referential entities that require the interpretation of an external experimenter to identify the referential meaning of the lexical items. This is the well known "Symbol Grounding Problem" [Harnad, 1990], which is related to the matter of "*how symbols get their meanings*" and "*how symbols are connected to the things they refer to*". The problem, as Harnad said, is analogous to trying to learn Chinese by using a Chinese/Chinese dictionary alone [Harnad, 1990]; this attempt would lead to a "merry-go-round", passing endlessly from one meaningless symbol to another, without acquiring the meaning of any of such symbols. Harnad and colleagues proposed the identification of a "grounding kernel" of concrete words that are learned earlier, from direct experience; the meanings of the rest of the words in the dictionary can be learned from definition alone, by combining the core words into subject/predicate propositions with truth values. In other words,

higher-order symbols, referring to more abstract knowledge, can be composed of grounded elementary symbols [Blondin-Massé et al., 2010].

In the last decades, the connectionist approach and artificial neural networks have received lots of attention as computational learning mechanism for natural language processing [Wermter et al., 1996]. According to the connectionist approach, neural network architectures (e.g. feed-forward and recurrent) permit the modelling of a number of functionalities related to language learning tasks through training according to specific learning rules. The connectionist approach of artificial neural networks, differently from the pure symbolic approaches, models mental phenomena as an emergent process of interconnected networks. Indeed, according to connectionism, cognition is not just symbol manipulation but it requires dynamic patterns of activity in a multi-layered network of interconnected units [McClelland et al., 1986]. Such dynamic patterns change according to the inputs and the applied learning rule [Harnad, 1990]. Recently, the combination and integration of connectionist networks with statistical and symbolic representations has led to an important field in natural language processing based on neural architectures. Indeed, symbolic approaches seem more suitable for formal and language-like tasks, while the connectionist ones at sensorimotor learning tasks [Harnad, 1990]. Most recently, the integration of connectionist networks with symbolic representations embodied in robotics platforms and combined with robotics methodologies has led to connectionist embodied models, in which cognitive processes are emergent from the sensorimotor interaction of an artificial agent with the environment. This approach appears to be promising in order to overcome the limitations of symbolic and pure connectionist models in the development of language learning systems.

2.2 Theories of Grounded Cognition

Grounded theories of cognition assume that knowledge, and cognitive processes in general, are grounded in perception and action systems; knowledge is represented

with modal symbols related to the perceptual states that produce them. Differently from symbolic approaches, grounded theories claim that perception and cognition are not independent systems, but they share a common representational system [Barsalou, 1999]. Some of the main grounded cognition theories proposed in the literature, which arose as reaction to standard theories of cognition (i.e. amodal symbol systems), are presented below:

- **Cognitive Linguistic Theories** deny the presence of a separate and autonomous module in the brain ("*language module*") responsible for language acquisition and refuse the separation of linguistic capabilities from the rest of cognition. A number of cognitive linguists have investigated the ways in which human beings perceive, categorise and conceptualise the world. The results of these investigations have suggested that human beings use basic bodily understanding of places, movement, forces, paths, objects and containers as "metaphors" (e.g. also known as image-schemas) for life, love, mathematics and all other abstract concepts [Lakoff and Johnson, 1980, Eynon, 2002]. In other words, the conceptualization of abstract entities requires the recruitment of the sensorimotor knowledge involved in the metaphors (or image-schemas) used for the grounding of such entities; that is, abstract concepts are grounded metaphorically in embodied and situated knowledge. Cognitive linguistics suggests that, without such "metaphors", there would be no abstract thought [Lakoff and Johnson, 1980, Eynon, 2002].

- **Cognitive Simulation Theories** focus on the role of *modal simulation*, *situated action* and *bodily states* in the grounding of cognitive processes [Barsalou, 1999]. According to Barsalou's Perceptual Symbol Systems theory (PSS), symbols are modal, sensorimotor, proprioceptive, and introspective and related to the perceptual states that produce them [Barsalou, 1999]; that is, symbols activate motor and sensory information (e.g. vision, audition, touch, etc.) tightly linked to the interaction with the world. When the body interacts with the environment (e.g. *sitting on a chair*), the brain captures and stores in memory

the neural activation patterns present during experience with entities (e.g. *how a chair looks like*, *the action of sitting*, etc.). Later, these perceptual symbols, when semantically related, are combined to form a simulator (concept). When knowledge is needed to represent a category (e.g. *chair*), these neural activation patterns are reactivated to simulate the concept [Barsalou, 2008]. The combination of simulators enables the formation of new concepts. In the framework of the cognitive simulation theories, another important approach has been proposed by Glenberg and colleagues [Glenberg, 1997]. Glenberg proposed that the meaning of a situation depends on a set of stimuli available for acting on objects (i.e. affordances) tuned on the individual's personal experience and according to the goal to be pursued [Glenberg, 1997]. For example, if the goal of a person is to change a light bulb, the meaning of the situation will arise from affordances related to a light bulb (e.g. *holding it in the hand*) "meshed" with the affordances of a chair (e.g. *it supports for reaching the bulb*) related to the goal to be pursued.

- **Social Simulation Theories** propose that the understanding of mental states in other people requires simulations of our own mind (e.g. to understand how someone else feels when disgusted [Goldman, 2006], we simulate how we feel when disgusted) and typically it requires the activation of mirror neuron circuits (i.e. neurons that have the property to fire both when an individual acts and when the individual observes another individual performing the same action) [Rizzolatti et al., 1996a]. From this perspective, simulation provides a general mechanism for establishing empathy (i.e. in a minimal sense empathy might simply mean the occurrence of a mirroring process) [Barsalou, 2008].

This thesis revolves around the embodied view of cognition applied to language. Indeed, cognition deals with the understanding of the human mind and the representation of knowledge and conceptualization. Hence, the embodied view of cognition affects language, as it makes use of concepts.

In the framework of the cognitive simulation theories, evidence in support of the role of simulation in language comprehension has been provided by behavioural and neurophysiological studies [Kaschak and Glenberg, 2000, Glenberg and Kaschak, 2002, Glenberg et al., 2008]. According to the embodied theory of meanings, known as the "Indexical Hypothesis", sentences become meaningful through grounding their interpretation in affordances [Kaschak and Glenberg, 2000]. The acquisition of the meaning of sentences requires three processes: (i) mapping words and phrases to their referents (i.e. perceptual symbols); (ii) deriving affordances from these referents; (iii) meshing these affordances under the guidance of syntax. Affordances [Gibson, 1977] permit the finding of the causal relation between objects, actions and effects, while grammar constraints the interpretation of sentences and directs the combination of affordances. The "Indexical Hypothesis" is supported by the phenomenon associated to language comprehension, known as the Action-sentence Compatibility Effect (ACE) presented in Glenberg and Kaschak [2002]. In the behavioural study that led to the observation of the ACE (i.e. modulation of the motor system during the comprehension of language), participants were faster in responding by pressing a button, when the direction of the arm movement and the action described by the processed word were compatible (e.g. making a movement with the arm away from the body to press the button and processing the sentence "*close the drawer*"). When a sentence implied an action towards the body (e.g., "*open the drawer*"), the participants were slower in responding moving the arm in the opposite direction. Moreover, the modulation of the motor system has been observed in neuroimaging and neurophysiological studies during the comprehension of concrete and abstract language [Glenberg et al., 2008]. The results of this neuroimaging and neurophysiological experiments support an embodied theory of meanings that relates the meaning of sentences to human action.

All the presented theories of grounded cognition are based on embodied concepts, which are modal and grounded in sensorimotor experiences. An important consequence of this embodied view of cognition, concerns language, as it makes use

of concepts. According to *embodied theories* of language, concepts are generated by modal symbols grounded in perception and action [Borghi et al., 2011].

2.3 The Development of Intelligence

In the developmental psychology literature, one of the most influential theories on the origins of intelligence in children has been proposed in [Piaget and Cook, 1952], where it has been argued that intelligence is rooted in sensorimotor knowledge. This is a general and comprehensive view of cognitive development, rather than a theory on specific cognitive capabilities. Piaget's theory of cognitive development identifies the cognitive processes that children use to construct their knowledge of the world in: (i) Schema, (ii) Assimilation, (iii) Accommodation, (iv) Organization and (v) equilibration. **Schemas**, which are representations that organize knowledge, are created in the brain while children seek to construct an understanding of the world (e.g. classification of objects by size, shape and colour). Schemas constitute the building blocks of intelligence, and they become more numerous, abstract and sophisticated during development. **Assimilation** is the process of integrating new perceptual and conceptual materials into an existing schema to understand new situations. **Accommodation** is the process that enables the creation of new schemas in case the existing ones are not suitable to capture experiences and situations. **Organization** is the process used by children to organize their experiences by grouping isolated behaviours and thoughts into a higher-order system. **Equilibration** represents the state of balance between assimilation and accommodation. Furthermore, Piaget's theory identifies four stages of cognitive development, during which children develop increasingly powerful and sophisticated cognitive skills, which are: (i) sensorimotor stage (birth-2 yrs), (ii) preoperational stage (2-7 yrs), (iii) concrete operational stage (7-11 yrs) and (iv) formal operational stage ($>$ 11 yrs). During the sensorimotor stage infants construct an understanding of the world by coordinating sensory experience (e.g. seeing, hearing) with motor actions (reaching, touching).

During this stage infants progress from reflexive and instinctual action present at the birth, to the beginning of dealing with the problem of symbolic capabilities towards the end of this stage. During the preoperational stage children develop the ability to represent objects and events using words and images. At this stage children also develop intuitive thought that allows them to begin using primitive reasoning. During the concrete operational stage children develop the ability to think logically about concrete problems and objects. During the formal operational stage, through the development of abstract and logical thought, children develop the ability to solve abstract problems.

Piaget's theory is a general view of cognitive development that remains one of the most influential hypotheses in child psychology and that can provide useful insights for the implementation of cognitive processes in artificial systems.

2.4 Developmental Stages of Language Acquisition in Humans

The acquisition of word meanings is a central topic in cognitive science. Indeed, amongst the various cognitive capabilities (e.g. memory, attention, perception, action, problem solving, intuition, mental imagery, etc.), language represents one of the most powerful tools available to human beings for communicating and exchanging information, ideas, thoughts and feelings with others through speech, signs, text and so on. A fundamental distinction among languages can be made in terms of the type of linguistic representation involved (e.g. auditory for speech, motoric gestures for sing language, tactile for languages such as Braille, etc.). Language through speech is one of the most characteristic abilities of the human species. The study of language (e.g. acquisition, comprehension, production, etc.) is central to many disciplines ranging from psychology and psycholinguistics (i.e. interactions of language with the human mind) to neuroscience and neurolinguistics (i.e. brain changes during language use) and sociology and sociolinguistics (i.e. relation between social beha-

viours and language) [Sternberg, 2009]. Most recently, the study of language has played a crucial role even in the field of computational linguistics and developmental robotics for the creation of computational models of natural language acquisition.

Language skills comprise verbal comprehension and production; language comprehension involves the ability to understand written and spoken linguistic inputs, while language production refers to the ability to produce linguistic outputs. Many scholars have agreed on the definition of some distinctive properties of language (i.e. communicative, arbitrarily symbolic, regularly structured, structured at multiple levels, generative and dynamic) [Sternberg, 2009]. Language, being regularly structured, that is, only particular sequences of words have meaning and different sequences yield different meanings, can be analysed at different levels: (i) phonology to analyse speech sounds, (ii) morphology and lexicon (i.e. the repertoire of morphemes in a given language) to study the structure of words, (iii) syntax for the study of the rules used to put words together and form meaningful sentences, (iv) semantics to study the meaning in language and (v) pragmatics to go behind the literal meaning of language. A central question in the study of language is how these different aspects of linguistic knowledge (also referred as modules) are organized and processed. Psycholinguistic studies on language modularity have suggested that there is a close interaction between these modules (e.g. phonology, morphology, syntax, semantics and pragmatics) during both the acquisition and the processing of language. Furthermore, language being generative, through the usage of syntactic rules can enable the creation of an unlimited number of new utterances.

In the research field of language acquisition, different theoretical stances on the development of language [Barrett, 1999] have been proposed. In particular, it is possible to distinguish between:

- **domain-specific vs domain-general theories**: domain-specific theories assume that cognitive processes are specialised for representing knowledge in specific domains, that is, there are many independent specialised knowledge structures (i.e. modularity of mind); on the other hand, according to the domain-

general theories, language development and cognitive processes in general, are dependent from processes that can handle different knowledge domains, that is, there is one cohesive knowledge structure.

- **nativism vs developmentalism**: while for nativists [Chomsky, 1979] some aspects of the language are innate in humans, according to developmentalists [Tomasello, 2003] linguistic capabilities are gradually acquired during the course of development through the usage of language (**Constructivism**).

Within the nativists, Chomsky proposed that humans have an innate Language Acquisition Device (LAD) that facilitates language acquisition. In the "Principles and Parameters" theory Chomsky [1979] proposed that linguistic knowledge consists of *innate universal principles* (e.g. grammars common to all languages) and *learnable parameters* associated to them (e.g. markers and switches specific for each language). Furthermore, the nativist stance is supported by the "Poverty of Stimulus" argument which states that the linguistic input does not contain sufficient information in itself to permit the induction of grammatical categories that thus must be innate [Barrett, 1999]. Contrary to the assumptions of the nativist linguistic theories, constructivist approaches to child language acquisition support the view that there is no need to assume the existence of innate language knowledge; for example, the "usage-based theory" of language acquisition makes the fundamental claim that language structure emerges from language use [Tomasello, 2009]. Children are active constructors of their own language system through implicit observation and learning of statistical regularities and logical relationships between the meaning of words and the words used.

The acquisition of language is complex because it involves different cognitive capabilities working together [Bloom, 2002]. Indeed, several cognitive capabilities can help children to construct a linguistic system from the received inputs. The ability to speak language develops over time. The most significant events in language development are concentrated during the first years of life of a child, when the brain matures and develops all its functionalities (e.g. creative thought, problem

solving, attention, abstract thinking, muscle movements, coordinated movements, smell, visual functions, language, reading, tactile sensation, sensory comprehension, etc.). The main developmental milestones for language acquisition can be summarized as follow [Cangelosi and Schlesinger, 2014]:

First 6 months Cooing:

Infants start to produce vowel sounds.

6-9 months Canonical babbling:

Infants start to produce phonemes (e.g. "bababa", "mamama"). This milestone comprises the production of distinct phonemes that characterize the primary language of the infant.

10-12 months Intentional communication, gestures:

Children start to show pre-linguistic skills (e.g. intentional communication and cooperation) by producing communicative gestures (e.g. pointing) and iconic gestures (e.g. a throwing motion to indicate a ball).

12 months Single words (holophrases), Word-gesture combinations:

Children acquire the capability to produce the first single words typically used to name or request objects, and to indicate their own actions or desired actions (e.g. the word "milk" can refer to the milk, to the act of spilling it, drinking it, etc.) [Tomasello and Brooks, 1999]. These kind of expressions are referred as "holophrases" that are single linguistic symbols functioning as a whole utterance.

18 months Two-word combinations, 50+ word lexicon size (vocabulary spurt):

After the first words are learned, a rapid increase in the child vocabulary occurs (i.e. "vocabulary spurt"). The increase in the child lexicon repertoire leads to the production of two-word utterances; this is when it begins the first understanding of syntax. However, before the capability to produce two-word combinations is fully developed, children go through a hybrid word/gesture

stage when they combine one gesture with a word to express combinations of meanings.

24 months Increasingly longer multiple-word sentences, Verb islands:

Children start to develop more complex syntactic competences. One of the most influential constructivist accounts of early grammatical development is provided by Tomasello's verb island hypothesis [Tomasello, 1992]; according to this hypothesis, children learn verb-specific constructions (e.g. verb + noun and the noun depends on the specific verb) and the level of complexity of different verb islands are due to usage-based experience. For example, for some verbs children might be able to use simple syntactic combinations of the verb with different nouns, while for other verbs children might have a richer syntactic use.

After 36 months Adult-like grammatical constructions, Narrative skills:

Children gradually develop adult-like syntactic constructions (e.g. Simple Transitives, Locatives, and Datives) [Tomasello and Brooks, 1999]. The acquisition of new syntactic skills leads to the development of more complex syntactic-morphologic constructions, more abstract and generalized grammatical categories, up to the formation of formal linguistic categories such as word classes.

An infant starting to learn a language is subject to the stream of perceptual-cognitive information about the world around him (i.e. the child starts to perceive entities through his/her senses) and the stream of spoken language (i.e. the child hears the sound of words and starts to associates a word to an identified physical entity) [Gentner, 1982]. In child psychology there are different studies that support the hypothesis that concrete words precede the acquisition of abstract words [Caramelli et al., 2004, Schwanenflugel, 1991]; considering that children learn through the sensorimotor interaction with the physical world, they first acquire concrete knowledge related to objects and situations and subsequently they learn more ab-

stract concepts. Indeed, studies conducted on children's early vocabulary acquisition have shown that, when children learn to speak, they first learn concrete nouns (e.g. object's names) and then abstract ones (e.g. verbs) [Gentner, 1982]. Gentner, through the formulation of the "Natural Partitioning Hypothesis" has argued that the linguistic distinction between nouns and verbs is due to the different perceptual-conceptual distinction between concrete and more abstract concepts [Gentner, 1982]. While concrete terms (e.g. nouns) refer to tangible entities (i.e. naturally individuated referents) characterized from an evident and direct mapping to the perceptual world and high imagery, more general and abstract words (e.g verbs) refer to intangible entities characterized from wicker perceptual constraints with the real world (i.e. verbs are linguistically more variable because they can refer to many events, situations and bodily states) [Gentner, 1982]. Hence, during the process of learning word meanings, the mapping of perceptual information into the linguistic domain is faster for concrete concepts than for abstract ones. Other studies have suggested that the development of abstract noun definitions follows the development of the concrete ones [McGhee-Bidlack et al., 1991]. Indeed, it has been shown that, while preschool children use functional responses (e.g. "a chair is to sit on") during the development of noun definitions, older children start to use indefinite place-holders (e.g. "a chair is something to sit on") and superordinate classes (e.g. "an apple is a piece of fruit") [McGhee-Bidlack et al., 1991]. Nevertheless, there are studies in which it has been proposed that the development of biological thought (i.e. distinction between inside animals and machines) might proceed from abstract to concrete instead [Simons and Keil, 1995]. Studies on children's expectations for what could be inside animals (i.e. animates) and machines (i.e. inanimates) have shown that children's expectations proceed from abstract to concrete [Simons and Keil, 1995]; indeed, children might have abstract expectations about the internal operating mechanisms of animals without concrete knowledge associated to them (concrete knowledge develops later). A special case of study is that of abstract social words such as "hi", "bye", "no" that have been found in the earliest production

vocabularies of toddlers [Tardif et al., 2008]. However, the study of such kind of abstract words is out of the scope of the work proposed in this thesis.

2.5 Abstract Words and Conceptual Knowledge

Herein the major theories proposed on the learning and representation of categories/concepts are briefly introduced [Kalkan et al.]:

- **Rule-based Theory**: Members of a category share common (perceptual) properties (e.g. colour, shape, etc.), and the membership for a category is based on satisfying established rules that permit to verify the common properties of the category. Following this approach categories have strict boundaries (i.e an item is either a member or not a member of a category) [Bruner and Austin, 1986].

- **Prototype-based Theory**: Categories are represented by "prototype" stimuli, which are used for judging the membership of other items. This approach assumes a more continuous way of categorization and less strict boundaries among categories [Rosch, 1973].

- **Exemplar-based Theory**: Concepts are represented by the exemplars of the categories stored in the memory. A new item is classified as a member of a category if it is similar to one of the stored exemplars in that category [Nosofsky et al., 1992].

As stated in Mervis and Rosch [1981], p.89:

> *A category exists whenever two or more distinguishable objects or events are treated equally. This equivalent treatment may take any number of forms, such as labelling distinct objects or events with the same name, or performing the same action on different objects.*

Abstract words are used in daily conversations among people to describe and explain events and situations that occur in their social and physical environment.

Nevertheless, the scientific study of concepts so far, has mainly focused on concrete concepts; hence very little is known about the development of abstract categories. Many scholars have suggested that the distinction between concrete and abstract words is not a dichotomy [Wiemer-Hastings et al., 2001]; that is, it is not possible to define a "clear cut" between words classified as concrete or abstract. There is instead a continuum, according to which all entities can be varied in their level of concreteness. For example, words that refer to a social role (e.g. "*physician*") might be classified more abstract than words that refer to a single object (e.g. "*book*") but less abstract than purely definitional words (e.g. "*democracy*") [Wiemer-Hastings et al., 2001, Borghi et al., 2011]. Furthermore, concrete words such "push" and "give" can be differentiated in their level of concreteness and motor modality; that is, a word like "push" is uniquely linked with the action of pushing by using the hand, while "give" implies multiple motor instances of the process of passing an object by using one hand, two hands, the mouth etc. [Cangelosi and Schlesinger, 2014]. Moreover, in [Altarriba et al., 1999] it has been proposed that words which refer to emotions should be categorized in a group of entities distinct from concrete and abstract words. This proposal was motivated by the fact that concrete, abstract and emotion words received different ratings in term of concreteness, imageability and context availability [Kousta et al., 2011]. However, concrete and abstract words can be differentiated according to the following factors:

- **Perceivability**: Abstract words, referring to entities that are distant from immediate perception, represent everything that is not physically defined nor spatially constrained (e.g. "*truth*", "*democracy*", "*happiness*", "*justice*"). These kind of words, contrary to concrete terms, do not have physical referents that can be seen or touched and it is not possible to interact with them. When concepts become more detached from physical entities and more associated with mental events, they become increasingly abstract [Barsalou, 1999, Paivio et al., 1968, Wiemer-Hastings et al., 2001].

- **Imageability and context availability**: According to the dual-coding the-

ory [Paivio et al., 1968] concrete and abstract concepts representations require a different involvement of memory; concrete concepts, being represented by activating a verbal and non-verbal system, require a major involvement of memory, while in case of abstract concepts, being represented in the verbal system only, the involvement of memory is inferior [Barsalou et al., 2008]. Further, abstract concepts evoke less imagery than concrete concepts [Wiemer-Hastings and Xu, 2005]. According to the context-availability theory concrete words activate a broader contextual verbal support than abstract words. This can be one of the reasons why abstract concepts acquisition is more complex than concrete concepts acquisition.

- **Hierarchical Categorization**: Conceptual knowledge can be organized in categories hierarchically structured. Traditionally, three levels of categorization have been proposed, namely the **subordinate**, the **basic** and the **superordinate** levels (e.g. "rocking chair/chair/furniture"). The subordinate level categories are characterized from a low degree of generality and from clearly identifiable, detailed and specific features (e.g. "rocking chair"). Subordinate level categories are included under basic level categories. The most relevant conceptual information relating to a category is stored at the basic level (e.g. "chair"). Basic level categories are included under superordinate level categories (e.g. "furniture") which are characterized from a high degree of generality and allow to store general information. As suggested in [Borghi et al., 2005], in the hierarchical organization of words categories, basic and subordinate concepts (e.g. "chair" and "rocking chair") refers to single concrete entities and elicit perceptual information; hence they can be considered more concrete concepts than superordinate concepts (e.g. "furniture") that can be associated to different intangible entities and elicit abstract information. In line with this position, many accounts of concepts formation have suggested that the first word categories acquired must be instance-based. For example, before a child can understand the meaning of the superordinate category "furniture",

he/she might start to learn the meaning of detailed, image-like representations of individual items, and then progress to increasingly abstract representations that embody non-perceptual information.

- **Mode of Acquisition (MOA)**: The difference between concrete and abstract concepts can be related to their mode of acquisition (MOA) which can be perceptual, linguistic, or by combining perceptual and linguistic information [Wauters et al., 2003]. In experiments with elementary school children it has been shown that MOA ratings gradually change with the school age progression, shifting from mainly perceptually acquired word meanings to mainly linguistically acquired concepts.

Abstract concepts pose a classical challenge for both symbolic and grounded theories of cognition. Scholars working in the field of classical theories of cognition (i.e. amodal symbol systems) have argued that grounded cognition approaches for knowledge conceptualization, using only sensorimotor representations of the external world, cannot represent abstract concepts that are not grounded externally. However, according to grounded cognition approaches, conceptual contents can be derived from the perception of internal states as well (introspection) [Barsalou, 1999, Barsalou and Wiemer-Hastings, 2005]. For example, in the study conducted by Barsalou and Wiemer-Hastings [2005] participants were asked to generate features for words varying in concreteness (e.g. truth, freedom, invention, bird, car, sofa, cooking, farming, carpeting). The results of this study have shown that abstract concepts focus on introspection. According to Glenberg and Kaschak [2002], abstract concepts contain motor information and hence it is possible to treat the problem of obtaining a representation associated to them by using modal systems. Given the current debate in the field, and the complexity of the matter, nowadays the task of representing abstract concepts through sensorimotor experience has been proved to be an extremely complex task.

Different theories proposed in psychology state that embodiment plays an important role even in representing abstract concepts. One of the main theories about

the embodiment of abstract language revolves around the concept of "metaphor". According to this approach, there are image-schemas derived from sensorimotor experience that can be transferred to experience which is not truly sensorimotor in nature [Lakoff and Johnson, 1980]. Abstract concepts can be grounded in concrete domains through "metaphors" [Lakoff and Johnson, 1980]. According to this hypothesis, human beings have an extensive knowledge about their bodies (e.g. eating) and situations (e.g. verticality) and they use such knowledge to metaphorically ground abstract concepts; for example, *love* can be understood as eating (e.g. "being consumed by a lover") and affective experience can be understood as verticality (e.g. "happy is up, sad is down") [Barsalou, 2008]. Hence, abstract concepts are represented through a metaphoric mapping. However, in order to fully represent abstract concepts, metaphors alone might be not sufficient and more features might be needed to distinguish among abstract concepts; furthermore the role of metaphors in the development of abstract words it has not been clarified yet.

Other studies have proposed that some abstract concepts arise from simulation processes of internal and external states [Barsalou, 1999]. In particular, abstract concepts require simulation that can capture complex multi-modal simulations of temporally extended events, with simulation of introspections being central [Barsalou, 1999]; introspection permits to access subjective experiences linked to abstract concepts [Wiemer-Hastings et al., 2001]. Indeed, considering that abstract concepts contain more information about introspection and events [Wiemer-Hastings et al., 2001], simulators for abstract words develop to represent categories of internal experience [Barsalou, 2009]. Hence, according to this approach, abstract concepts, differently from concrete ones, require the activation of situations and introspections.

Other scholars have suggested that sentences, including both concrete and abstract words, are understood by creating a simulation of the actions that underlie them [Glenberg and Kaschak, 2002]. In particular, through behavioural and neurophysiological studies it has been shown that even the comprehension of abstract

words activates the motor system [Glenberg et al., 2008]. Hence, according to this approach, abstract concepts, similarly to concrete ones, can be grounded in perception and action.

2.5.1 Situated Conceptualization

Concepts are the elementary units of reason and linguistic meaning [Gallese and Lakoff, 2005]. Further, it has been proposed that a concept is knowledge about a particular category; for example, the concept "*birds*" is represented by the knowledge about the category "*birds*" that represents the bodies, behaviours and origins of the respective entities [Barsalou et al., 2003]. In case of abstract concepts, such knowledge is more detached from physical experience [Borghi et al., 2011]. There are different research studies in which it has been shown that *situations* and *situated action* play an important role in the conceptual representation of both concrete and abstract language [Barsalou and Wiemer-Hastings, 2005, Schwanenflugel, 1991]. According to these findings, in order to understand the meaning of "*chair*" for example, it is necessary to acquire information not only about the physical properties of the object, but also about the usage of the object in relevant situations. Even abstract concepts appear to depend heavily on situations and situated action [Schwanenflugel, 1991]. The processing of abstract concepts is facilitated when a background situation contextualizes it [Barsalou and Wiemer-Hastings, 2005]. Nevertheless, situations in which abstract concepts occur are retrieved less easily than situations in which concrete concepts occur, because abstract concepts can be associated with a larger variety of situations. As a matter of fact, while for conceptual representation of concrete words there is a circumscribed region in which the situation occurs and the focus is on situations in which object are presented and used, for abstract concepts the focus is on events and introspection and hence their content is distributed across several situations. Nevertheless, according to studies reported in [Barsalou, 1999], it seems possible to simulate introspective experience and then there is no reason for believing that abstract concepts can not be simulated.

2.5.2 Multimodal Theories of Knowledge Representation

The traditional symbolic and embodied theories of conceptual representation proposed in literature, rely on a single kind of representations (amodal for symbolic theories and modal for embodied approaches). Nevertheless, some of the grounded theories of cognition proposed in literature for knowledge representation rely on multiple systems for *perception* (e.g vision and audition), *action* (e.g. movement and proprioception) and *introspection* (e.g. mental states) [Barsalou, 2008]. Indeed, recent findings support the view that conceptual processing rely on multiple representational systems for which linguistic and sensorimotor information are both activated [Louwerse and Jeuniaux, 2010]. These results are in line with the most relevant theories regarding concrete and abstract concepts knowledge representation that are the "dual-coding" theory [Paivio et al., 1968] and the "context-availability" theory [Schwanenflugel, 1991]. According to the dual-coding theory, while concrete words are represented by activating a verbal (i.e. linguistic) and non-verbal system (i.e. imagistic system), abstract words are represented in the verbal system only [Paivio et al., 1968]. For the "context-availability" theory, concrete words, differently from the abstract ones, activate a broader contextual verbal support and they have stronger semantic relations with the context represented by other words that make their processing faster than the processing of abstract words. Contrary to the dual-coding theory, the context-availability theory does not assume the access to a distinct system (i.e. non-verbal system) and both concrete and abstract concepts are represented in a single verbal system.

Recently, along the same line of the dual-coding theory (i.e. knowledge represented by multiple systems), the Language and Situated Simulation (LASS) theory [Barsalou et al., 2008] and the Words As Tools (WAT) theory [Borghi and Cimatti, 2009] have been proposed. According to the LASS theory, both the sensorimotor and linguistic system are activated during language processing. Furthermore, for the processing of abstract concepts it has been proposed that linguistic information might be more relevant than for concrete concepts [Barsalou et al., 2008]. According

to the WAT theory [Borghi and Cimatti, 2009], words represent tools that permit to act in the social world. In facing the challenge of abstract word representation, the authors of the WAT proposed the existence of two simultaneous cognitive source for word meanings, one of which is individual and related to embodied individual experience, and the second one, which is a socially embodied one [Borghi and Cimatti, 2009]. While concrete word meanings can be grounded through embodied individual experience, in case of abstract words the knowledge is embodied in the use of the social word [Borghi and Cimatti, 2009]. Further, abstract words acquisition often implies complex linguistic explanations and repetitions; on the contrary, the acquisition of concrete words appears much easier and often occurs within a single episode of hearing a word spoken and perceiving the corresponding entity. Furthermore, while concrete words evoke more sensorimotor information, abstract words elicit more verbal linguistic information. Most recently, it has been proposed that concrete and abstract concepts contain different types of information that is, experiential information (i.e. sensory, motor and affective) and linguistic information (i.e. linguistic co-occurrence); sensory-motor information is more preponderant for concrete concepts, while affective information plays a greater role for abstract concepts [Kousta et al., 2011]. Following this proposal, abstract words have a processing advantage over concrete words, considering that abstract words tend to be more emotionally loaded [Kousta et al., 2011]. The novelty of this approach is that emotion is considered to be another type of experiential information playing an important role in representing abstract words.

Taken together, these studies suggest that the meaning of words is grounded by activating the multi-modal experience related to the conceptual referent of words and, linguistic experience plays an important role in shaping such conceptual knowledge. In the experimental studies proposed in this dissertation, abstract words, which are not directly linked to the physical and perceptual world, are grounded by reusing the sensorimotor knowledge related to the conceptual referent of such words, which is shaped through language. The endeavour of the proposed studies

is to ground abstract words in the sensorimotor system in an indirect way. The arrangement of sensorimotor and perceptual knowledge leads to the grounding of higher order concepts [Harnad, 1990]. In particular, the idea exploited in these studies is that some concepts can be grounded via direct sensorimotor experience and identified through linguistic labels. Such labels can be used to combine perceptual symbols and form new categories which cannot be learned via direct sensorimotor experience (i.e. symbolic instructions permit to combine perceptual symbols) [Harnad, 2010]. As argued in [Barsalou, 1999, Glenberg and Kaschak, 2002] conceptualization can guide action to produce new knowledge. In the studies proposed in this dissertation, conceptualization is driven via linguistic instructions.

2.6 The Neural Basis of Language Processing

Many scholars have suggested that the evolution of the neural basis of human language and its properties like speech, syntax and lexicon, derived from Darwinian mechanisms [Lieberman, 2002]. An interesting hypothesis about the role that mirror neurons could have played in language evolution has been formulated by Rizzolatti and Arbib [Rizzolatti and Arbib, 1998]. They suggested that the language evolved from the capability of recognizing actions made by others, that is the action-recognition system has constituted the basis for language development. The progressive evolution of the mirror system (responsible for action recognition in others) from ancestors to humans led to the evolution of language and communication (i.e. from sign language to speech).

Investigations on language processing in neuroscience are based on several techniques that permit the measurement of brain activity while processing linguistic stimuli. One of the most common approach consists in measuring event-related potentials (ERP) by using elettroencephalography (EEG) when a stimulus is presented to subjects. Other techniques used for measuring the brain activity in response to stimuli are the functional Magnetic Resonance Imaging (fMRI) and the Positron

Emission Tomography (PET) that analyse images of changing blood flow in the brain associated with neural activity. Transcranical Magnetic Stimulation (TMS) is commonly used for stimulating neurons in a specific part of the brain and then measuring muscles activities.

Traditionally, language processing has been considered to be located in the Broca's and Wernicke's brain areas. Recently, deficiencies of this traditional view have been noticed through studies with patients affected by aphasia (i.e. deficits in the comprehension and formulation of language caused by dysfunction in specific brain regions) and the permanent loss of language. Neuropsychological studies in brain-lesioned patients and brain imaging studies have shown that language may involve various cortical areas according on the type of language-related semantic information being processed [Pulvermüller et al., 2001]. Further, clinical evidence has shown that the permanent loss of language does not occur even when Broca's or Wernicke's areas have been destroyed [Lieberman, 2002]. The results of these studies have led to the intuition that the neural basis of human linguistic ability are complex and they involve other cortical structures, other than Broca's and Wernicke's areas [Lieberman, 2002]. Such structures form part of the neural circuits implicated in the lexicon, speech, and syntax development [Lieberman, 2002]. Hence, the neural system for language is widely distributed in the brain.

Moreover, until recently the cortical systems for language and action control have been considered to be organized in modules independent from each other and characterized from different cortical bases; that is, the motor and premotor cortex control action while the Perisylvian network (i.e. Broca's and Wernicke's areas) is responsible for language. This view has been supported by studies on some neurological disease that affected specific language or action functions while maintaining normal performances in other cognitive domains and by many brain imaging studies and connectionists models. Contrary to this view, recent studies support the hypothesis that information about language and action might interact with each other. Neurophysiological studies have shown that during action verb processing

(i.e. verbs that denote actions) different brain areas are activated depending on the effector (e.g. arm/hand, leg/foot, mouth) involved in the processed action verb [Pulvermüller et al., 2001]. Indeed, it has been shown that the motor cortex has a somatotopic organization for legs, arms and mouth articulators [Pulvermüller et al., 2001]; that is, there are different areas in the motor cortex specifically associated with the control of the movement of legs, arms and mouth. For action words that are semantically related to the action described in the word, the neural representation of the action in the motor cortex has to also include the semantic neurons of the corresponding word. These findings support the existence of different distributed networks for words that describe actions relate to legs, arms and mouth articulators (i.e. semantic somatotopy model). This leads to the important consequence that the perception of action words, like "pick" for example, activates the same cortical area involved for the control and the execution of the pick action [Pulvermüller et al., 2001].

Theories of associative learning (i.e. the process by which an association between two stimuli or a behaviour and a stimulus is learned) have suggested that the representations of words, frequently co-presented with non-linguistic stimuli (e.g. vision or audition) include the co-activation of neurons into their representations so that whenever such words are perceived, the mental images can be immediately aroused [Pulvermüller et al., 2001]. For example, when a child learns to perform an action and simultaneously hears from the caregiver the corresponding word that describes the action, the link in the cortex between that word and the corresponding motor area becomes stronger. According to theories of associative learning, the representation of action words referring to movements with a particular part of the body (e.g. leg, arm, or face) should include neurons involved in programming the respective actions. Hence, the neural representations of such words are distributed over language areas and additional areas related to the word's meaning [Pulvermüller et al., 2001]. This proposal has received support from several experiments. In [Perani et al., 1999] it has been found that different word classes (i.e. nouns and verbs) led to widely

distributed signs of activity. Positron Emission Tomography (PET) studies have been used to measure cerebral activities during tasks requiring to read nouns and verbs (concrete and abstract) for lexical decision [Perani et al., 1999]. The results of these studies have suggested that verbs and nouns processing requires the activation of different brain areas; the left temporal lobe plays an important role in processing nouns, while the left frontal lobe is involved during verbs processing. Furthermore, according to [Perani et al., 1999] concrete and abstract words activate different brain areas; evidence that the processing of abstract words produces higher activation in the left hemispheric areas of the brain has been provided by Perani et al. [1999]. Additionally, in [Pulvermüller, 1999] it has been found that the processing in isolation of abstract concepts (i.e. concepts not occurring in situations) is localized in the left frontal area of the brain (close to the Broca's area that is responsible for words generation).

Many studies have supported the existence of a link between the mirror neuron system and language processing. Indeed, a prediction of embodied theories of language learning is that when individuals listen to action-related sentences their mirror neuron system is activated. Mirror neurons, originally discovered in the premotor cortex of monkey, the so called F5 area, have the property to fire both when an individual acts and when the individual observes (i.e. visual stimuli) another individual performing the same action [Rizzolatti et al., 1996a]. Many researchers share the idea that the monkey F5 area is the homologue of Broca's area in the human brain. Recent studies have shown that mirror neurons, besides having response properties to visual stimuli, also have acoustic properties. These audio-visual mirror neurons discharge not only when the action is executed or observed, but also when its sound is heard [Buccino et al., 2005]. Transcranial Magnetic Stimulation (TMS) and behavioural studies for understanding whether listening to action-related sentences modulates the activity of the motor system have been carried out [Buccino et al., 2005]. In the TMS experiments, motor evoked potentials (MEPs) from hand and foot muscles have been recorded, while participants were listening to sen-

tences expressing hand/arm action, foot/leg action, and abstract content. Results showed that listening to hand/foot action-related sentences induced a decrease of MEP amplitude recorded from hand/foot muscles. In behavioural studies, participants were asked to answer with the hand or the foot while listening to hand/foot action-related sentences. Coherently with the TMS findings, the behavioural data showed that reaction times were slower when participants responded with the same effector involved in the listened action; the processing of language with a motor content activates the same sectors of the motor system where the involved effector is represented [Buccino et al., 2005]. These results support the involvement of the motor system in the processing of action-related sentences.

In addition to the studies proposed in linguistics and psychology, recent studies in cognitive neuroscience have suggested that conceptual knowledge is embodied and that the sensory-motor system has the right kind of structure to characterise both sensory-motor and more abstract concepts [Gallese and Lakoff, 2005]. An increasing body of evidence has shown that language understanding implies a mental simulation (i.e. imaging) and understanding and imaging use the same neural substrate [Gallese and Lakoff, 2005]. According to the hypothesis formulated in [Gallese and Lakoff, 2005], understanding requires the formation of a mental simulation of action or perception, using many of the same neurons as actually acting or perceiving. Indeed a major finding in neuroscience has suggested that imagining and acting use a shared neural substrate. To understand the meaning of the concept "grasp" for example, one must at least be able to imagine oneself or someone else grasping an object. Gallese and Lakoff [2005] have argued that the same thing may apply to all other action concepts, to object concepts, and to abstract concepts with conceptual content that is metaphorical [Gallese and Lakoff, 2005].

2.7 Combinatoriality of Language and the Motor System

In contrast to other forms of communication, language is a discrete combinatorial system that permits the conveyance of new messages and concepts by combining simple words together [Pinker, 2010]. Indeed, a finite number of words (i.e. lexicon) can be combined and permuted, according to specific structural rules (i.e. grammar), in order to convey new meanings. The acquisition of lexicon and grammar are both necessary to produce new sentences and their related meanings. Nevertheless, in the process of language development, lexicon acquisition (with subsequent generalization and decomposition properties) constitutes an important prerequisite for higher-order grammar learning. Indeed, the acquisition of lexicon and its related meanings precedes the emergence of more abstract syntactic structures which can be obtained through a gradual transition from lexical semantics [Cangelosi et al., 2010]. According to [Fodor and Lepore, 2002], compositionality in language and mind is due to the fact that complex symbols inherit their syntactic and semantic properties from a series of primitive symbols.

Recent evidence has suggested that the human motor system is also hierarchically organized [Arbib et al., 1998, Mussa-Ivaldi and Bizzi, 2000]; that is, low level motor primitives can be integrated and recombined in different sequences in order to generate a rich "*grammar*" of motor behaviours that can enable the execution of novel tasks. Indeed, Mussa-Ivaldi and Bizzi [2000] have suggested that motor learning consists of tuning the activity of a relatively small group of neurons that constitute a "*module*". The combination of such "*modules*" may be a mechanism for producing a vast repertoire of motor behaviours in a simple manner. In other words, more complex human behaviours can be seen as the result of the integration of motor primitives organized in hierarchical structures. Furthermore, it has been proposed that the sensorimotor system has the right kind of structure to characterise both sensorimotor and more abstract concepts [Gallese and Lakoff, 2005].

Collectively these studies suggest that language and the biological motor system are based on hierarchical recursive structures that can be exploited to ground the meaning of language in sensorimotor experience. Indeed, the modular organization of the biological motor system has been shown to be based on hierarchical recursive structures which have linguistic analogues in grammatical/syntactical structures [Cangelosi et al., 2010]. These observations and insights have inspired the development of a similarly organized artificial system that combines low level motor primitives for grounding the meaning of language in action and perception. More complex behaviour and their related meanings can be achieved by integrating different motor primitives together. In this framework language and its combinatorial structure provide a tool for organizing motor primitives and perceptual knowledge in order to bootstrap more complex cognitive behaviour in artificial agents.

Chapter 3

Artificial Intelligence and Language Modelling

Artificial Intelligence (AI) is the branch of computer science, emerged in the mid twentieth century, which deals with the design and implementation of computational models that attempt to simulate the mechanisms underlying cognition. One of the long term goals of AI is to build intelligent machines, up to the human level of intelligence, that can pass the Turing Test [Turing, 1950]. The Turing Test was proposed by Alan Turing in 1950 to deal with the question of whether machines can think. Turing proposed the "Imitation Game", which involved a person, a machine and an interrogator located in separated rooms. The interrogator, asking questions to the person and the machine, at the end of the game according to the received answers has to distinguish the person from the machine. Nevertheless, at present artificial intelligent systems are still far from achieving the human level of intelligence and hence, passing the Turing Test.

AI research can be carried out by employing two different approaches, namely top-down and bottom-up. The top-down approach, in line with symbolic views of cognition, considers the intelligence of a machine as a high-level phenomenon that does not depend on how low-level operations that produce it are implemented. Indeed, this methodology ignores the neurological interconnections that underlie

intelligence, and assumes that a machine must be supplied with an internal representation of the essential features of the world in which it operates. On the contrary, the bottom-up approach explores the aspects of cognition that can be recreated, by employing neural networks for example. Indeed this approach focuses on actions and behaviours that produce intelligence, rather than on representations and functions. Research in AI, especially in the case of the bottom-up methodologies, is tightly linked to embodied cognitive science, considering that finding in cognitive science can inspire better artificial simulations of the human mind. In turn, artificial intelligence can provide more accurate models of the human mind, which can produce interesting predictions that can be tested through experimental and behavioural studies by cognitive scientists. Indeed, cognitive modelling can provide a powerful tool for investigating and understanding how motor behaviours and symbols manipulation capabilities can be integrated to bootstrap higher-level language representations. However, the comparison between results produced by cognitive modelling and neuroscience research requires that cognitive modelling respects (i) **neurobiological constraints**: the model's neural system should be endowed with at least some crucial characteristics of the human neural system, (ii) **embodiment constraints**: the model should be endowed not only with a brain which is similar to that of humans, but also with a sensorimotor system similar, at least in some respects, to a human sensorimotor system and (iii) **behavioural constraints**: the model should reproduce and replicate the behaviours produced during empirical experiments [Caligiore et al., 2009]. According to the Tri-Level Hypothesis, cognitive processes, and hence cognitive models that attempt to reproduce them, can be analysed at three different levels [Marr, 1982]: (i) **computational level** to identify the knowledge computed during the cognitive process, (ii) **algorithmic level** to analyse the mechanisms involved during the computational process and (iii) **implementation level** to simulate the identified algorithms.

Some of the problems addressed by AI are knowledge representation, reasoning, problem solving, planning, learning, natural language processing, motion and

manipulation; and this list is by no means exhaustive. In this chapter current research in artificial intelligence for knowledge representation and natural language processing is described. Some of the approaches proposed in this research field such as symbolic and subsymbolic models (e.g. connectionism), statistical models, and embodied connectionism are described in the next sections of this chapter. A description of the hardware and software architecture of the iCub robotics platform closes the chapter.

3.1 Symbolic Models

The first attempts to create models of language were influenced by early AI techniques and approaches such as symbolic models of knowledge representation and logical reasoning (i.e. deduction, adduction and induction) [Alishahi, 2010]. In the traditional approach to artificial intelligence, informally defined as "Good Old Fashioned AI (GOFAI)", natural language processing, rooted in linguistic analysis of semantics, syntax, pragmatics and context, is based on symbolic computation. In [Russell et al., 1995] it has been proposed that communication via language between a sender and a receiver requires seven component steps. When a Speaker (S) wants to inform the Hearer (H) about the proposition (P) using Words (W), the following seven processes take place:

- Intention: the Speaker (S) decides to communicate a Preposition (P) to the Hearer (H).
- Generation: the Speaker (S) plans how to translate the proposition (P) into an utterance that will enable the Hearer (H) to infer the proper meaning of (P). At the end of this process the Speaker (S) generate the word (W).
- Synthesis: the Speaker (S) produces a string of sounds corresponding to the word (W).

- Perception: the Hearer (H) perceives the sounds corresponding to the word (W) and decodes it into a string (i.e. speech recognition).
- Analysis: the Hearer (H) analyses the input string through three main stages:
  - Syntactic Interpretation (or parsing): to analyse the syntactic structure of the string and build a parse tree for the input string.
  - Semantic Interpretation: to extract the literal meaning of the string.
  - Pragmatic Interpretation: to give the proper meaning to the string according to the context.
- Disambiguation: the Hearer (H) infers that the Speaker (S) wanted to convey the preposition (P). If the Hearer (H) assigns to (P) the interpretation that the Speaker (S) intended to convey, than communication is successful.
- Incorporation: the Hearer (H) decides whether to believe or not to the proposition (P).

According to the linguistic tradition proposed in Chomsky [1979], a symbolic model of language is based on an abstract rule-based grammar which specifies the set of valid sentences. In such models language processing is governed by specific principles and rules (i.e. grammar), and ambiguities are resolved using parse trees (i.e. syntax); these are known as symbolic grammar models. Grammar (i.e. finite set of rules that specifies a language) and syntax (i.e. analysis of grammar) through the proper ordering of words elicit the meaning of sentences. A symbolic model for a fragment of English can represent the linguistic knowledge to be acquired through the following generative grammar:

$$S \rightarrow NP \quad VP | S\,Conjunction\,S$$

$$NP \rightarrow Pronoun \,|\, Name \,|\, Noun \,|\, \cdots$$

$$VP \rightarrow Verb \,|\, VP \quad NP \,|\, \cdots$$

Grammar permits to combine the words of a specified lexicon into phrases (e.g. Sentence = Noun Phrase + Verb Phrase). These distinct parts of sentences are hierarchically related. For example, in the grammar described before, the sentence (S), located at the highest level of such hierarchical organization, is composed of a Noun Phrase (NP) and a Verb Phrase (VP). In this example, words which are part of the lexicon are ungrounded because they require the interpretation of an external user. Indeed, in the framework of symbolic models, knowledge representation consists of discrete and disjoint symbols which are organized in a list structure that is grammatical and combinatorial. Knowledge is represented by manipulating symbols according to specific structural rules (e.g. in the form **IF** (**A** *is true*) **THEN** (**B** *is true*) where **A** and **B** are propositions whose truth or falsity has to be determined), making use of logical techniques such as deduction, induction, expert systems (that include deduction and induction) or other forms of reasoning. Following the symbolic approach, possible ways of representing knowledge are semantic networks, which are models of data representation that include: (i) nodes representing particular concepts or elements of the world, (ii) arcs representing the relationships between the concepts or elements, and (iii) scripts where knowledge is organized by attributes and associated procedures. Symbolic modelling often refers to an explicit formalization of knowledge which is represented in terms of symbols, producing circular definitions much like those found in a dictionary [Harnad, 1990], and their propositional relations.

One of the most prominent symbolic theory of acquisition, induction and representation of knowledge is the Latent Semantic Analysis (LSA). This is a corpus-based statistical method that represents the meaning of words in high dimensional space based on word patterns of co-occurrence with other words [Landauer and Dumais, 1997]. This method analyses the relationships between a set of documents and the terms they contain and it assumes that words that are close in meaning will occur close together in text. To give a flavour of how LSA represents the meaning of words, a small example is presented (Fig.3.1); the first step is to represent words in

a matrix (X) in which each row stands for a unique word and each column stands for the context in which the word is used (Fig.3.1) [Landauer and Dumais, 1997].

Example of text data: Titles of Some Technical Memos

c1: *Human* machine *interface* for ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

$\{X\} =$

	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Figure 3.1: Word by context matrix X taken from [Landauer and Dumais, 1997]

During the next step, LSA applies Singular Value Decomposition (SVD) to the matrix X which is decomposed into the product of three other matrices ($X = WSP$). Next, the original matrix (X) is reconstructed ($\hat{X}$) based on just two dimensions [Landauer and Dumais, 1997]. Each cell of the matrix $\hat{X}$ contains the frequency with which the word of its row occurs in the context denoted by its column [Landauer and Dumais, 1997].

Evidence against the predictions made by the LSA theory has been provided in [Glenberg and Robertson, 2000]; by using sentences with similar LSA values, authors found that participants distinguished sentences depending on the perceptual characteristics of the objects. For example, after presenting the context-setting sentence "*Marissa forgot to bring her pillow on her camping trip*", participants judged more sensible the sentence "*As a substitute for her pillow, she filled up an old sweater with leaves*" than "*As a substitute for her pillow, she filled up an old*

sweater with water", although the words "leaves" and "water" are similarly far from "pillow" in terms of LSA values [Scorolli and Borghi, 2008]. By taking into account the perceptual characteristics of the objects, a pillow made by a sweater filled up with leaves for the participants of the experiment seemed not usual but more sensible and imaginable than a pillow made by a sweater with water. These results contributed to the formulation of the Indexical Hypothesis, according to which '*the meaning of words in sentences is emergent: meaning emerges from the mesh of affordances, learning history, and goals*' ([**Glenberg and Robertson, 2000**], **pag. 388**). Indeed, many symbolic models of language make use of sophisticated algorithms and techniques for representing knowledge, but they mostly ignore the role of experience.

However, symbolic approaches have had notable success in solving some tasks requiring logical reasoning, like for example playing chess; for example the IBM's Deep Blue chess computer in 1997 beat the international grandmaster Gary Kasparov. Nevertheless, Deep Blue's "intelligence" is extremely narrow in scope, considering that the system wouldn't be able even to recognise a chess piece or to carry on a conversation about the game won. Indeed, many other tasks in everyday life do not necessarily require the application of logical and systematic reasoning but instead other unconscious pattern recognition such as "intuition" for example.

3.2 Subsymbolic and Hybrid Models

Subsymbolic models emerged as an alternative to symbolic approaches. In subsymbolic models, such as artificial neural networks (based on the biological neural network metaphor), genetic algorithms (based on ideas of Darwinian evolution) and particle swarm optimisation (based on observations of bird flocking and other social behaviours), knowledge is represented by continuously valued (i.e. analogical) symbols. Subsymbolic systems, unlike symbolic approaches, do not require to provide to the system explicit formalization of knowledge through structural rules, but instead knowledge is represented by numerical patterns, which define the relations between

inputs and outputs of the system.

The best known subsymbolic systems are based on artificial neural networks (i.e. connectionism). Connectionist models are networks consisting of interconnected units, characterized from activation levels, which can transmit signals to other units along weighted connections [McClelland et al., 1986, Rumelhart et al., 1986]. Each unit computes its own output signal by: (i) weighting each of its input signals by the strength of the connection along which the signal is coming in, (ii) summing the weighted input signals, and (iii) feeding the result into a linear/non-linear output function, usually a threshold [Pinker and Prince, 1988]. The learning process in these models consists of adjusting the strengths of connections and the threshold-values, usually minimizing the distance between the actual output of the model and the corresponding target output value [Pinker and Prince, 1988]. Following the connectionist approach, a cognitive process is represented through a large number of interconnected neurons, which perform parallel computation. Connectionists models providing distributional representation and parallel processing of knowledge represent a powerful tool for modelling language acquisition and processing.

Connectionist models based on artificial neural networks have been extensively used as computational learning mechanisms for natural language processing (considering that in this approach knowledge associated to language can be learned from instances of usage), and they have been shown to model successfully a whole variety of language learning tasks. For example, in [Elman, 1990] a simple recurrent neural network model employs an additional input layer, so called context layer, which stores a copy of hidden units from the previous training step. The presence of recurrent links that feed back hidden units to the context units endows the network with a dynamic memory. Indeed, through this additional input layer the model has memory of the activation values of hidden neurons at the previous time step and it can use this information when processing the next input. In [Elman, 1990], this ungrounded model based on simple recurrent neural networks is used in a set of simulations ranging from the temporal version of the XOR problem (i.e. the logical

operation of "exclusive disjunction" also known as "exclusive or" that outputs true whenever both inputs differ) to the task for discovering syntactic/semantic features for words.

The modelling approach based on connectionism has been extensively used for the grounding of categories and naming tasks. For example in [Harnad et al., 1991] a three layer neural network model has been proposed for sorting lines into three categories (i.e. "short", "middle", "long"). Other connectionist models, in addition to the direct grounding of symbols, investigated the symbol grounding transfer, which refers to the process of transferring the grounding of basic categories to new symbols acquired via linguistic descriptions. For example, in [Greco et al., 2003] a connectionist architecture for category learning has been proposed; the network learns combinations of different shapes and colours, and new categories are learned via linguistic descriptions that permit to combined the symbols directly grounded in perception to create higher-order categories. However, connectionist models of language acquisition cannot easily scale up to larger data and the knowledge acquired through these models is not always easy to interpret and evaluate.

In connectionist models learning the meaning of a word is a matter of establishing a connection between a set of stimuli and verbal labels. For example, a feed-forward neural network can receive perceptual inputs associated to presented entities and in the hidden units it can learn the conceptual representation of the pattern in input through the application of the back-propagation algorithm [McClelland et al., 1986]. Similar concepts are represented by similar activation in the hidden units. The conceptual representations that develop through the training of the network are related arbitrarily to the perceptual states that activate them; hence these symbols that represent knowledge are amodal and arbitrary [Barsalou, 1999]. However, more recently it has been proposed that neural networks embodied in robotic platforms can be good candidates for modelling the acquisition of language [Sugita and Tani, 2005, Cangelosi and Riga, 2006, Marocco et al., 2010]. Indeed, associative networks that represent information in both the perceptual and cognitive domain, grounding

knowledge in perceptual systems, are not amodal [Pulvermüller, 1999].

Hybrid models, that combine the symbolic and connectionist approaches, have been proposed for the acquisition of language. Connectionist modules permit to ground basic symbols into perceptual categories and symbolic modules serve for the manipulation of such symbols.

3.3 Statistical Models

The statistical approaches to cognition permit to combine the descriptive power of symbolic models with the experience-based properties of connectionism. Many probabilistic models of language acquisition can be considered as a more sophisticated version of symbolic models where each rule has associated a probability. In the framework of statistical models, Bayesian networks can be considered to constitute a statistical account of the multi-modal information stored in the dynamic systems that generate simulations and guide situated action [Barsalou, 2008]. A Bayesian network is a graphical model that encodes probabilistic relationships among a set of random variables X:

$$X = \{X_1, \cdots, X_n\} \tag{3.1}$$

More specifically, a Bayesian network is a directed acyclic graph in which nodes represent random variables and arcs indicate probabilistic dependencies between nodes. Each node X_i is associated with a **conditional probability**:

$$P(X_i|Pa(X_i)) = \frac{P(X_i, Pa(X_i))}{P(Pa(X_i))} \tag{3.2}$$

that represents the probability that X_i occurs when $Pa(X_i)$ has already occurred. $Pa(X_i)$ represents the parent node of X_i. In equation (3.2), $P(X_i, Pa(X_i))$ can be substituted with:

$$P(X_i, Pa(X_i)) = P(Pa(X_i)|X_i)P(X_i) \tag{3.3}$$

and obtain:

$$P(X_i|Pa(X_i)) = \frac{P(Pa(X_i)|X_i)P(X_i)}{P(Pa(X_i))} \tag{3.4}$$

For computing $P(Pa(X_i)|X_i)$ the Bayes' rule (also known as inverse probability) is applied:

$$P(Pa(X_i)|X_i) = \frac{P(X_i|Pa(X_i))P(Pa(X_i))}{P(X_i)} \tag{3.5}$$

where $P(Pa(X_i)|X_i)$ is the probability of a hypothesis $Pa(X_i)$ given X_i, and $P(X_i|Pa(X_i)$ is the probability of X_i assuming that the hypothesis $Pa(X_i)$ is valid. $P(Pa(X_i))$ and $P(X_i)$ are the **prior probabilities** of the hypothesis $Pa(X_i)$ and evidence X_i, respectively. The goal of the Bayesian inference is to find the hypothesis that maximizes $P(Pa(X_i)|X_i)$. Bayesian networks provide a compact representation of the **joint probability** over all the random variables in the network. The joint probability represents the probability that two or more events occur together or in succession. Given $X = \{X_1, \cdots, X_n\}$ random variables, the joint probability is defined in the form below:

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} p(X_i|Pa(X_i)) \tag{3.6}$$

The process of using a Bayesian network to compute probabilities is called Bayesian inference. Nevertheless, inference in Bayesian networks is feasible in case of small networks, while it takes very long time in large networks. Bayesian networks can be constructed from domain knowledge by applying the following steps: (i) identify the variables in the interested domain, (ii) determine the direct influence relationships among variables in the domain, and (iii) determine the conditional probabilities given the structure of the Bayesian network. When for the construction of a complete Bayesian network the domain knowledge is not sufficient, the network can be learned from data.

In Bayesian networks language acquisition can be formulated as an induction

process. Indeed, inductive inference enables humans to make powerful generalizations from sparse data when learning about word meanings and many other aspects of the world [Tenenbaum et al., 2006]. A framework based on Bayesian theory for modelling inductive learning and reasoning has been proposed in [Tenenbaum et al., 2006], where one of the implemented tasks consisted in learning words (or category labels) by applying the Bayes' rule.

A Bayesian probabilistic model for learning semantic representations of concrete and abstract words has been proposed by Andrews et al. [2009]. They identified two statistical data types from which semantic representations of words can be learned [Andrews et al., 2009]. In particular, they argued that semantic representations of words can be derived from an optimal statistical combination of experiential data and distributional data. Experiential (or sensorimotor) data are sensorimotor and they are collected through the interaction of the body with the physical world; on the contrary, distributional (or linguistic) data describe the statistical distribution of words in language. In this framework, experiential and distributional data are both non-trivial source of information for obtaining semantic representations of words. Indeed authors have argued that a probabilistic model based on the combination of sensorimotor and linguistic data is a better predictor of human performance than a model based on one source of information only.

Cognitive models based on Artificial Neural Networks and Bayesian Networks have shown that the brain is sensitive to the statistical structure of experience [Barsalou, 2008]; for both approaches, if the processing occurs in a modular system separate from the brain's modal systems, then they remain ungrounded like traditional symbolic approaches.

3.4 The Developmental Robotics Approach

In most of the literatures so far, cognitive processes have been mainly investigated in the context of separate research areas. However, recent studies have shown that

mental processes are deeply influenced by the structure of the body and its interaction with the environment [Barsalou, 2008, Glenberg and Kaschak, 2002]. These new findings are now altering the relationship between different disciplines, ranging from computer science to robotics, cognitive science, developmental psychology and neuroscience, which are now working together to build a new interdisciplinary science. Development represents a key factor not only in disciplines like psychology for the investigation of the physical and cognitive human development that occurs throughout the entire life to better understand how people change and grow, and for the evaluation of children to determine if they have a developmental disability, but nowadays it is also very important in disciplines like robotics to achieve autonomous and intelligent behaviours.

The first time the word *robot* appeared was in 1921, in the title of the play R.U.R. (Rossum's Universal Robots) of the Czech writer Karel Čapek; in that context the word *robot* had the literal meaning of "serf labor". In 1979 the Robot Institute of America defined a *robot* as: '*a reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through various programmed motions for the performance of a variety of tasks.*' (**Robot Institute of America, 1979**)

Nowadays such definition of robot is applicable to industrial robots only, which are machines that can work in a structured environment and that can be employed for repetitive and precise tasks requiring transportation, manipulation or measurement. Recently new trends in robotics have started to work at the design of autonomous agents that can be employed in unstructured environments and that can be reactive to possible dynamic changes. The design of autonomous robots by pre-programming all the necessary behaviours for interaction is a quite challenging task, because is not possible to foresee and plan in advance all the possible situations that can happen.

By following the classical approach to robotics, the design of an autonomous robot requires the implementation of three independent functional modules (i.e.

sense-plan-act paradigm): (i) a perceptual system that through the usage of sensors extracts useful information from the environment, (ii) a planner for scheduling a sequence of actions that enable to achieve a specific goal and (iii) a motor system that executes the motor actions for the implementation of the desired behaviour. In this framework, the best example of the classical approach to autonomous robotics is represented by the Shakey robot [Nilsson, 1984] developed at Stanford Research Institute (SRI) for the Defense Advanced Research Projects Agency (DARPA). One of the main drawbacks of this approach resides in the lack of adaptiveness to unexpected modifications of the environment, due to the modular structure of the system and the independence of the layers responsible for sensing, planning and acting.

An alternative approach to robotics has been proposed by Rodney Brooks in 1991, when he introduced the behaviour-based robotics, arguing that the sense-plan-act paradigm used in the classical approach was not suitable for the construction of real working robots. Brooks proposed a new paradigm according to which the building blocks of an intelligent system must be simple sensorimotor behaviours that incorporate their own perceptual, modelling and planning requirements, on the top of which more sophisticated behaviours can be built. One of the drawbacks of this approach concerns the integration of different behaviours in order to obtain the control strategy of the overall system.

Recently, Cognitive Developmental Robotics (also known as Epigenetic Robotics) taking inspiration from developmental mechanisms studied in children by psychologists and cognitive neuroscientists, has started to focus on the modelling of different brain and behavioural processes in humanoid robots. In contrast to purely computational modelling methods, cognitive robotics focuses on the design of artificial architectures which integrate perception and action, capable of autonomous learning, decision-making and communication. This is an innovative approach to robotics that presents a strong interdisciplinary character and aims to overcome current limitations in robots design. Indeed, according to the developmental paradigm, instead of building robots that construct and maintain complex internal representations,

artificial agents are endowed with some basic perceptual and motor skills that can be subsequently re-arranged and integrated to interact in new scenarios. In other words, simple perceptual and motor skills can be reused to bootstrap the learning of more complex behaviours and robots can be flexible in the face of changing conditions in the environment. Indeed, cognitive abilities in humans develop over time layering over previous stages of development that may be a necessary way to manage complexity [Metta et al., 2001]. In line with this view, the aim of developmental robotics is not to model the end product of intelligence, which would be akin to adult level of intelligence, but the developmental process itself. The field of cognitive developmental robotics, still has to establish its definition, design principle, and methodology; however according to Asada et al. [2001] cognitive developmental robotics: '*aims to understand the cognitive developmental processes that an intelligent robot would require and how to realize them in a physical entity*' (**Robotics and Autonomous Systems, 2001**.). A more recent analysis of developmental robotics models and architectures has been proposed by Cangelosi and Schlesinger [2014].

Emerging theories on artificial cognitive systems can contribute to the current knowledge in neuroscience and psychology, and in turn, scientific and technological advances in cognitive robotics can have an important impact in developmental psychology and cognitive neuroscience, where humanoids can be used to formulate and test new hypotheses on cognitive functions in the study of human behaviour [Sandini et al., 2007].

3.4.1 Grounded and Embodied Connectionist Models

In the last decades, grounded models of language acquisition arose as reaction to purely symbolic approaches. The major novelty introduced by grounded models is the attempt to ground the meaning of words in referents in the real world; for example, the meaning of the word "round" is grounded in the visual features of entities, "push" in motor control structures, "heavy" in haptic features, and so on [Roy, 2005a]. Connectionist models embodied in robotics platforms represent a

powerful tool to study communication and language in artificial systems through a grounded approach. Indeed, connectionist architectures can be employed as control systems (i.e. "artificial brain") of robotics platforms; by following this approach, taking different kinds of sensory information as inputs, the network architecture activates the robot's motor joints according to the elaborated output. In such models, the external world plays an essential role in shaping the language used by cognitive systems. Indeed, linguistic abilities, which develop through the direct interaction of artificial agents with the environment, are grounded in the perceptual and sensorimotor knowledge of agents. This guarantees that symbols related to language are linked to perceptual internal representations.

For example, a framework for grounding nouns through the integration of action and perception (i.e. motor and sensor primitives) has been presented in [Roy, 2005a,b]. This framework has been used in a series of conversational robots that were able to translate spoken commands such as "hand me the blue one on your right" into situated action. These robots were endowed with a three-dimensional "mental model" of the physical environment updated according to the linguistic, visual or haptic inputs. According to this framework proposed for the grounding of language, words that refer to actions (e.g. verbs like "push") are grounded in sensorimotor control structures, while words that refer to perceptual properties (e.g. adjectives like "red") are grounded in sensory expectations associated with specific actions (e.g. "red" is a colour category linked to the motor program for directing active gaze towards an object). Furthermore, in this framework object names (e.g. nouns like "ball") are grounded in the perceptual properties of objects and in all the motor affordances that may affect objects. This model is consistent with the notion of schemas proposed by Piaget [Piaget and Cook, 1952], according to which the meaning of words is grounded in both perceptual features and motor programs.

Other robotics models have focus on the acquisition of language through the interaction with human users. For example in [Dominey et al., 2009] robotic technology including vision and motion planning were integrated together with aspects

of cooperative behaviour and language-based communication, in order to provide a coherent system for adaptive human-robot interaction. A user through spoken language can interact with a humanoid robot to command in real-time sequences of behaviours. The robot can either receive action commands (e.g. "Left open", "Give it to me", "Right close", etc.) or control commands (e.g. "Learn", "OK", "Macro", etc.). Through this system the robot can react to language commands and learn in real-time new behaviours by combining pre-existing motor skills.

Models to study the emergence of shared lexicons through biological and cultural evolution mechanisms have been proposed in [Cangelosi, 2001, Cangelosi and Parisi, 2002]. In these models, a population of artificial agents, initialized to use random languages, after an iterative process of communication and "language games", converges towards the usage of a shared lexicon. The paradigm of "language games" for language acquisition has been used extensively by Luc Steels and collaborators according to whom [Steels, 2001]:

> *A language game is a sequence of verbal interactions between two agents situated in a specific environment. Language games both integrate the various activities required for dialogue and ground unknown words or phrases in a specific context, which helps constrain possible meanings* (**Intelligent Systems, 2001**)

In [Steels, 2001] language games are proposed as a parading to solve the challenge of integration and grounding for human-robot dialogue. To implement the language games idea, Steels and colleagues employed different experimental platforms including different generations of Sony robots. For example, in the "Talking Heads" experiment [Steels et al., 2002], agents look at a white-board containing coloured geometric figures, which the robots use as subjects of a language game; this experiment has demonstrated that a shared lexicon gradually emerges to describe a world made of coloured shapes. Such model has been extended in [Steels and Kaplan, 2002] to study the emergence of communication between humans and the Sony AIBO robot; it has been shown that any kind of concept acquisition can be

used (e.g. a single object, or an action or property of the situation). One peculiar aspect of the approach proposed by Steels and collaborators is the importance of social mechanisms in the grounding and emergence of language.

Other studies have focused on developmental aspects (like "intrinsic motivation") as factors that favour the acquisition of language. Architectures based on intrinsic motivation make use of particular types of reinforcement learning in which rewards are provided not from external means but through internal evaluation [Oudeyer and Kaplan, 2006]. For example, in [Oudeyer and Kaplan, 2006] a computational model and a robotic experiment have tested the hypothesis that children discover communication by exploring and playing within their environment. The experiment, that focused on the role of intrinsic motivation and active exploratory behaviour, has shown that "intrinsic motivation" toward the experience of novel situations, which 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.

Other models have focused on the learning of semantic combinatoriality from the interaction between linguistic and behavioural processes. For example, in [Cangelosi and Riga, 2006] a simulated robot learns to perform via imitation a set of basic actions, which can be recalled by their names. The combination of words associated to such basic actions leads to the acquisition of higher-order concepts. The results of this experiment have shown that the simulated robot is capable to perform concrete actions and understand each action's name. Another example of semantic combinatoriality is given in [Sugita and Tani, 2005]; experiments on a real wheeled robot equipped with a two degree of freedom arm and a vision system have been presented. In this experiment the robot learns a set of behaviours by interacting with objects that are associated with two-words sentences consisting of a verb to refer to the behaviours and a noun to refer to the objects. The robot, controlled by a recurrent neural network with parametric bias nodes (RNNPB), is trained through learning via demonstration. The RNNPB model is based on a

Jordan simple recurrent neural network [Jordan, 1986] with parametric bias nodes (PB) in the network's input layer for modulating its own dynamic function. The RNNPB controller consists of two modules which are responsible for behavioural and linguistic tasks, respectively. The PB nodes containing some shared neurons between the two modules enable the interaction of the two modules. The learning of the model is supervised and performed through back-propagation through time (BPTT); two different mechanisms are used for the connection weights modification and PB vector modification. The robot interacts with three coloured objects (i.e. "red", "blue" and "green") on each of which it can perform three different behaviours (i.e. "pointing at", "pushing" and "hitting"). The robot is also trained to learn and recognize language commands. After the training, the robot has exhibited the ability to translate linguistic commands into the correspondent situated action and to produce the appropriate language output associated to the performed behaviour. In this model the robot represents the meaning of words and the corresponding behaviours in a compositional manner. Furthermore, in [Yamashita and Tani, 2008] the emergence of functional hierarchy in a multiple time-scale neural network model has been presented; a humanoid robot stands in front of a workbench, where a goal object of cubic shape is placed. The task for the robot consists of autonomously learning five basic behaviours. The results of this experiment have shown that the humanoid robot can learn to generate object manipulation behaviours in a compositional way; basic behaviours, such as touch/lift/move objects are sequentially combined by utilizing inherent time constant differences (i.e. slow and fast units) in the employed neural network model. The results of this experiment have suggested that multiple time-scales (i.e. primitives are represented by fast context units whose activity changes quickly, while sequences of primitives are represented by slow context units whose activity changes slowly) are an essential factor for the emergence of functional hierarchy in neural systems. In [Morse et al., 2010] a robotic model based on the embodiment of language acquisition has been presented; such robotic model supports the hypothesis presented by [Smith and Samuelson, 2010] that con-

siders the body posture central to the linking of linguistic and visual information. The model proposed by Morse et al. [2010] has been used to replicate the Smith and Samuelson [2010] child psychology experiments. Indeed, the participation of a humanoid robot in a psychology experiment permitted highlighting the role of body posture and spatial location while learning object names. The results of the robotic experiments have confirmed that body posture affects the linking of linguistic and visual information. Additionally, it has been shown that changing posture from sitting to standing can disrupt such ability [Morse et al., 2010].

The grounded and embodied connectionist models presented in this section have shown that cognitive robots can be been successfully employed for learning words that refer to concrete objects and actions. Although abstract concepts appear to play a central role not only in human cognition but also for the development of intelligent agents that can autonomously create categories and use language, building intelligent systems that can learn their meaning is still a challenging task for cognitive developmental robotics. The work presented in this thesis aims to propose a mechanism for the grounding of abstract words in robots through the implementation of neuro-robotic models, where the meaning of higher-order concepts is obtained through the hierarchical organization of basic sensorimotor concepts; in this dissertation it is proposed that such hierarchical organization of concepts can be a possible account for the acquisition of abstract words in cognitive robots.

3.5 The iCub Robotic Platform

The principles behind developmental robotics have also inspired the design of humanoid robotics platforms. One of the most prominent example of robots built by following this approach is the iCub humanoid [Metta et al., 2008]. The iCub, designed by the RobotCub Consortium, is an open-source robotic platform for research in embodied human cognition, artificial intelligence, and cognitive and brain inspired robotics research (Fig.3.2(a)). More specifically, the iCub robot, which represents

the state-of the-art humanoid in Europe, has been designed to support research in the themes of learning, control, cognition and interaction.

(a) (b)

Figure 3.2: The iCub: real robotic architecture (a) and iCub simulator (b)

3.5.1 Hardware Description

The iCub robot is 104 cm tall and its overall weight is 22 kg; its dimensions are similar to those of a three and half year old child. The iCub has a PC104 machine located in the head, which can communicate with actuators and sensors, and small micro-controller boards located in the torso. The robot is equipped with a body cover; lines of red LEDs representing mouth and eyebrows are mounted behind the face panel for making facial expressions. Considering that the robot originally was not designed for autonomous operation, it was not equipped with on-board batteries or processors; instead an umbilical cable provides power and a network connection.

The iCub kinematic structure consists of several rigid bodies connected through joints which allow motion (e.g. rotation, complex motion) between the connected bodies. Hence, joints determine the Degrees of Freedom (DOF or mobility) of the

system; DOF correspond to the number of independent parameters that define the system configuration. Each joint is driven by actuators (i.e. motors) and typically the number of DOF of the joint identifies the number of actuators needed to define the system configuration. The iCub actuators were selected according to the torque requirements for each joint, which were calculated by using the Webots [Michel, 2004] simulator that is based on ODE (Open Dynamic Engine) which is an open source library for simulating three-dimensional rigid body dynamics. The actuators adopted for the iCub are based on a combination of brushless Direct Current (DC) motors with speed reducers; this solution was preferred to other options (conventional DC brushed motors) because of their higher robustness and reliability. Although brushless motors offer higher performance and efficiency than brushed motors, they need complex electronic control. However, large joints (as for example the shoulder) have brushless motors, while small joints (as the hand) have brashed motors. The iCub has 53 DOF distributed on the head, torso, legs, arms and hands [Parmiggiani et al., 2012].

DOF and Actuators

- **head**: 6 DOF (3 neck, 3 eyes). The three DOF of the neck enable a serial pitch, roll and yaw configuration. The three neck joints are driven by brushed DC motors. The two cameras are moved by a three DOF eyes mechanism which allows both tracking and vergence behaviours. The eyes movement is enabled by three DC brushed motors.

- **torso**: 3 DOF. Two base motors actuate jointly the pitch and roll axes whereas a third motor group drives the yaw joint.

- **legs**: 6 DOF in each leg (3 hip, 1 knee, 2 ankle). The first DOF of the hip is driven remotely by means of a cable drive actuated by a motor which is located in the lower torso assembly. The DOF in the knee, is actuated by the knee exion/extension motor, and a two DOF ankle are actuated by a brushless motor housed in the lower leg segments and by a smaller motor group placed

directly on the foot.

- **arms**: 7 DOF in each arm (3 shoulder, 2 elbow, 2 wrist). The three brushless motors driving the shoulder are housed in the upper-torso frame. The brushless motor driving the elbow is housed at the center of the elbow assembly.

- **hands**: 9 DOF in each hand (3 for the thumb, 2 for the index, 2 for the middle finger, 1 for the coupled ring and little finger, 1 for the adduction/abduction). Seven out of the nine motors driving the hand joints are placed in the forearm assembly. Given the limited amount of space available in the hand, brushed DC electric motors were employed. These electric motors are coupled to speed reducers to obtain the desired torques. Two motors, placed directly inside the hand assembly, are used for adduction/abduction movements of the thumb and of the index, ring and small fingers.

The iCub has been specifically designed to maximize the number of degrees of freedom allocated to the hands, with the constraint of the overall small size. Originally the iCub legs have been designed mainly for crawling; currently, new foot design is seeking to enable the iCub for bipedal locomotion. Additionally, the iCub can stand on top of the iKart, a mobile base for the robot which mounts six wheels, a high performance i7-CPU, wireless connection and high performance Li-ion batteries, that can be controlled using a standard interface.

A controller can make robotics joints to behave as desired. The control of a single joint, as shown in figure 3.3, requires several components: (i) a digital microprocessor that consists of a micro-controller and a processor with special interfaces, (ii) an amplifier that drives the actuator and turns the control signals into power signals, (iii) an actuator (e.g. electric motor Direct Current (DC) that can be either brushless or with brushes), (iv) the mechanical system to be controlled (e.g. the joint of robot) and (v) the sensor that measures the output produced by the system and feeds it back to the microprocessor where it is compared with a reference value for the computation of the new control signal.

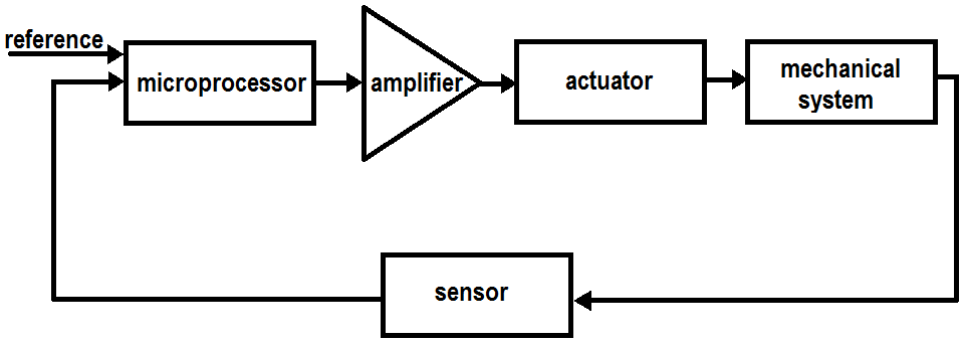


Figure 3.3: Block diagram for the control loop of a single joint

When the iCub robot is switched on, it immediately starts to move. By running the motor interface the robot perform self-calibration to reach its calibration position.

Sensors

The iCub is equipped with different types of sensors:

- 2 digital cameras (located in the head)
- 2 microphones (located on the side of the head)
- encoders: for positional control
- inertial sensors (e.g. accelerometers and gyroscopes, located in the head) measure the three components of linear accelerations and angular velocities
- 4 force/torque sensors (2 in the upper arms and 2 in the legs)
- distributed pressure sensing capacitive skin system based on a modular triangular structure in two forms:
  - 108 tactile sensors in the fingertips and palm mainly used for collision detection
  - generic body skin on the forearm (in a new version of the robot, tactile sensors will be embedded in the fingertips, the palms, the forearms, the upper arm segments, the torso, the upper leg segments, the knees, the lower leg segments and the feet). The tactile sensors can be used for better and safer human-robot interaction

3.5.2 Software Architecture

The iCub software architecture, largely written in C++ programming language, is based on YARP (Yet Another Robot Platform) [Metta et al., 2006], which is an open-source and multi-platform framework for humanoid robotics, consisting of a set of libraries, protocols and tools that support distributed computation and that can be used for inter-process communication on a local network. YARP, adopted by the RobotCub consortium as the middle-ware for the iCub humanoid robot, permits to decouple devices from software architecture and to exchange information between the user code and the robot with its environment. The core components of YARP are:

- *libYARP_OS*: for interfacing with the operating system and to provide some basic services (e.g Thread, Semaphore, etc.). This library also provides easy network communication using the YARP Port Network
- *libYARP_sig*: for common signal processing tasks (visual, auditory)
- *libYARP_dev*: for interfacing with common devices drivers used in robotics (sensors and actuators)

YARP has a command-line interface that permits to perform several operations such as give status information, make and break connections between ports, and send/receive data to/from ports. YARP also provides an image viewer to visualize image transmitted in standard network format.

The iCub software repository contains many software modules and applications that can be used for controlling the robot through the YARP interface. The documentation and low level code developed for the iCub robot is available as open-source code. More specifically, the iCub software repository contains modules, graphical user interfaces (e.g. "robotMotorGui" for moving the joints of the iCub robot using sliders, "iCubSkinGui" to display the output of fingertip/skin tactile sensors, etc.) and libraries ("iKin" for forward-inverse kinematics, "iDyn" for forward-inverse kinematics and dynamics, "actionPrimitives" for primitive actions like reach, grasp,

etc.). Furthermore, from the collection of several modules it is possible to obtain useful applications.

The Action Primitives library, based on the YARP Cartesian Interface has been used in the first experiment proposed in this dissertation for implementing motor primitives like reach, grasp, etc, and for combining them to form higher level actions that permit to execute more complex tasks without considering the motion control details [Pattacini et al., 2010]. The functions contained in the library permit to: (i) move the arm of the robot to a specific pose (i.e. position and orientation), (ii) execute a predefined fingers sequence and (iii) wait for a specific time interval. For producing an action, the corresponding request item is "*pushed*" in the actions queue by using the function *pushAction ("params")* that allows to insert in the action queue a sequence of elementary actions to be executed.

In the second experiment presented in this dissertation, the Cartesian Controller, available in the iCub software repository, has been used in order to implement the desired motor primitives by solving the inverse kinematic problem and to control the robot's arm. The Cartesian controller consists of two modules [Pattacini et al., 2010]:

- Solver: through a non-linear optimizer, which takes into account all the imposed constraints, determines the arm joints configuration that permits to achieve the desired pose (i.e. end-effector position and orientation). The solver uses the IpOpt [Wächter and Biegler, 2006] software package to solve the following non-linear optimization problem (i.e. inverse kinematic):

$$q = \arg\min_{q \in \mathbb{R}^{10}} \left(\frac{1}{2} \parallel \alpha_d - K_\alpha(q) \parallel^2 + w \cdot \frac{1}{2} \parallel q_{rest} - q \parallel^2 \right) \tag{3.7}$$

$$\text{subject to} \quad \begin{cases} \parallel x_d - K_x(q) \parallel^2 \; < \epsilon \\ q_L < q < q_U \end{cases}$$

  where q is the desired joints vector that has 10 components in case the 7 joints of arm and the 3 joints of torso are controlled. x_d and α_d represent the desired

position and orientation, respectively. K_x and K_α are the forward kinematic maps for the position and orientation, respectively. q_{rest} is used to keep the torso as close as possible to the vertical position while moving and w is a positive factor $w < 1$ that weights q_{rest}. q_L and q_U represent the physical bounds of the joints and ϵ is a small number in the range between $[10^{-5}, 10^{-4}]$.

- Controller: it computes the velocity of the motors that generate a human-like quasi-straight trajectory of the robot end-effector

3.5.2.1 The iCub Simulator

The iCub Simulator (Fig.3.2(b)) is an open-source multi-platform computer simulator, licensed under General Public License (GPL) [Tikhanoff et al., 2008, 2011]. The simulator is based on the Open Dynamic Engine (ODE) library that simulates rigid bodies and the collision detection algorithms to compute the physical interaction with objects, and the OpenGL/SDL library that provides a rendering engine designed to reproduce as accurately as possible the physics and the dynamics of the real robot. The simulated iCub, constructed collecting data directly from the robot design specifications in order to achieve an accurate replication (e.g. height, mass, degrees of freedom) of the iCub prototype developed at the Italian Institute of Technology in Genoa, is composed of multiple rigid bodies connected via joint structures. The simulator permits the testing of algorithms in order to verify their correctness prior to use the physical robot. Considering that the the simulated and real robot are provided with the same software interfaces, minimal changes to the code permit to transfer the developed algorithms from the simulated iCub to the real robot. The iCub simulator has a configuration file that permits to set the desired iCub parts activation before running the simulator. Keyboard and mouse are used for the manual navigation of the environment. The simulator allows to create static and dynamic object of different shape in the environment and also to import 3D models on it. Additionally, it is possible to get and set the position of created objects and to rotate them.

The iCub Simulator (Fig.3.2(b)) has been used for testing the developed algorithms prior the use the real robotic architecture. Subsequently, experiments have been run on the real iCub humanoid robot, adopted as robotic platform.

Chapter 4

Neural Network Algorithms for Modelling and Analysing Language

Following the bottom-up approach to artificial intelligence, briefly introduced in Chapter 3, aspects of cognition and intelligence including language representations can be reproduced by using artificial neural networks (ANNs). An ANN is a computational model inspired by the organizational structure of the human brain, composed by a large number of units (referred as neurons) and connection weights (or synaptic links) that decide the strength of connections between units.

The first model introduced for artificial neurons, which is still used in neural network modelling, was proposed in McCulloch and Pitts [1943] and called Threshold Logic Unit (TLU). Few years later, in 1949 the psychologist Donald Hebb introduced a rule (later defined as Hebbian Learning) for learning connections between neurons; the rule implied that connections between two neurons are strengthened when both neurons are active at the same time [Hebb, 1949]. In 1958 Frank Rosenblatt introduced an algorithm for supervised classification of inputs [Rosenblatt, 1958] known as perceptron. In 1969, the first artificial neural network model was presented by Minsky and Papert [1969]. Advances in neural network processing were achieved

through the introduction of the back-propagation algorithm [Werbos, 1974, Parker, 1985, LeCun, 1986, Rumelhart et al., 2002]. In 1986 thanks to David E. Rumelhart and James McClelland the parallel distributed processing, making use of the back-propagation algorithm, became popular under the name of connectionism [McClelland et al., 1986].

In this chapter some of the main artificial neural network models and learning algorithms, which will be used in this thesis, are presented. Further, the chapter contains a description of the methods used for the analysis of the internal dynamics of the models implemented for carrying out the experimental studies presented in Chapter 5, Chapter 6 and Chapter 7 of this dissertation.

4.1 Artificial Neural Network Models

An artificial neural network, in analogy with the biological neural system, is a non-linear parallel processing computational model that consists of simple interconnected units, called neurons, that can exchange information by means of connections that can be active or inhabited according to the value of their corresponding weights. Artificial neural network based models are ideally suited when is not possible to define an algorithm for task completion. Some of the fields in which neural networks find application to solve different types of problems are:

- Classification: according to a measure of similarity/dissimilarity similar input patterns are associated together (e.g. Pattern recognition, clustering, feature extraction, image matching)
- Regression and Prediction: inferring unknown data by relying on historical data (i.e. extrapolation)
- Optimization: minimize a specific cost function with respect to some constraints
- Control: as robotic controller, a neural network by establishing a relation

between inputs (e.g. sensors) and outputs (e.g. actuators) signals of the system, can control the behaviour of a robot

The three main classes of network architectures, that is, (i) single-layer feed-forward, (ii) multi-layer feed-forward and (iii) recurrent architectures are presented in the next sections of this chapter.

4.1.1 McCulloch-Pitts Model and Perceptron

In 1943 McCulloch and Pitts proposed a computational model which was a neural network implementation of propositional logic. The McCulloch-Pitts model for an artificial neuron consists of: (i) one or more input units $X = [x_1, x_2, \cdots, x_n]$ where $X \in \mathbb{R}^n$, (ii) an internal activation function $f(\Sigma)$ and (iii) one output y (Fig.4.1).

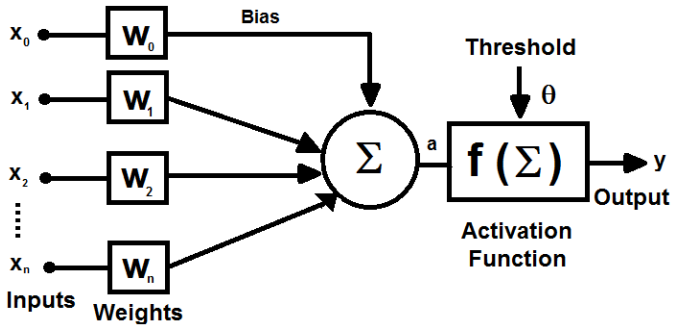


Figure 4.1: Model of an artificial neuron proposed by McCulloch and Pitts

Each neuron represents a multiple-input, multiple-output (MIMO) system that receives n signals from the inputs, produces one output signal and transmits it to all the other units. In particular, the input signals (x_i) traverse weighted connections (w_i) and generate an internal activation signal a, which is a linear weighted sum of the input signals to which is added the bias value (w_0) (Eq.4.1).

$$a = \sum_{i=1}^{n} (w_i \cdot x_i) + w_0 \tag{4.1}$$

From equation (Eq.4.1) it is possible to notice that the relation between input and output depends from the variation of the synaptic weights w_i that models the

synaptic efficacies of inter-neuron synapses. Positive weights correspond to excitatory synapses, while negative weights model inhibitory synapses. The activation value a of the neuron is subsequently transformed through the activation function $f(\Sigma)$. The activation of the McCulloch-Pitts model is regulated by a step function (Fig. 4.2), which implies that the output y of a neuron is either activated or deactivated, depending on whether the threshold value θ is reached or not:

$$y = f(\Sigma) = \begin{cases} 1, & if \quad a \geq \theta \\ 0, & if \quad a < \theta \end{cases}$$

The step activation function in the McCulloch-Pitts model determines a binary classification of the inputs that are categorised into one of two possible groups (1 or 0).

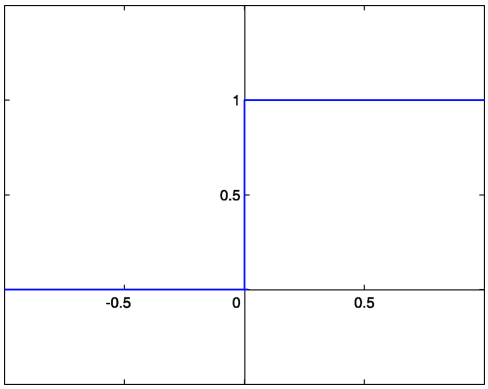


Figure 4.2: Step activation function profile. Source wikibooks.org

Indeed, such model can be used as a linear separator, considering that produces two categories in the input space. In 1958 the perceptron neural network model, consisting of a set of neurons based on the McCulloch-Pitts model and distributed in the input and output layer (**single layer perceptron**), was introduced [Rosenblatt, 1958]. The connection weights in a single layer perceptron are learned by applying the *delta rule* [Widrow et al., 1960]. Simulations with the single layer perceptron showed that this model could easily implement the major logic functions (e.g. AND, OR, NOT). The implementation of the AND function is described; the input patterns consist of two signals (x_1, x_2) weighted by (w_1, w_2) equal to 1 and the bias

value w_0 is set to -1.5. The output unit employs a threshold activation function, which in case the activation value "a" is greater than zero it sends in output one, while if "a" is less than zero it sends zero. From the activation table associated to the logic AND implemented by the perceptron model (Tab.4.1) it is possible to observe that the output of the model is activated in correspondence of the input ($x_1 = 1$, $x_2 = 1$) only.

x_1	x_2	activation	y
0	0	$(0 \cdot 1) + (0 \cdot 1) - 1.5 = -1.5$	0
0	1	$(0 \cdot 1) + (1 \cdot 1) - 1.5 = -0.5$	0
1	0	$(1 \cdot 1) + (0 \cdot 1) - 1.5 = -0.5$	0
1	1	$(1 \cdot 1) + (1 \cdot 1) - 1.5 = 0.5$	1

Table 4.1: Activation table for the logic AND implemented by the perceptron model

From a geometrical prospective, the perceptron model that implements the AND logic function represents a linear operator in the input space that seeks to find a hyper-plane the separates the input space into two categories (Fig. 4.3).

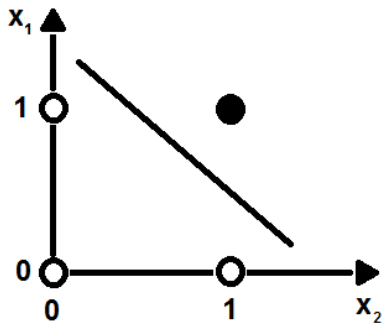


Figure 4.3: Geometrical representation of the input space for the AND logic function

Indeed, this model separates the point of coordinates $(1, 1)$ from the other three points $(0, 0)$, $(0, 1)$ and $(1, 0)$. Hence, the perceptron model classifies the input patterns in one of two possible classes. One of the limitations of the perceptron network was noticed by Minsky and Papert [1969] that published a mathematical analysis of the perceptron to point out that such model was not able to classify input patterns not linearly separable in the input space. To illustrate this limitation, Minsky and Papert used the XOR (i.e. exclusive or) logic function that is a typical

example of non-linearly separable function. This function takes two input arguments with values in $[0, 1]$ and returns one output in $[0, 1]$. The output is 1, if and only if, the two inputs have different values (Tab.4.2).

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Table 4.2: Truth table for the logic XOR function

From the geometrical representation of the input space for the XOR logic function it is possible to notice that this function is not linearly separable (Fig.4.4) and therefore the perceptron model cannot separate the point of coordinates (1, 1) and (0 ,0) from the other two points (0, 1) and (1 ,0).

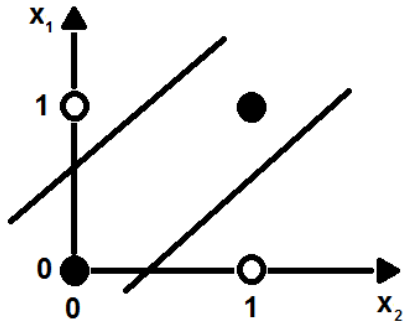


Figure 4.4: Geometrical representation of the input space for the XOR logic function

For solving non-linear separably problems the multi-layer perceptron (or MLP) model was introduced. MLP is a more general network architecture, where hidden layers are added between input and output layers. In parallel, alternative methods have been proposed for non-linear separably problems. A generalization of the single layer perceptron topology (SLPT), called recursive deterministic perceptron (RDP), was introduced in [Tajine and Elizondo, 1998]. To construct a RDP several growing methods were proposed. These methods consist of incrementally adding Intermediate Neurons (IN) to the topology; each of these IN represents a SLPT and they have a similar function to that of the hidden units in the back-propagation

algorithm. The resulting topology is a feed-forward multilayer neural network that permits to deal with non-linearly separable problems.

4.1.2 Multi Layer Perceptron and Recurrent Architectures

A multi layer perceptron is an artificial neural network model in which neurons are arranged in multiple layers (an input and output layer with one or more hidden layers) to constitute a directed graph, with each layer fully connected to the next one (Fig.4.6 (a)). Neurons of a multi layer perceptron are of three different types: (i) input neurons that receive the information to be processed, (ii) output neurons that contain the results of the computation and (iii) hidden neurons that are in between input and output neurons and do not directly receive inputs nor send outputs to the external environment. Except for the input layer, each neuron has a non-linear activation function which must always be normalizable and differentiable. One of the most used activation function in MLP is the sigmoid (or logistic function in case the sigmoid ranges from [0, 1]).

$$y = f(\Sigma) = \frac{1}{1 + e^{-\beta}} \tag{4.2}$$

where β is the slope parameter. The profile of a logistic function is shown in figure 4.5. The popularity of sigmoid functions in neural networks is also due to the fact that their derivatives are easy to calculate, which turns out to be very useful during the computation of the weight updates in certain training algorithms.

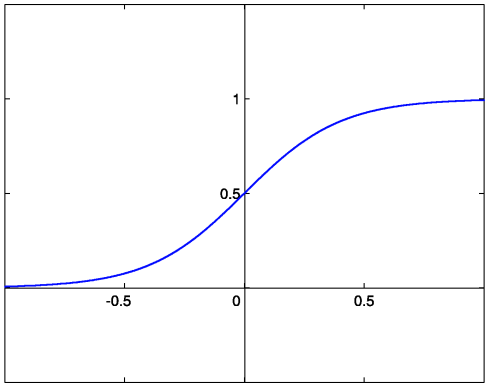


Figure 4.5: Sigmoid activation function profile. Source wikibooks.org

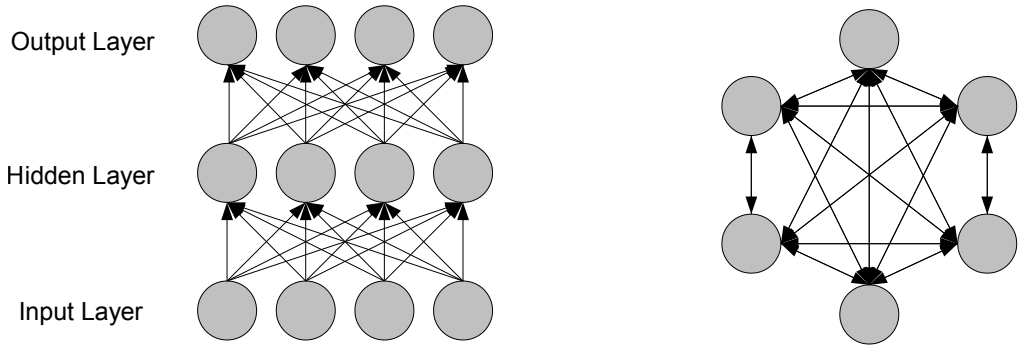
According to their topology, MLPs can be distinguished in different classes.

Feed-forward NNs consist of a set of neurons distributed in the input, hidden and output layers (Fig.4.6 (a)), for which the information flows in one direction only (i.e. *forward*) from the input units to the output ones, without feedback loops. These networks do not have internal memory and they can learn a *static mapping* between input (X) and output (Y).

$$Y = f(X) \quad X, Y\, static\, patterns$$

Recurrent NNs are characterized from a *bidirectional* flow of information, which is possible through the presence of recurrent connections (feedback) that go backward from output to input units providing to the network internal memory (Fig.4.6 (b)). This kind of networks are suitable for modelling dynamic temporal behaviours. They are dynamical systems that can learn a *non-static mapping* between the Input (X) and Output (Y). This characteristic makes recurrent neural networks particularly suitable for sequence processing.

$$Y(t) = f(X(t)) \quad X, Y\, time - varying\, patterns$$



(a) (b)

Figure 4.6: Topology of a: feed-forward (a) and fully recurrent neural network (b)

An example of recurrent neural networks is provided by the Hopfield network [Hopfield, 1982], which is a fully connected feedback network in which symmetric

inter-neuron synapses guarantee that the network energy function decreases monotonically; this type of network is mainly used as an associative memory or to solve optimization problems. Hopfield network with asymmetric inter-neuron synapses are used in networks with periodic and chaotic behaviour.

Recurrent networks in which the feedback signal is only in one of the layers of the network are called Simple Recurrent Neural Networks (S-RNN); examples of simple recurrent neural networks are Elman and Jordan architectures (Fig.4.7). An Elman network [Elman, 1990] is a three-layer perceptron with the addition of context units in the input layer and recurrent connections, with weights fixed to a constant value equal to one, from the hidden layer to the context units (Fig.4.7 (a)). The recurrent connections have the role to keep a copy of the value of the neurons in the hidden layer at the previous instant time $(t-1)$. Indeed, the context units at the time (t) contain a copy of the hidden units at the time $(t-1)$. This enables the Elman network to "remember" its previous state which permits the performance of tasks which require the prediction of time sequences that cannot be obtained with a conventional feed-forward network.

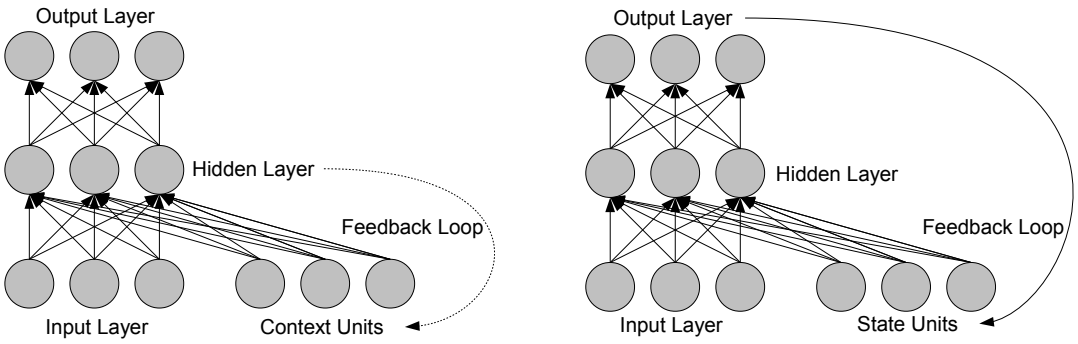


Figure 4.7: Topologies of simple recurrent neural networks: Elman (a) and Jordan networks (b)

Jordan networks [Jordan, 1986] are similar to Elman architectures but instead of context units they have state units that contain a copy of the output layer (Fig.4.7 (b)). At each time step, the inputs are propagated in the same way of feed-forward networks, including the application of the learning algorithm (usually

back-propagation). Jordan networks (as well as Elman architectures) are Discrete-Time Recurrent Neural Networks (DT-RNN) in which the processing occurs in discrete steps and each neuron computes its output spontaneously. For DT-RNN the relation between inputs and outputs is governed by a functional equation $f(x)$ that can be both linear or non-linear. An interesting learning algorithm that can be applied to Jordan networks is the "teacher forcing algorithm" that instead of feeding the state units with the actual output of the network, it feeds the desired target output value as the network runs; hence, this algorithm forces the output units to assume the correct states, even as the network runs. This algorithm has advantages in terms of convergence of the learning [Pearlmutter, 1990].

Contrary to the DT-RNN, in Continuous Time Recurrent Neural Networks (CT-RNN) inputs and outputs are functions of continuous time variables and the relation between inputs and outputs is governed by a differential equation in time [Pineda, 1987] rather than a functional equation. Hence, neurons have a temporal response that relates the state of the network to inputs. A Continuous Time-RNN implements a feature of biological neurons, namely that the activities of neurons are determined not only by current synaptic inputs but also by the past history of neural states. Due to this characteristic according to which activation changes continuously, the CT-RNN can better model mechanisms for producing continuous sensorimotor sequences than DT-RNN models.

Another important example of neural network based model is provided by Self Organizing Maps (SOM, also known as Kohonen maps) that are used for unsupervised learning [Kohonen, 1982]. In SOM neurons are interconnected in a grid and groups of neurons self organize in specific regions; nearby locations in the map represent inputs with similar properties. Similarly to other techniques (e.g. Principal Component Analysis), SOMs permit to reduce the dimensions of data [Fodor, 2002].

4.2 Supervised Learning

Connectionism, in contrast to some symbolic models used for knowledge representation, has the advantage to be based on learning methods. Generally speaking, models based on learning are designed to support automated knowledge acquisition, fault tolerance, and induction [Wermter et al., 1996]. This is particularly important in the field of natural language processing, considering that learning enables several language-related tasks, such as speech recognition, spoken language understanding, machine translation and information retrieval [Wermter et al., 1996]. Furthermore, models based on learning permit to design more flexible, scalable, adaptable and portable natural language systems [Wermter et al., 1996].

In MLP models, a learning algorithm is a mathematical method that computes the update of the synaptic weights that better approximate a desired function. Indeed in a neural network based model, synaptic weights represent the most important factor in determining its function. A learning algorithm can be (i) supervised, (ii) unsupervised and (iii) reinforcement learning. In **supervised algorithms** a neural network, by receiving pairs of inputs and target outputs (i.e. training examples) that describe the relations between inputs and outputs and represent the knowledge/experience about the task, through the learning process has to find the function that permits to match the training examples. In **unsupervised learning**, only the input patterns are provided to the network; the learning seeks to find hidden structures in data and to understand how data are organized. In **reinforcement learning**, data are generated by software agents that through interactions with the environment seek to maximize reward functions.

One of the most common supervised learning method for training neural networks is the Back-Propagation algorithm (BP) [Rumelhart et al., 2002]. The learning through BP consists in: (i) providing the input patterns X to the network, (ii) calculating the corresponding output Y, (iii) and computing the error signal E by comparing the output Y with the desired target output values $\hat{Y}$. Then, the error of the network is propagated backward and the connection weights of neurons are

updated. Before the training through back-propagation can start, it is necessary to:

- define the topology of the network (e.g. number of neurons in input and output, number of hidden layers, activation function, etc.) that depends on the specific task and set the value of some important parameters for the training process (e.g. learning rate and momentum)
- collect the training set of the network that describes the relations between inputs and outputs of the network. The sample of the training set are divided into two independent sets: the training set used to train the network and the testing set used to test the performance of the neural network
- initialize the weights to small random values (typically in the interval [-1, +1] or [-0.5, 0.5])

Given a feed-forward neural network with $X \in \mathbb{R}^N$ neurons in input, $\hat{Y} \in \mathbb{R}^M$ target outputs and $Y \in \mathbb{R}^M$ neurons in output:

$$X = [x_1, x_2, \cdots, x_N]$$
$$\hat{Y} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_M]$$
$$Y = [y_1, y_2, \cdots, y_M]$$

the back-propagation algorithm consists of the repeated application of two stages: forward propagation and backward propagation (Fig.4.8).

- **Forward propagation**: the network receives the input patterns X and calculates the corresponding outputs Y by using (Eq.4.1) and (Eq.4.2). In the next algorithm's step the output signal (Y) of the network is compared with the desired output values ($\hat{Y}$) to calculate the error signal (E) (Eq.4.3).

$$E = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{Y}_j^i - Y_j^i)^2 \tag{4.3}$$

  The error E is a cost function defined on the observations of the system; the minimization of the error E leads to the minimization of the difference between

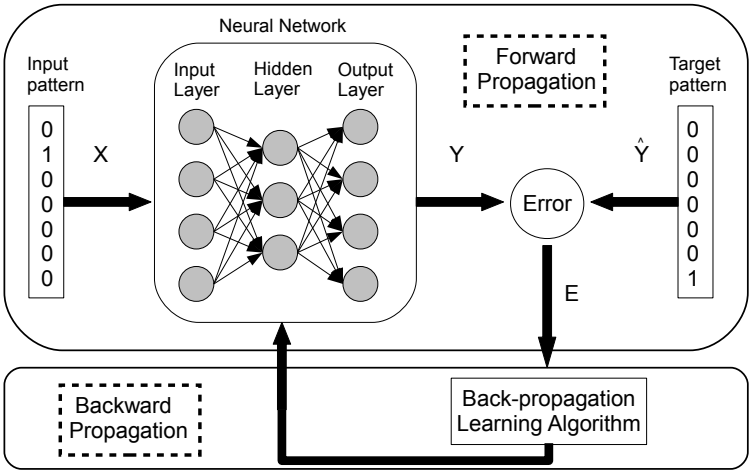


Figure 4.8: Illustration of the Back-propagation learning method

the output of the network (Y) and the desired output values ($\hat{Y}$) (i.e. total Mean Square Error MSE). In order to minimize this cost function, the *gradient descent* method is used. The gradient descent algorithm consists in selecting a starting point (initial guess) in which to calculate the gradient (i.e. partial derivatives) of the function in order to find a "descent direction" (negative value of the gradient), and hence move to a new point along the identified descent direction and calculate the gradient of the function in this new point (Fig. 4.9). This process is repeated until the algorithm eventually converges where the gradient is zero. When the error signal for each neuron is computed, the weights coefficients can be updated.

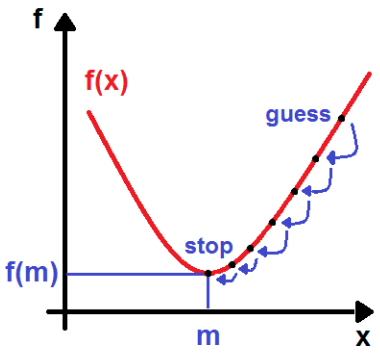


Figure 4.9: Illustration of the gradient descent method

- **Backward propagation**: the error of the network is propagated backward

from the output layer through the other layers. This is done by recursively computing the gradient of each neuron. The weights w_{ij} that connect the neuron i to the neuron j are updated by using (Eq.4.4):

$$\triangle w_{ij} = \eta \cdot \delta_i \cdot x_i \tag{4.4}$$

where η is the learning rate and δ_i is calculated by using (Eq.4.5) for the output units:

$$\delta_i = (\hat{y}_j - y_j) \cdot y_j \cdot (1 - y_j) \tag{4.5}$$

and (Eq.4.6) for the hidden units:

$$\delta_i = \sum_{j=1}^{M} (\delta_j \cdot w_{ij}) \cdot y_j \cdot (1 - y_j) \tag{4.6}$$

The BP algorithm is considered to have converged whether the absolute rate of change in the MSE is sufficiently small (e.g in the range [0.1, 0.01]). The successful learning enables the model to perform properly a desired task and to generalize well, that is, the model behaves correctly on new instances of the learning task. However, the system cannot generalize in case of over-training, which can arise when the training set is too big; to avoid over-training stopping criteria can be added to the learning algorithm. The lack of generalization in the system can also arise when there are too many hidden neurons in the network and the capacity for computation exceeds the dimensionality of the input space. This is analogous to having a system of equations with more equations than free variables: the system is over specified and cannot generalize well. On the other hand, in the case where there are not enough hidden neurons in the network, the system might be unable to properly fit the input data. In machine learning there are several methods to verify the degree of generalization of the network; cross-validation is one of this methods and it consists in dividing the data contained in the training set into two mutually exclusive sets:

the training set and the test set. The training set is the larger data set used to train the model, while the test set is the smaller data set used to validate the model. This process is repeated with different subsets, until each object of the data set is used once for the test set. Furthermore, the training set might consists of data of different types that have different ranges of values which can affect the learning process. Hence, normalization can be used to scale the data either in the interval $[0, 1]$ (E.q.4.7):

$$Norm(x_i) = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{4.7}$$

or $[-1, +1]$ (E.q.4.8):

$$Norm(x_i) = \left(\frac{x_i - min(x_i)}{max(x_i) - min(x_i)} - 0.5 \right) * \ 2 \tag{4.8}$$

where x_i represents the data to be normalized and $min(x_i)$ and $max(x_i)$ are the minimum and maximum values that the data can assume over the training process.

Important parameters to be set during the learning process are the learning rate η and the momentum μ. The learning rate, that typically can assume values in the interval [0, 1], represents the step-size used in the gradient descent algorithm that affects the speed at which the algorithm converges to a minimum solution; if the learning rate value is too small the convergence of the learning algorithm is extremely slow, while if it is too large the algorithm might not converge. The momentum, that can assume values in the interval [0, 1], is used to prevent the learning algorithm to converge to a local minimum and to speed its convergence.

One of the drawback of the back-propagation algorithm is that it requires a continuous supervision and it could converge to a local minimum.

4.3 Machine Learning and Data Analysis

The application of machine learning algorithms generates data that can be analysed quantitatively and qualitatively in order to understand them. The proper analysis

of collected data is fundamental for discovering patterns that can answer important research questions. Indeed, when neural networks are designed to model some cognitive functions it is fundamental to understand how networks operate; this requires to examine the structures of the network's internal representations. The analysis of internal representations of a neural network model is a complex task, considering that the weights learned by the model are usually difficult to be interpreted. The traditional techniques used in this framework are hierarchical clustering and principal component analysis [Bullinaria, 1997].

The Hierarchical Cluster Analysis (HCA) of each input patterns permits the mapping of the activity of the hidden layer by creating a hierarchy of clusters based on a selected distance measure (e.g. Euclidean distance). In case a neural network model is operating efficiently, the cluster analysis reveals that related input-output patterns are closely clustered (i.e. low Euclidean distance between related patterns). This approach can identify interesting relations between data that might be not so obvious otherwise.

Considering that the hidden layer of neural network based models usually represents a multi-dimensional system, the visualization of such points and the analyse of the trajectories between states by using traditional techniques is difficult to be performed. To this end it is possible to perform the Principal Component Analysis (PCA) of the hidden layer that permits to perform dimensional reduction with the minimum loss of information. This procedure was used by Elman [1993] to reduce the dimensionality of the hidden units and then to construct the phase state graph of the principal components.

EXPERIMENTAL STUDIES

Overview

This section intends to provide an introduction to the performed experimental studies. The aim of the experiments proposed in the next chapters is to find a general mechanism that enables the grounding of abstract action words through sensorimotor experience in a humanoid robot. To this end, a number of neuro-robotic models, that permitted the investigation of the relations between symbolic knowledge (i.e. *language*) and sensorimotor experience (i.e. *perception* and *action*) were implemented.

In contrast to purely computational modelling methods and classical natural language processing methodologies, in the approach adopted for carrying out these studies language is considered to be embodied in perceptual and sensorimotor knowledge. Hence, cognitive humanoid robots provide a powerful platform for testing the design of artificial cognitive architectures that integrate perception and action, capable of autonomous learning, decision-making and communication.

In the performed experiments two different sets of words were taught to a humanoid robot; in the first two experiments presented in this thesis, the robot is trained to learn words related to general actions (e.g. "ACCEPT" and "REJECT"). In the third experiment, in addition to the name of general actions (e.g. "USE" and "MAKE"), the name of objects/tools (e.g. "KNIFE", "HAMMER", "BRUSH", etc.) used during interaction in the environment, were taught to the robot. Indeed, the linguistic instructions provided to the robot consisted of action and object

names.

In the next chapters three experiments will be presented:

- Chapter 5 presents a feed-forward model for the encoding of higher-order words (e.g. "ACCEPT" and"REJECT") as integration of motor primitives. The training of this model was effective for teaching a humanoid robot the meaning of words that lack of a direct concrete referent, although some limitations of the model were evident. Considering that this model was based on a feed-forward architecture, the activation of the action primitives could not be temporally specified. A temporal specification of actions implemented in the second model presented in Chapter 6 permitted the increase of the combinations of actions for the generation of more complex movements. Furthermore, in this model simplified representations of actions were used as input/output to the neural controller. The activation of a one-hot node resulted in the robot's execution of pre-determined actions.

- Chapter 6 presents a recurrent model that extends the feed-forward architecture for the encoding of higher-order words presented in Chapter 5. In this model the sequences of linguistic inputs, temporally specified and consisting of verbs only, led to the acquisition of higher-order concepts (e.g. "ACCEPT", "REJECT") grounded on basic motor primitives (e.g. "PUSH", "PULL"). Higher-order symbolic representations were indirectly grounded in action primitives directly grounded in sensorimotor experience.

- Chapter 7 presents a recurrent model that integrates multi-modal inputs (i.e. language, vision and proprioception) and that takes into account a more realistic representation of the sensorimotor knowledge associated to the iCub robot. More complex actions (e.g. "CUT", "HIT", "PAINT", etc.) were built by integrating low level motor primitives (e.g. "PUSH - PULL", "LIFT - LOWER", "MOVE LEFT - MOVE RIGHT") iterated for a certain number of time steps. The simplified representations of actions ("one-hot" encoding)

used in the other two experiments were replaced with the joints values recorded from the iCub robot right arm. Furthermore, the new model was scaled up to handle a larger action repertoire resulting from different combinations of joint activations, and the visual input captured from the robot's cameras has been included as an input unit of the model. Indeed, in this model the acquisition of lexical categories is achieved by integrating three different modality inputs: proprioceptive input (joint values), visual input (object features) and linguistic instructions (sentences consisting of a verb and a noun). Through the implementation of this model, the hierarchical organization of concepts directly linked to sensorimotor experience permitted the acquisition of higher-level words and categories.

Chapter 5

A Study on the Learning of Higher-order Concepts

Cognitive robots have been successfully used for the learning of concrete concepts and lexicons [Cangelosi and Riga, 2006, Cangelosi et al., 2006, Sugita and Tani, 2005, Yamashita and Tani, 2008, Dominey et al., 2009]. For instance, in Cangelosi and Riga [2006] it has been shown that cognitive robots are capable of performing concrete actions and understanding each action's name. Nevertheless, building intelligent systems that can understand the meaning of abstract words is still a challenging task for cognitive developmental robotics. Abstract concepts such as "truth", "democracy", "happiness", "justice", etc. refer to intangible entities not physically defined and/or spatially constrained (e.g. mental states), which cannot be perceived through the senses [Wiemer-Hastings and Xu, 2005]. This is why grounding abstract words is still a highly challenging and problematic task in cognitive robotics.

One important property of human language, which inspired the proposed studies, is "combinatoriality" [Pinker and Prince, 1988], that is the possibility of producing new concepts from the combination of simple words. The process of transferring the meaning of words directly grounded in sensorimotor experience, to words generated via linguistic combinations, is called "symbol grounding transfer" [Cangelosi

and Riga, 2006], and follows the mental simulation model of concept combination [Barsalou, 1999, Cangelosi and Schlesinger, 2014].

This chapter presents the first model on the learning of higher-order abstract concepts in a robotic platform and the related results [Stramandinoli et al.]. The model, based on Artificial Neural Networks (ANNs), grounds abstract language in the iCub sensorimotor experience. The mechanism of the "symbol grounding transfer" is adopted as a training strategy of the model implemented for the acquisition of higher-order concepts. Such neuro-robotics model, exploiting the "combinatorial" property of language, enables the learning of higher-order concepts by combining words directly grounded in sensorimotor experience. The target of this study is the acquisition of the meaning of words like "ACCEPT", "REJECT", "PICK", which describe general actions. The acquisition of such higher-order concepts develops trough an incremental training mechanism. A set of basic motor primitives (e.g. "MOVE ARM AWAY", "MOVE ARM TOWARD", "OPEN HAND", "CLOSE HAND", etc.) are initially taught to the iCub through the "direct grounding mechanism"; then, new symbols related to more abstract words like "KEEP", "GIVE", "RECEIVE", are learned through the combination of the words directly grounded in motor primitives.

An extension of the proposed model, which will allows the investigation of the relations between abstract symbolic representations (i.e. language) and sensorimotor knowledge (i.e. actions), is be presented in Chapter 6.

5.1 Theoretical Background

A broad range of social psychology studies, which demonstrated embodiment effects and the tight coupling between the cognitive and motor systems, have been described in [Barsalou et al., 2003]. Some studies have demonstrated that social stimuli induce bodily states (e.g. postures, arm movements and facial expressions); in other studies it has been shown that bodily states produce emotions (e.g. a push-

ing movement associated with avoidance produces a negative affect) [Barsalou et al., 2003]. In further work it has been shown that when bodily states are compatible with cognitive states, processing is optimal; that is, processing a positive stimulus is faster when performing an approaching arm movement (e.g. pulling) than an avoidance movement (e.g. pushing) [Barsalou et al., 2003]. Embodiment effects which reflect a "pattern-completion inference mechanism" that supports situated action have been proposed in [Barsalou, 2003]. A pattern completion inference mechanism uses perception to activate situated conceptualizations that produce predictions of associated embodiment effects. According to this view, representations of familiar situations that contain embodiments become established in memory (e.g. receiving a gift, feeling positive affect, and smiling). When part of this situation occurs (e.g. receiving a gift), it activates the remainder of the situational pattern, producing associated embodiments (e.g. smiling). Similarly, if smiling is engaged, representations of situations that contain it are activated, producing associated pattern components (e.g. positive affect, generosity); thus, an agent draws inferences from the simulation that go beyond the given information [Barsalou, 2009].

Different neurophysiological studies have shown that motor simulations generate prediction about the meaning of words [Pulvermüller, 2005, Buccino et al., 2005, Pulvermüller et al., 2005, Barsalou, 2008, 2009]. Furthermore, behavioural studies have also supported the effect of physical actions in comprehension [Glenberg and Kaschak, 2002, Zwaan and Taylor, 2006, Richardson et al., 2003].

5.2 Overview of the Experiment

A model based on ANNs for grounding the meaning of abstract words in the sensorimotor experience of a cognitive robotic platform is presented. This preliminary study has been developed on a software environment for the iCub robot [Tikhanoff et al., 2011]. Words like "ACCEPT", "REJECT", "PICK", that express general actions and characterized from an evident sensorimotor component have been taught

to the robot. The meaning of such words is grounded through an incremental training mechanism; in particular, the iCub is first trained to learn a set of basic motor primitives through the mechanism of direct grounding; subsequently, the grounding is transferred from basic symbols to new ones, the latter obtained as combination of elementary words. Specifically, at the beginning of the training the simulated robot learns to perform a series of action primitives (e.g. "PUSH", "PULL", "GRASP", "RELEASE", etc.) and then, through the process of the grounding transfer, by combining action primitives, the robot acquires more abstract concepts (e.g. "KEEP", "GIVE", "RECEIVE"). The goal of this study is to show that the grounding of higher-order categories can be obtained as a combination of categories directly grounded in sensorimotor experiences.

5.2.1 Model Description

According to the embodied connectionist approach, linguistic abilities develop through the direct interaction between cognitive agents and the physical world they interact with. This study takes inspiration from the model proposed by Cangelosi and Riga [2006] in which two simulated robots, a teacher and a learner, were trained to learn a set of basic action primitives. The teacher was preprogrammed to show to the learner how to perform a set of action primitives. The training of the learner required two mechanisms; the first is the *direct grounding* of basic words, during which the agent, by observing the teacher, learns a set of basic action primitives and their corresponding name via direct sensorimotor experience. The second mechanism is the *grounding transfer* process when the grounding of basic words is transferred to higher-order words via linguistic description [Cangelosi, 2005]. In particular, the training of the robot consisted of three incremental stages:

(i) Basic Grounding (BG): the robot learns by imitation to perform basic action primitives and their corresponding names (e.g. *CLOSE_LEFT_ARM*, *CLOSE_RIGHT_ARM*, *MOVE_FORWARD*)

(ii) Higher-order Grounding 1 (HG1): combining basic action primitives (e.g. *GRAB [is] CLOSE_LEFT_ARM [and] CLOSE_RIGHT_ARM*) via linguistic description the robot acquires new words

(iii) Higher-order Grounding 2 (HG2): the robot learns high-order words through the combination of action primitives and higher-order action words (e.g. *CARRY [is] GRAB [and] MOVE_FORWARD*)

In the proposed study the robot first learns a series of motor primitives (e.g. PUSH, PULL, GRASP, RELEASE) that subsequently are combined to acquire higher-order action words (e.g. KEEP, GIVE, RECEIVE). Finally, the robot acquires new higher-order concepts (e.g. PICK, ACCEPT, REJECT) by combining motor primitives and the higher-order action words previously learned. At the end of the experiment, the robot is capable to categorise abstract symbols by experiencing sensorimotor actions.

In Section 5.3 a feed-forward neural network model that implements the learning of higher-order words in the iCub robot is presented. The linguistic input provided by an experimenter guides the autonomous organization of the robot's sensorimotor knowledge; sequences of linguistic inputs lead to the development of higher-order concepts grounded on basic concepts and actions.

5.3 Feed-forward Network for the Acquisition of Higher-order Concepts

The robot's neural network controller is a three layers feed-forward network fully connected, with a sigmoid activation function with unity slope $\lambda = 1$ (Eq.5.1).

$$f(x) = \frac{1}{(1 + e^{-\lambda x})} \tag{5.1}$$

The network has 14 input units that encode the name of the motor and action primitives taught to the robot (Fig.5.1). The hidden units consist of 8 neurons that

are fully connected with both the input and output layer (Fig.5.1). The number of neurons in the hidden layer was selected training several networks and estimating the generalization error of each of them. The output units consist of 8 neurons that encode an abstract representation of motor primitives. According to the activated linguistic input, the network selects in output which motor primitive (or a sequence of them) has to be activated in order to obtain the desired behaviour. The output of the network is the input for an iCub module that implements the execution of motor primitives.

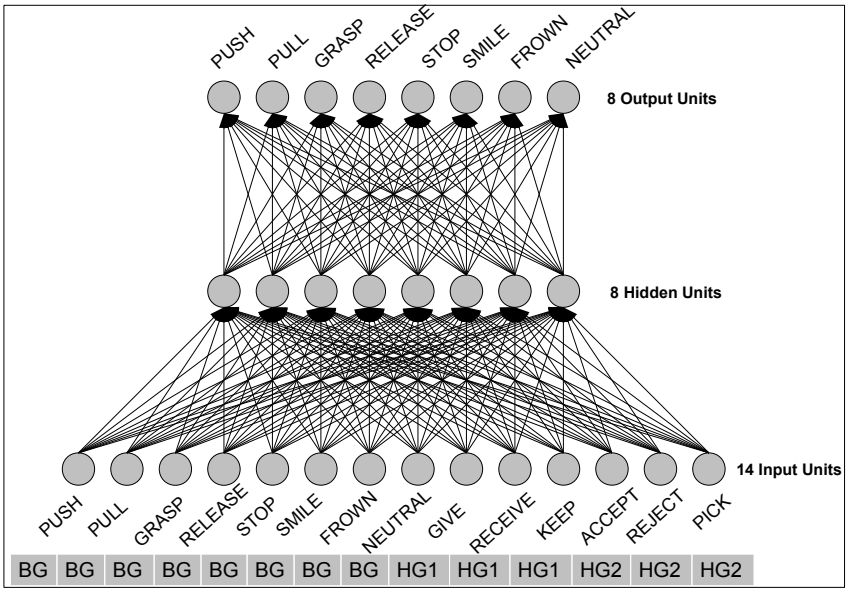


Figure 5.1: Feed-forward architecture for learning words associated to action primitives. ©2011 IEEE

The name of the motor and action primitives taught to the robot were encoded as binary vectors for which the "one-hot" encoding was adopted (See Table 6.1 for an example of such encoding).

Indeed the neural network model, developed in C++ programming language, was linked to the iCub simulator. The execution of motor primitives was implemented by using the Action Primitives library available in the iCub software repository [Pattacini et al., 2010]; this library provides a set of primitives that can be easily combined in order to obtain more complex behaviour (more details regarding the

Action Primitives library are provided in Section 3.5.2).

5.3.1 Neural Network Training

Inspired by the model presented in [Cangelosi and Riga, 2006], the training of the presented model is incremental and it consists of three 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 robot learns the names associated to the motor primitives in input to the neural network, which are "PUSH", "PULL", "GRASP", "RELEASE", "STOP", "SMILE", "FROWN", "NEUTRAL". The "STOP" word is used to make the robot understand the end of a command. The words "SMILE", "FROWN", "NEUTRAL" are intended as bodily states rather than emotional.

The network was trained by using the back-propagation supervised learning algorithm. The weights of the network were initialized to random values in the range $[\pm 0.5]$ and the back-propagation algorithm run for 10000 iterations, with learning rate (α) equals to 0.2 and momentum (β) equals to 0.9. The learning rate and momentum, in general, can assume values between the range $[0, 1]$. In the proposed simulation, a small value of α slows the convergence rate of the algorithm but helps to ensure that the global minimum is not missed. To control the convergence rate of the algorithm a small learning rate value was coupled with a bigger value of momentum; in particular, a big value of β increased the convergence speed of the algorithm. The back-propagation algorithm calculates the weight corrections that permit to reduce the distance of the actual outputs from the target outputs.

As performance function for the artificial neural network model, the Root Mean Square Error (RMSE) - the square root of the Mean Square Error (MSE) - was selected. The RMSE is defined as follows:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(e_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

where $\hat{y}_i$ is the target output and y_i is the network output.

The simulation parameters used for the training of the feed-forward network and the Root Mean Square Error value (RMSE) calculated at the end of each training stage are shown in Table 5.1.

Training Stage	**No. Iterations**	**Learn Rate**	**Momentum**	**RMSE**
BG	10000	0.2	0.9	0.005840
HG1	10000	0.2	0.9	0.005620
HG2	10000	0.2	0.9	0.005042

Table 5.1: Simulation parameters for the training of the feed-forward neural network and RMSE values. ©2011 IEEE

The BG training stage runs for 10000 iterations and, as shown in figure 5.2, after 5000 runs the value of the error is smaller than 0.02.

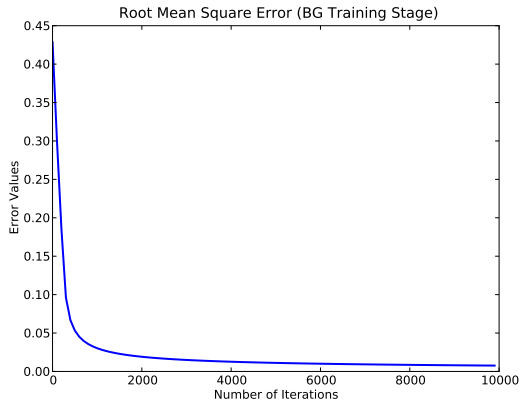


Figure 5.2: Root Mean Square Error after the BG training stage. ©2011 IEEE

The HG1 and HG2 training phases implement the grounding transfer process. During these two training stages the grounding of basic words, acquired via direct sensorimotor experience, is transferred to higher-order words via linguistic description that, in the neural controller implementation, is simplified as providing a binary vector (one-hot encoding) to the network. The grounding transfer consists of multiple steps, depending on the number of motor primitives that are combined to obtain a more complex behaviour. For example, in order to transfer the grounding from the basic actions GRASP and STOP to the higher-order word *KEEP* (i.e. *KEEP [is] GRASP [and] STOP*) two steps are required, one for each motor primitive involved

in the linguistic description. Each of these steps consists of two phases (Fig.5.3):

- the network receives as input the action primitives words contained in the linguistic description of the higher-order word and computes the corresponding output without applying back-propagation algorithm (feed-forward phase without learning).
- the network receives as input the name of the higher-order word and as target the output of the network calculated during the feed-forward phase (back-propagation learning).

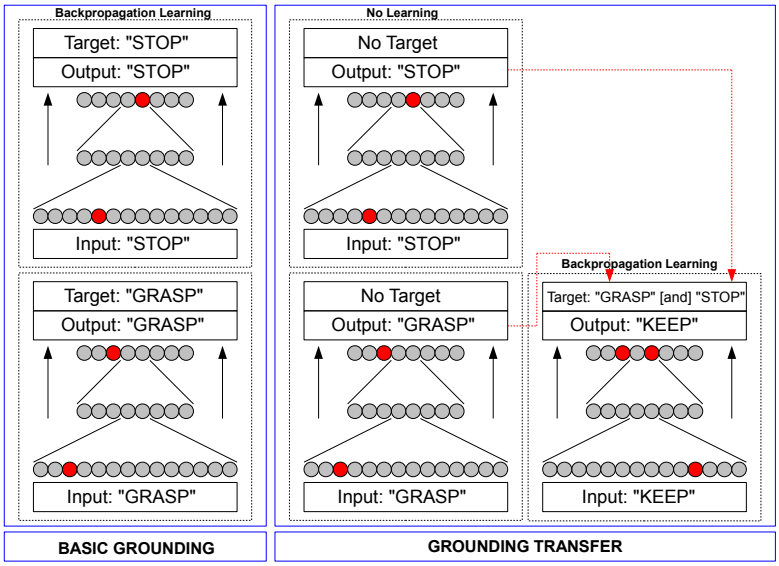
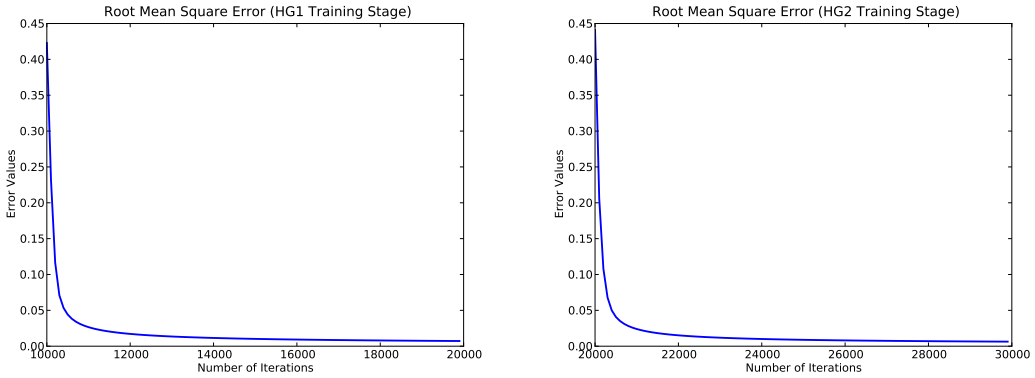


Figure 5.3: Representation of the grounding transfer mechanism. ©2011 IEEE

This described mechanism is adopted during both HG1 and HG2 training stages. In the HG1 stage the robot learns three new higher-order action words (*GIVE*, *RECEIVE*, *KEEP*) by combining only basic action primitives. In order to obtain the transfer of grounding from basic actions to higher-order words, the network calculates separately the output corresponding to the words contained in the linguistic description (*GRASP*, *STOP*) and stores it. Then, the network receives as input the higher-order word *KEEP* and as target the outputs previously stored.

The HG1 and HG2 training stages run for 10000 iterations each and, as shown in figure 5.4(a) and 5.4(b), after 5000 runs the value of the error is smaller then

0.02. During the HG2 stage, the robot learns higher-order behaviour (*ACCEPT*, *REJECT*, *PICK*) consisting of the combination of motor primitives and higher-order action words (e.g. *ACCEPT [is] KEEP [and] SMILE [and] STOP*).



(a) (b)

Figure 5.4: Root Mean Square Error after the HG1 (a) and HG2 (b) training stages. ©2011 IEEE

At the end of the training all the motor primitives, higher-order words and higher-order behaviour were successfully learned. Indeed, simulation results have shown that the network performs correctly the mapping between inputs and outputs.

5.3.2 Robot Simulation

The proposed neural network model, depending on the linguistic input received, it outputs a combination of the name of the motor and action primitives to be executed. Indeed, the output of the neural network model triggers the action primitives to be executed; such primitives were implemented by using the Action Primitives Library [Pattacini et al., 2010] available in the iCub software repository. The library provides a set of functions for the execution of actions that can be combined to perform more sophisticated tasks; it relies on the YARP Cartesian Interface [Metta et al., 2006] that 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 that allows to move the robot arm to a desired position (inverse kinematics), the library uses a non-linear optimization

technique implemented by the IpOpt software package [Wächter and Biegler, 2006].

A software module that relies on the YARP middle-ware and the Action Primitives library has been developed in order to execute action primitives associated to words. The output of the neural network module selects which action primitive has to be executed in order to perform the desired higher-order behaviour (Fig.5.5).

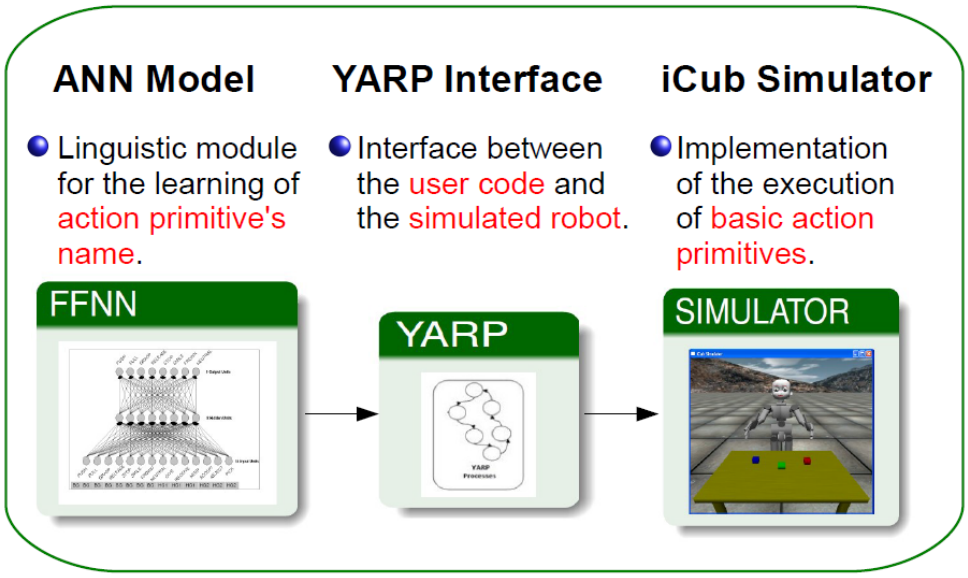


Figure 5.5: Software architecture for the learning of words: Neural Network controller, YARP interface and iCub Simulator

Some of the action primitives implemented are shown in figure 5.6; for example, figure 5.6(b) shows the PUSH primitive. Considering that the simulated iCub and the real one have the same software interface, the reproduction of the experiments with the physical robot does not require any particular modification of the code linking the neural network with the real robot (though extra work is required to handle with visual input stream and motor performance).

For the implementation of the iCub facial expression (i.e. SMILE, FROWN and NEUTRAL), the Face Expressions application available in the iCub software repository was used. The iCub head has an expression system that consists of LEDs for the display of facial features (LEB - the Left Eyebrow subsystem, REB - the Right Eyebrow subsystem and M - the Mouth subsystem) and a servomotor (EL subsystem) for the activation of the eyelid movements. To send commands to the

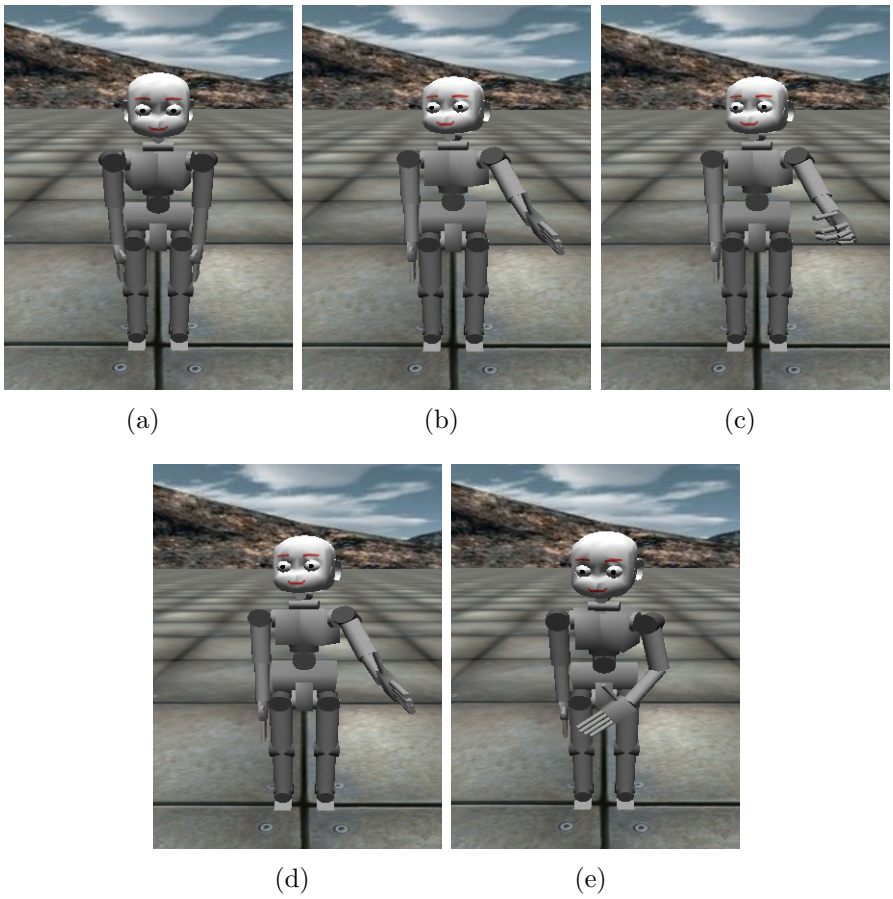


Figure 5.6: Execution of basic action primitives on the iCub: home position (a), PUSH (b), GRASP (c), RELEASE (d) and PULL (e). ©2011 IEEE

iCub head expressions system a low-level interface can been used. Such interface receives commands in the format of ASCII characters, sent over a serial connection, which define the state of the individual subsystems. The interface can be operated at the low-level through a YARP write console that permits to set a specific face expression by sending a string of the ASCII characters in the following format:

S30 (‘S’ for the eyelids - servo)

L02 (‘L’ for the left eyebrows)

R02 (‘R’ for the right eyebrows)

M64 (‘M’ for the mouth)

The numbers that follow the letters S, L, R and M indicate the led ports to be

turned on to display the desired face expression.

5.4 Discussion

In this chapter one of the first attempt to model the grounding of higher-order words in the iCub humanoid robot was presented. The higher-order words used in this study have a general meaning; they refer to the way in which objects can be manipulated and describe their motion (e.g. grasp, receive, etc.). A model based on feed-forward artificial neural networks was used for the grounding of higher-level concepts obtained as a combination of simple motor primitives directly grounded in sensorimotor experience. Such model produced effective results in teaching a humanoid robot the meaning of higher-order words, although some limitations of the model were also evident; in particular, the activation of the action primitives could not be temporally specified. A temporal specification for action executions is not only important for the control of the robot; it also permits to increase the combinations of actions in order to generate more complex movements. This in turn directly affects on the number of meanings that can be specified for different words. For example, in this model is not possible to distinguish 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.

As an extension of this preliminary study, in Chapter 6 the implementation of a recurrent model enabled the temporal specification for action executions that permitted to increase the combinations of actions in order to generate more complex movements. Furthermore, words characterized by a major level of abstractness are grounded in the experiments presented in Chapter 7. In the current study the neural network controller receives linguistic inputs and outputs a combination of the name of action primitives to be executed. In Chapter 7 this model is extended in order

to include in the neural network controller the encoding of motor outputs and to control the motor behaviour of the robot.

Chapter 6

Learning Higher-order Concepts through Temporal Sequences of Motor Primitives

In this chapter a neuro-robotics model based on artificial neural networks that investigates the relations between the development of symbol manipulation capabilities and sensorimotor knowledge in the iCub humanoid robot is presented. To overcome the limitations of the feed-forward model in terms of time specification and combinatorial ambiguity proposed in Chapter 5, a new model based on recurrent neural networks was implemented. Recurrent neural networks have been used since the beginning of the connectionist era for addressing language related research [Hinton and Shallice, 1991, Elman, 1990]; they 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 modelling short term memory in artificial systems. In this framework, the use of recurrent neural networks permits the learning of higher-order concepts based on temporal sequences of motor primitives.

6.1 Neural Network Architecture

The proposed neural network model (Fig.6.1) takes inspiration from the architecture discussed in [Botvinick and Plaut, 2006]. In preliminary tests, this architecture produced more stable and reliable learning results than other network topologies 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 the ones adopted in the feed-forward model presented in Chapter 5.

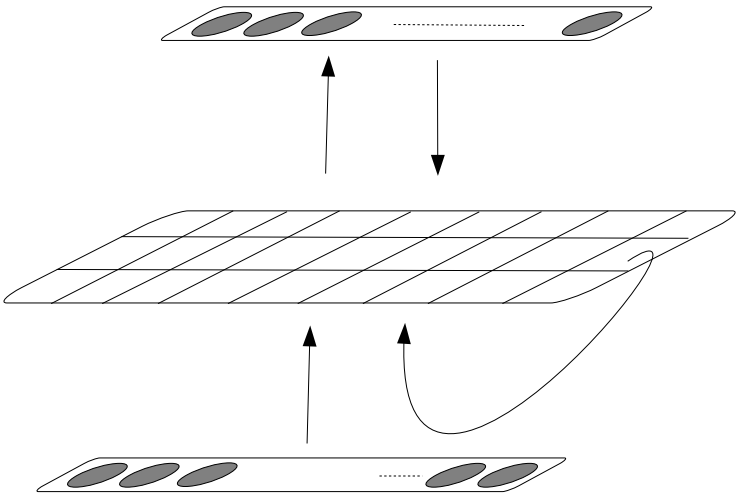


Figure 6.1: Recurrent architecture for the learning of higher-order concepts.

The input of the network is a localistic encoding of 13 words (13 input units); each word is represented as a binary vector for which the "one-hot" encoding was adopted. The output of the network is a localistic encoding of 7 motor primitives (7 output units); each action is represented as a binary vector for which the "one-hot" encoding was adopted. Similarly to the feed-forward model presented in Chapter 5, a simplified representation of motor primitives was adopted(See Table 6.1 for an example of such encoding). The actual execution of the actions was delegated to the Action Primitives Library [Pattacini et al., 2010] according to which action primitives were formed executing motor primitives in sequence; for example the "GIVE" action consists in executing the motor primitives "GRASP", "PUSH" and "RELEASE" in sequence. In Chapter 7 the simplified representations of ac-

tions ("one-hot" encoding) were replaced with the joints values recorded from the iCub robot right arm. The hidden layer of the proposed architecture consists of 27 units. The number of neurons in the hidden layer has been selected training several networks and estimating the generalization error of each of them. The input layer is connected to the hidden layer, which in turn is connected 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.

6.2 Training of the Model

The training set of the network, differently from the feed-forward architecture presented in Chapter 5, consists of sequences of temporal patterns that encode the abstract representation of actions. The words that directly refer to motor primitives activate a single output pattern that represents the action to be performed by the robot in response to a verbal command. Higher-order words activate sequences of motor primitives (Tab.6.1).

Similarly to the feed-forward model, the training of the recurrent neural network is performed by means of the back-propagation algorithm described in Section 4.2. The back-propagation algorithm used for the training of the recurrent artificial neural network-based model was extended from the incremental to the batch mode. Through the training in batch mode, all the inputs in the training set were applied to the network before the weights were updated; hence, all weight updates were summed over the presentation of the whole training sequences and subsequently, the accumulated weight updates were performed.

The formation of higher-order concepts, which refer to words whose meaning is obtained as a combination of motor primitives, is obtained by using the training methodology presented in Section 5.3.1. After the neural network learns the associations between basic grounding words and motor action primitives, the following stages that lead to the acquisition of combinatorial meaning are performed on the

model:

- the network receives as input the action primitive words that form the linguistic description of the higher-order word that the robot has to learn (e.g., *GIVE [is] GRASP [and] PUSH [and] RELEASE*).

- the motor outputs corresponding to the action primitive 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 in order to form a sequence of primitive actions (note that the sequence, *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; hence back-propagation is applied to minimize the distance of the input from the output target.

Following this approach, the meaning of words 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 are expected to create different temporal patterns according to the different motor actions that define the "meaning" of a given word. In Table 6.1 the encoding of some of the words in input to the model and the abstract representation of motor outputs are shown.

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

Table 6.1: Training set sample corresponding to the higher-order word REJECT for the recurrent neural network model

The words associated to motor primitives, learned during the basic grounding stage, and the linguistic descriptions used during the higher-order HG1 and HG2 training stages for grounding the meaning of higher-order words are shown below:

1. Basic Grounding words (BG):

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

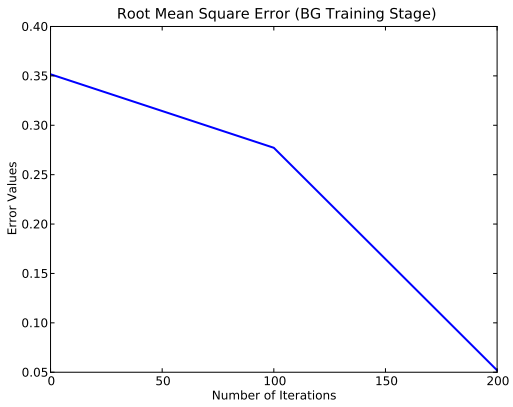
REJECT [is] GIVE [and] FROWN

KEEP [is] PICK [and] NEUTRAL

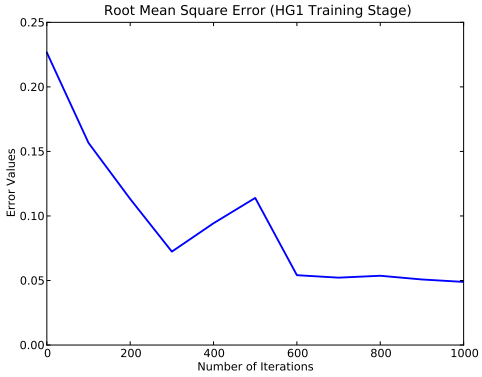
This training methodology is extremely flexible and permits to freely add novel words to the known vocabulary of the robot, or to completely rearrange the word-meaning associations.

6.3 Simulation Results and Observations

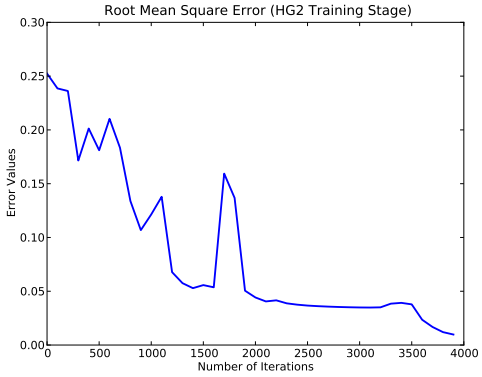
As described in Sections 5.3.1 and 6.2, the training mechanism of the network consists of three incremental stages. Figures 6.2(a),(b),(c) show the Root Mean Square Error (RMSE) calculated at the end of each of these training stages.



(a)



(b)



(c)

Figure 6.2: 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 patterns; hence, as it can be observed from figure 6.2(a), the network is able to learn this mapping in few iterations (i.e. 200). The HG1 and HG2 training stages require more training cycles, considering that in these stages the task is much more complex than the mapping learned during the BG stage. Indeed during the HG1 and HG2 stages, the network has to learn the mapping of single input patterns corresponding to the higher-order words that have to be learned, with the entire sequences of temporal

motor primitives, 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 figure 6.2(b),(c).

After the training, tests performed on the simulated iCub robot showed that the neural controller is able to correctly select and activate the proper sequence of motor primitives in relation to a word given in input. In the current model, the implementation of the mechanism for words meaning acquisition takes inspiration from the Perceptual Symbol Systems (PSS) theory proposed in [Barsalou, 1999]. Indeed, during the HG1 and HG2 stages the robot constructs higher-order concepts (e.g. *GIVE*) by reactivating the model internal representations 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 higher-order concepts are directly grounded on the basic motor primitives that constitute the meaning of the basic words [Cangelosi and Riga, 2006].

In order to better understand the internal organisation of the network and its dynamics, the activation of internal units in time of the model have been analysed. Since the neural network creates a hierarchical structure of meanings based on the combinations of basic concepts, the expectation was that similar internal representations would have been 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. For the proposed model, in order to show that similar hidden units patterns where activated according to similar primitive actions (a kind of pre-motor activation), a cluster analysis on the internal activations was performed (Fig.6.3). The results of such analysis, as shown in figure 6.3, were ineffective and showed complex internal dynamics, with very sparse clusters.

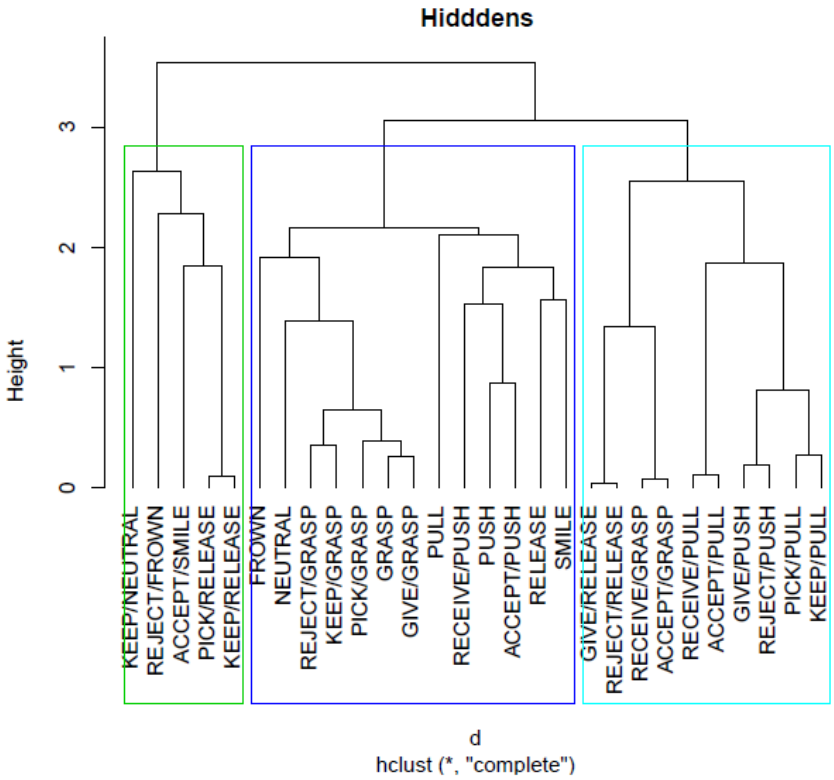


Figure 6.3: Cluster analysis of the internal activations of the model

For the formation of clusters, as measure of dissimilarity between pairs of observations, the Euclidean distance was used. Figure 6.4 shows the colormap of the cluster similarities, which is a symmetric matrix in which each element represents the dissimilarity between pairs of observations.

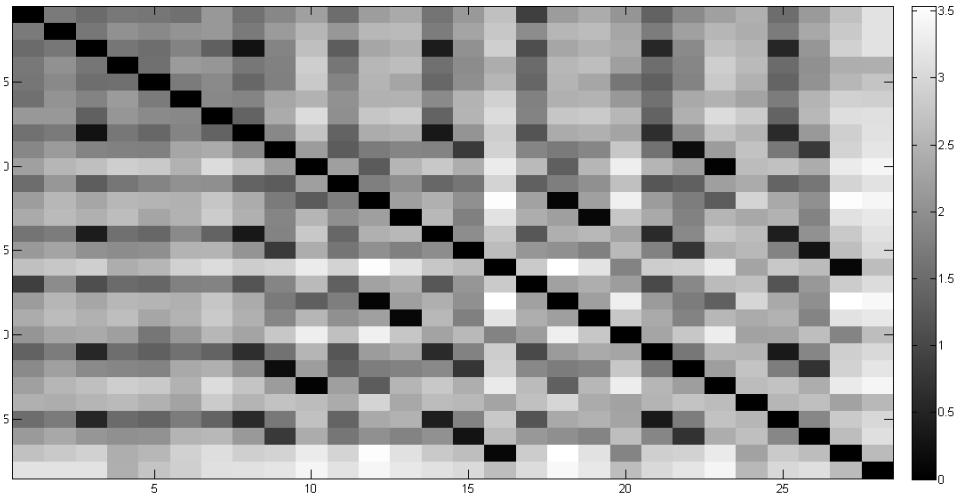


Figure 6.4: Matrix of similarities between pairs of observations

To reduce the dimensionality of the space defined by the 27 hidden units, the Principal Components Analysis (PCA) was performed on the hidden activation values in time. This used the activation patterns of elements of each sequence. Figure 6.5 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).

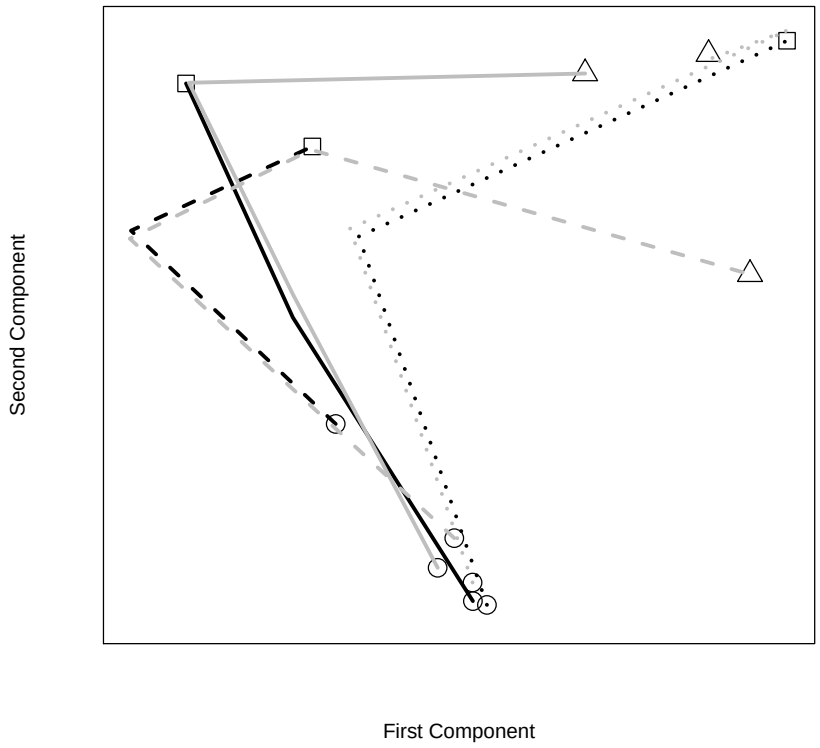


Figure 6.5: Trajectories of various patterns in time within the phase space of the first two principal components. Circles represent the starting point of a sequence while squares and triangles represent the end point of a time sequence of HG1 and HG2 levels respectively

Figure 6.5 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 the expectation, that internal representations for a given action are similar when motor patterns have similar outcomes, but different for different motor sequences.

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 [Fogassi et al., 2005]. 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 particular, Fogassi et al. [2005] studied neurons active in association with grasping movements of two monkeys. They tested two main conditions in which the monkey performed: (i) REACH, GRASP, BRING THE FOOD TO THE MOUTH sequence (the monkey ate the food) and (ii) REACH, GRASP, PLACE THE FOOD IN A CONTAINER sequence. During the second condition, the monkey was rewarded with food after accomplishing the task. The results of this test showed that the same neurons discharged differently during the "grasping for eating" and the "grasping for placing" conditions. Analysing the activation of neurons in the inferior parietal lobule (IPL) Fogassi and colleagues found that:

- IPL neurons coding a specific motor act (e.g. "grasping") had different activations patterns according to the action in which the motor act was embedded ("grasping for eating", had a different activations pattern than "grasping for placing")

- IPL neurons discharged when monkeys observed an experimenter performing the action. Neurons responded differently when the same act was embedded in different actions

- IPL neurons fired during the observation of an act and before the execution

of the subsequent action. These neurons allowed the observer to understand the agent's intention

Results of this study showed that the main factor that determines the discharge intensity of neurons is the goal of the action. Most IPL neurons code the grasping act differently according to the final goal of the action in which the grasping act is embedded. Authors carried out control experiments to check that this difference in the discharge intensity of neurons is not due to other factors like for example the force used to grasp an object or the difference in movements kinematics, or motivation.

Furthermore, neuro-computational studies have supported the results presented in [Fogassi et al., 2005]. In [Chersi et al., 2006, 2010] a computational model of neurons in the IPL area of the brain has been presented; this computational model is based on the following hypothesis:

- IPL neurons are organized in chains of simple motor acts (e.g. "reaching", "grasping", "eating", "placing") that encode a specific action with a particular goal (e.g. bringing food to the mouth, placing an object in a container)
- Chains can be constituted on motor neurons, mirror neurons, or both
- The same chain can be used for executing an action but also, by exploiting the properties of mirror neurons, for understanding an action executed by other agents

Motor acts are connected in motor chains with a specific final goal and the initial input for activating one of the motor chains is provided by the Pre-Frontal Cortex (PFC) that is believed to play an important role in action planning; the PFC contains the representation of the final goals of actions ("eating" or "placing"). The appropriate chain is selected by evaluating contextual information (e.g. visual information like the presence of a container). Starting from the evidence that the language processing of sentences that express a motor content modulates the activity

of the motor system [Pulvermüller et al., 2001], Chersi and colleagues hypothesised that the processing of action-related sentences involves the activation of the chain (i.e. motor sequence) of motor neurons directly involved in the sentence. Additional evidence suggests 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, 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 in different movements.

In order to investigate whether the model presented in this chapter shows the same dynamics of the chain model, 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, the activation values for each hidden unit were plotted as a raster matrix of 9×3 elements (Fig.6.6). Each cell of the raster matrices shown in (Fig.6.6) represents the activation value of the corresponding hidden neuron (a black cell corresponds to a neuron with activation value equals to 1, while white cells correspond to neurons with activation value equals to 0). Results of such visual elaboration, highlighting the relation between the internal representation of hidden units recorded during the Basic Grounding (e.g., *PULL*) and the internal representation of the same concept embedded in high-level words (such as *RECEIVE*, *ACCEPT*, *PICK* and *REJECT*) are shown in (Fig.6.6(a)) from which it is possible to observe that, by visually comparing the representations recorded during the BG and HGs stages, the former are very often quite different from the others, and only a small fraction of neurons is activated similarly in all the cases. This is different in case of words that share part of their 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*, it is possible to notice that, although some of the activations are in common, the two representations

appear quite different.

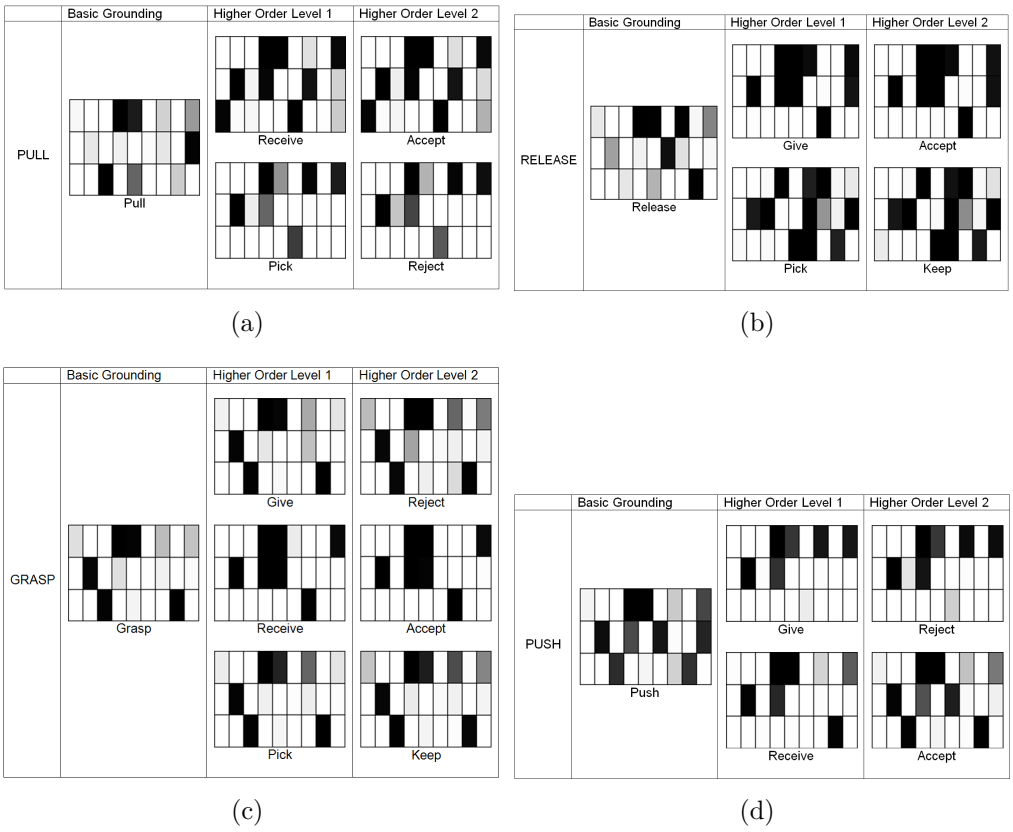


Figure 6.6: Visual elaboration of activation values of the hidden units as a matrix of 9×3 elements

These observations provide indication that the hypothesis formulated by Chersi et al. [2006] can be a general mechanism that explains the way in which recurrent neural networks represent and reuse hierarchical concepts.

6.4 Discussion

In this chapter a neural network controller for investigating the relations between higher-order symbolic representations and sensorimotor knowledge in the iCub robot has been presented. The neural network controller, based on recurrent networks, enabled the learning of higher-order concepts based on sequences of low-level primitives. Indeed, differently from the model presented in Chapter 5 based on a neural network architecture corresponding to a simple feed-forward multi-layer perceptron (MLP) that did not consider temporal feedbacks and with a hidden layer and sig-

moid activation, in this chapter in order to capture temporal dependencies among sequential data recurrent neural networks were used.

Simulation results showed that higher-order symbolic representations can be indirectly grounded in action primitives, which are themselves directly grounded in sensorimotor experience. Through the analysis of the network dynamics for the proposed recurrent architecture it has been observed that motor primitives show different activation patterns according to the action's sequence in which they are embedded; that is, for example the motor primitive "PUSH" has different activation patterns according to the action sequences that has to ground. These simulation results are consistent with empirical neuroscience and computational neuroscience studies on action representation that showed that the goal of an action changes the substrate of neurons involved in the action processing [Fogassi et al., 2005, Chersi et al., 2006, 2010].

In Chapter 7 a more realistic representations of the perceptual and sensorimotor knowledge is included in the proposed model. Instead of using abstract representations of actions, the output of the new model directly controls individual joints of the robot degrees of freedom. Hence, the model can be easily scaled up to handle a large action repertoire, resulting from various combinations of joint activations.

The proposed neuro-robotic modelling approach, which enables the learning of hierarchical higher-order representations based on combination of sensorimotor primitives, can be used to investigate the sensorimotor bases of abstract concepts. This can support the understanding of the incremental contribution of embodied knowledge in the continuum between concrete words (e.g. push, pull), which are directly grounded in actions and perceptual experience, and abstract words (e.g. use, make), for which the sensorimotor grounding is based on an indirect grounding mechanism.

Chapter 7

Grounding Abstract Action Words through the Hierarchical Organization of Action Primitives

Building on the premise that the brain contains modal symbols and representations, which are directly related to the perceptual states that produce them and which work together to create cognition, this chapter presents an embodied multi-modal robotics model that enables the grounding of abstract action word meanings. In particular, the focus of the presented study is on the modelling of the grounding of words as "USE" and "MAKE" in perceptual and sensorimotor experience developed during the interaction of a humanoid robot (i.e. iCub) in the real world. The scope of the presented study is twofold; on the one hand, the carried out study enables the iCub to ground the meaning of abstract action words and scaffold more complex behaviour through the sensorimotor interaction in the environment. On the other hand, the proposed model permits the investigation of the relation between the development of conceptual knowledge (i.e. language) and perceptual and sensorimotor categories (i.e. perceived objects and execution of actions) acquired by the iCub humanoid robot. Indeed, the implementation of an embodied computational model enables the first grade of language development (lexicon acquisition) in the iCub robot and the

investigation of the relations between embodied sensorimotor categories (continuous domain) and the representation of lexicon (discrete/logical domain).

Among the different lexical categories (i.e. noun, verb, adjective, adverb and preposition), abstract action words represent a class of terms distant from immediate perception that describe actions (i.e. verbs) with a general meaning and which can be referred to several events and situations [Barsalou, 1999, Paivio et al., 1968, Wiemer-Hastings et al., 2001]. As described in Chapter 2, according to the classic theory of categorisation, conceptual knowledge can be organized in categories hierarchically structured [Gallese and Lakoff, 2005]. For example, in the hierarchy "furniture/chair/rocking chair", "furniture" is a superordinate word (e.g. generalization) while "rocking chair" is subordinate to the basic word "chair". Basic and subordinate words (e.g. "chair", "rocking chair"), refer to "single" entities and they can be seen as more concrete words than the superordinate ones (e.g. "furniture"), which refer to sets of entities that differ in shape and other perceptual characteristics [Borghi et al., 2011]. Further, categories like "furniture" that do not have corresponding motor programmes for interacting with them, represent more general and abstract concepts. According to such hierarchical organization of lexical categories, abstract action words refer to higher-order and general concepts. Indeed, abstract action words, which do not have corresponding physical referents, cannot be directly linked to sensorimotor experience through a one-to-one mapping with their physical referents in the world.

The meaning of words like "USE" and "MAKE" is general and it depends on the context in which such words are used. Indeed, language is situated in the context in which it occurs [Barsalou et al., 2003]. For example, in a scenario in which a person is interacting with a set of tools, the meaning of "USE" is specified by the particular tool employed during the interaction (e.g. "USE [a] KNIFE", "USE [a] BRUSH"), while the meaning of "MAKE" depends on the outcome of interactions (e.g. "MAKE [a] SLICE", "MAKE [a] HOLE"). Furthermore, as described in Chapter 2, conceptualization is embodied [Barsalou et al., 2003, Gallese and Lakoff,

2005]; that is, concepts are formed via sensorimotor experience and through the integration of multi-modal inputs. Indeed, in the proposed study, the iCub is enabled to ground abstract action words (like "USE" and "MAKE") in perception (e.g. object categories like "KNIFE", "HAMMER", "PENCIL", etc.) and actions (e.g. sensorimotor categories like "CUTTING", "HITTING", "DRAWING", etc.) through object-body interactions in the physical environment. Linguistic instructions provided by a human tutor can guide the iCub to organize the knowledge directly grounded in perception and sensorimotor experience to derive the meaning of more abstract concepts. Hence, the acquisition of concepts that refer to abstract action words can be driven by the integration of proprioceptive and visual information. The integration of low-level capabilities (e.g. perceptual and sensorimotor skills) with multi-modal symbols, enables the hierarchical organization of concepts that leads to the grounding of abstract action words. The implementation of an embodied computational model, that accounts for the acquisition of abstract action words in the iCub, can contribute to the investigation of the relations between perception, action and language representations.

7.1 Background of the Experiment

During the process of language development, the acquisition of lexicon and of its related meanings precedes the emergence of more abstract syntactic structures which can be obtained through a gradual transition from lexical semantics [Tomasello, 2009]. Indeed, lexicon acquisition constitutes an important prerequisite for learning the syntactic structures that govern language. In contrast to other forms of communication, language is a combinatorial system that permits the conveyance of new messages and concepts by integrating simpler words together. A finite number of terms (i.e. lexicon) can be combined and permuted, according to specific structural rules (i.e. grammar) in order to convey new meanings [Pinker, 2010]. Recent evidence has suggested that the human motor system is also hierarchically organized;

that is, low level motor primitives can be integrated and recombined in different action sequences in order to perform novel tasks [Mussa-Ivaldi and Bizzi, 2000]. Collectively, these studies suggest that language and the biological motor system are based on hierarchical recursive structures that can serve to ground the meaning of language in sensorimotor experience [Cangelosi et al., 2010].

By exploiting the combinatorial organization of language and the motor system, the architecture proposed in this chapter integrates simple motor primitives and words in order to create the semantic referents of terms that do not have a direct mapping to the perceptual world [Stramandinoli et al., 2012]. The semantic referents of these words are formed by recalling and reusing the sensorimotor and perceptual knowledge grounded during previous experience and interactions in the physical environment. A "grounding kernel" of words directly linked to sensorimotor experience [Harnad, 2010], combined in hierarchical structures through language, permits to indirectly ground the meaning of abstract action words. New concepts are formed through linguistic definition alone by involving a form of higher-order concepts that are based upon the combination of simpler word representations. Such a hierarchical organization of concepts can be a possible account for the acquisition of more abstract and general words in cognitive robots.

Studies presented in neuroscience [Pulvermüller et al., 2001, Hauk et al., 2004, Tettamanti et al., 2005, Buccino et al., 2005] and the behavioural sciences [Buccino et al., 2005, Scorolli and Borghi, 2007] have demonstrated that language is embodied in perceptual and sensorimotor knowledge. According to this embodied perspective, language skills develop together with other cognitive capabilities and through the sensorimotor interaction of an agent with the environment. In the investigation and studies related to the embodiment of language in sensorimotor experience, particular attention has been given to action words (i.e. verbs referring to actions). Indeed, through electroencephalography (EEG) recordings it has been shown that action words processing causes differential activation along the motor strip in the brain, with strongest in-going activity occurring close to the cortical representation

of the body parts (e.g. hands, legs, lips) primarily used for carrying out the actions described by the processed verbs [Pulvermüller et al., 2001]. Further studies have shown that action word meanings have correlates in the somatotopic activation of the motor and premotor cortex [Hauk et al., 2004]. Moreover, transcranical magnetic stimulation (TMS) studies and behavioural experiments have shown that the processing of action-related sentences modulates the activity of the motor system and, according to the effector used in the action described by the processed action word, different sectors of the motor system are activated [Buccino et al., 2005]. More recently, a review on the sensorimotor grounding of language has been presented in Pulvermüller and Fadiga [2010]. Neuroimaging investigations have found specific motor activations when subjects understand speech sounds, word meanings and sentence structures. Furthermore, studies involving TMS and patients with lesions affecting inferior frontal regions of the brain, have shown the contributions of motor circuits to the comprehension of phonemes, semantic categories and grammar. Additionally, in Pulvermüller [2003] it has been shown that lexical and grammatical structures of language are processed by distributed neuronal assemblies with cortical topographies that reflect lexical semantics.

Psychological studies and theories along the same line of research have been proposed. According to the perceptual symbol systems (PSSs) theory, conceptualization requires the sensorimotor simulation of past experience [Barsalou, 1999]. For example, when a person thinks about an object, the neural patterns in the brain that have been formed during earlier experiences done with the object, are reactivated. The neural underpinnings of this simulation could be found in wide neural circuits that involve canonical and mirror neurons [Rizzolatti et al., 1996b]. Furthermore, the embodied theory of meanings known as the Indexical Hypothesis, holds that sentences become meaningful through grounding their interpretation in affordances [Kaschak and Glenberg, 2000]; that is, the meaning of words in sentences is emergent from the mesh of affordances, learning history, and goals [Glenberg and Robertson, 2000]. More specifically, in language comprehension studies [Glenberg and Kaschak,

2002], it has been observed that sentences are understood by creating a simulation of the actions that underlie them (Action-sentence Compatibility Effect).

Taken together, these studies suggest that conceptualization and language representations are formed through sensorimotor experience; that is, language representations are not related to abstract amodal symbols but they are grounded in perception and sensorimotor knowledge (i.e. perceptual multi-modal symbols). Despite all the aforementioned multidisciplinary studies, the interaction between language comprehension and action is not yet fully understood. The aim of this study is to create a cognitive architecture that enables the iCub humanoid robot to acquire the meaning of abstract action words, and further, that can contribute to the elaboration of a theory on the relations between perception, motor behaviours and language representations.

7.1.1 Lexicon Development and Embodied Conceptualization

Studies conducted on children's early vocabulary acquisition have shown that, when children learn to speak, they first learn concrete nouns (e.g. object's name) and then the abstract ones (e.g. verbs) [McGhee-Bidlack et al., 1991]. While concrete language refers to tangible entities characterized by a direct mapping to the perceptual world, more general and abstract terms are only indirectly related to perceptual inputs [Barsalou, 1999, Wiemer-Hastings and Xu, 2005]. This is why the problem of abstract concept acquisition cannot be simply resolved by directly linking words to the entities and concepts to which they refer. Nevertheless, the transition from highly concrete concepts to the abstract ones is gradual. That is, the categorization of concrete and abstract terms cannot be simply regarded as a dichotomy [Wiemer-Hastings et al., 2001] but there is instead a continuum in the level of abstractness according to which all words can be categorized.

Recent studies have provided evidence for supporting the idea that the conceptualization is embodied. Categorization is not just related to objective properties

of objects but also to the sensorimotor interaction with the physical environment [Gallese and Lakoff, 2005]; this means that an object is categorized not only in terms of its perceptual and visual properties, but also according to the motor programs (i.e. affordances) that can be performed on/with it. In a hierarchical organization of categories, like for example "furniture/chair/rocking chair", the category in the middle of this hierarchy, called "basic-level" category [Rosch, 1999], tends to be learned earlier and to be remembered more easily than other words in the hierarchy. The reason why this happens, as remarked in [Gallese and Lakoff, 2005, Arbib et al., 2008], is that the basic-level categories have corresponding mental images and human beings have motor programs to interact with them (while it is not the case for categories like "furniture" for example). Further support to the embodiment of concepts has been provided. As exposed in [Arbib et al., 2008], many concepts can be defined as "sit" and "chair" via the multi-modal integration of different input signals (e.g. vision, proprioception, language, etc.). Words must link to non-verbal experience, which is both perceptual (vision) and behavioural (action) [Arbib et al., 2008]. Along the same line of research, different studies ranging from behavioural experiments and neuroscience to computational modelling, have investigated the integration of vision, action and language through an embodied approach. In particular, in [Caligiore and Fischer, 2013], it has been suggested that vision, action and language form an integrated and dynamic system that is attuned to the constraints of its bodily implementation. Furthermore, the embodiment of cognition is supported by different behavioural studies, which have shown that seeing objects automatically activate plans for actions directed toward it [Tucker et al., 1998]; that is, the observation of an object can activate the motor activity related with it (i.e. object's affordances) [Tucker et al., 1998]. Analogous results have been observed in case of linguistic stimuli; that is, object names induce similar action planning effects as seeing the objects themselves [Tucker and Ellis, 2004, Borghi et al., 2004].

In the study proposed in this chapter, a cognitive architecture is implemented to enable the iCub humanoid to ground the meaning of abstract action words in per-

ception and sensorimotor experience. This permits to analyse the relations between objects, actions and language representations. In the experiment, the robot interacts in the environment with tools that permit the performance of goal-oriented actions. As proposed in [Gibson, 1977], for human beings tools are detached objects that afford manipulation: an elongated object of moderate size, graspable at one end and weighted at the other, if used to hit or strike, it is a hammer; a rigid object with a sharp dihedral angle and a blade that affords cutting and scraping, it is a knife; a writing tool that leaves traces when applied to surfaces and thus affords trace-making, it is a pencil [Gibson, 1977]. The affordances for object manipulation include the visual cues indicating that an object or a portion of it constitutes a suitable target for a stable grasp [Oztop et al., 2004]. Recent studies with human participants have suggested that the internal representation for a new tool used by the brain might be encoded in terms of specific past experiences which consist of brief feed-forward movement segments used in the initial exploration of the tool [Mah and Mussa-Ivaldi, 2003]. Subsequently, a tool task is solved by dynamically combining these sequences [Mah and Mussa-Ivaldi, 2003].

The studies presented in this section provided useful insights for the development of the experiment presented in this chapter.

7.2 Related Computational Models

Recently, cognitive robotics models have started to investigate some of the issues related to language development. However, attempts to model the acquisition of abstract categories in robots are in fact non-existent. Different models have focused on the acquisition of words related to objects and actions but they did not address the problem of grounding abstract categories. For example, Sugita and Tani [2005] propose a model for the acquisition of the meaning of simple linguistic commands. A mobile robot acquires the meaning of two-words sentences through the translation of linguistic commands into context-dependent behaviours. In [Yamashita and

Tani, 2008] a humanoid robot learns to generate object manipulation behaviours by a functional hierarchy which self-organizes through multiple time-scales in the neural activity of the neural network based model. In [Dominey et al., 2009] a model for the learning of a cooperative assembly task has been presented; a user can guide the robot through an arbitrary, task relevant, motor sequence via spoken commands and the robot can acquire on the fly the meaning of novel linguistic instructions and new behavioural skills by grounding the new commands in combinations of pre-existing motor primitives. In [Farkaš et al., 2012] a model for the learning of actions oriented toward objects in the iCub robot peripersonal space has been proposed; the model can generalize novel action-target combinations with randomized initial arm position and it can adapt its behaviour in case the action-target changes during motor execution. In [Kalkan et al.] the interactions of a robot with its environment have been used to create concepts typically represented by verbs in language. Authors have argued that verbs typically refer to the generation of a specific type of effect rather than a specific type of action. In the model they propose, behaviours are represented in terms of the produced effects. In [Yürüten et al., 2012] a model for the learning of adjectives and nouns from affordances has been presented; the iCub humanoid robot is enabled to learn nouns and adjectives from sensorimotor interactions and to predict the effects of the interaction with objects (e.g. labelled as verbs). The categorization of objects in the model proposed in [Yürüten et al., 2012], is done in terms of the functional view of the object rather than in terms of objects appearance. All the presented models focused on the learning of different lexical categories (e.g. adjectives, nouns and verbs) which can be directly mapped into physical referents in the real world (i.e. concrete concepts).

This chapter presents a novel embodied cognitive robotic model for the grounding of abstract action words through the multi-modal integration of different input signals (i.e. vision, proprioception and language). In particular, a concept like "USE" has been defined in terms of the actions that can be performed with selected tools (e.g. "CUT" and "KNIFE" or "PAINT" and "'BRUSH", etc.). Therefore, the

grounding of abstract action words has been achieved by linking non-verbal knowledge, both perceptual (vision) and behavioural (action), to words [Arbib et al., 2008]. The proposed model represents the first attempt in grounding the meaning of general words in perceptual and sensorimotor experience.

By exploiting the hierarchical recursive structures, observed in both language and the biological motor system, the implementation of an embodied computational model permitted to ground the meaning of abstract action words through the hierarchical organization of motor primitives and perceptual knowledge. The study aims the investigation of how compositional actions and symbol manipulation capabilities can be integrated to bootstrap higher-level language representations. In the proposed model motor primitives, integrated and hierarchically organized, enable the execution of more complex behaviour and therefore scaffold the emergence of higher-level capabilities. In such scenario, sequences of linguistic inputs, provided by an external experimenter to guide the organization of the robot's knowledge, can be interpreted in terms of the robot internal language and motor repertoire. This leads to the development of higher-order concepts grounded on simple words and action primitives. In the proposed framework, the learning and representation of compositional lexicon and its integration with embodied sensorimotor categories, developed during object-body interactions, is fundamental for bootstrapping the process of language acquisition. Novel lexical terms can be continually acquired throughout the course of the robot's development, during new sensorimotor interactions with the environment, through linguistic descriptions.

As an extension of the model presented in Chapter 6, a neural network model that takes into account the sensorimotor features of the iCub robot was implemented. In the model presented in Chapter 6 sequences of linguistic inputs, consisting of verbs only, led to the development of higher-order concepts (e.g. "ACCEPT", "REJECT") grounded on basic motor primitives (e.g. "PUSH", "PULL"). Higher-order symbolic representations were indirectly grounded in action primitives directly grounded in sensorimotor experience. Simulation results have shown that motor primitives have

different activation patterns according to the action's sequence in which they are contained. By exploiting the results presented in Chapter 6, for the implementation of the new model more complex actions (e.g. "CUT", "HIT", "PAINT", etc.) were built by integrating low level motor primitives (e.g. "PUSH - PULL", "LIFT - LOWER", "MOVE_LEFT - MOVE_RIGHT") iterated for a certain number of time steps. Additionally, in the architecture proposed in this chapter, more realistic representations of the sensorimotor inputs were included. Abstract representations of actions ("one-hot" encoding binary vectors) used in Chapter 6, were replaced with the joints values recorded from the iCub robot right arm. Furthermore, the new model has been scaled up to handle a large action repertoire resulting from different combinations of joint activations, and the visual input captured from the robot's cameras has been included as an input unit of the model. Differently from the previous architecture, in this experiment the execution of actions required the interactions with a number of objects/tools (e.g. "KNIFE", "HAMMER", "BRUSH", etc.) and the linguistic instructions provided to the robot consisted of action and object names. Indeed, in the model proposed in this chapter the acquisition of lexical categories is achieved by integrating three different modality inputs: proprioceptive input (joint values), visual input (object features) and linguistic instructions (sentences consisting of a verb and a noun). Through the development of this study, the hierarchical organization of concepts directly linked to sensorimotor experience permitted the acquisition of higher-level words and categories.

7.3 Model Description

According to embodiment, intelligence and mental processes are deeply influenced by the structure of the body and by motor abilities. Therefore the integration of a neural network model into a robotic platform can be beneficial for enabling the process of the grounding of language. In this study, partial recurrent neural networks (RNNs) were used to model the mechanisms underlying motor and linguistic

sequence processing in the iCub robot. The use of RNN enabled the learning of higher-order concepts based on temporal sequences of motor primitives. Indeed, the network was trained with dynamical sequences of I/O patterns which allow the robot to learn actions that develop in time (temporal sequences) through the tuning of the neural network parameters (connection weights). The proposed architecture, based on a 3-layer Jordan simple recurrent neural network [Jordan, 1986], is presented in figure 7.1.

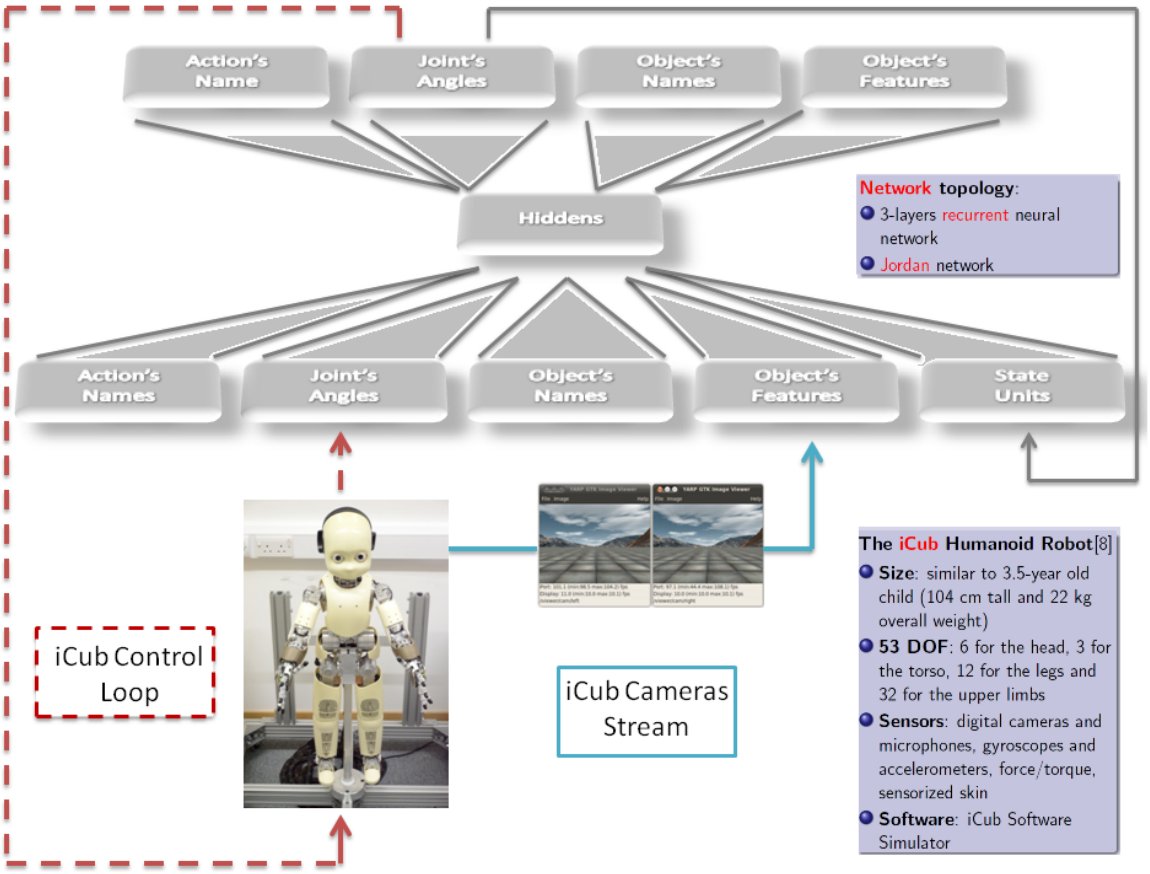


Figure 7.1: Illustration of the implemented multi-modal neural network model

A Jordan network, which has feedback connections from the output to the input units, is a discrete-time RNN in which the processing occurs in discrete steps and the relation between input/output units is governed by a functional equation that can be either linear or non-linear. In a Jordan network, activations of the output units of the network at time $t-1$ are available to the input units at time t (through the state units), via connections which may be modified during the training. The feedback of the output neurons allows the network's input units to see the previous

output, and hence the subsequent behaviour can be shaped by previous responses.

Considering that language is inherently multi-modal, in the sense that it uses many input modalities linked together (e.g. sight, hearing, touch, motor actions, etc.), it follows that in the brain there is no single "*module*" for language [Gallese and Lakoff, 2005]. For example, according to this proposal, the concept "*grasp*" gets its meaning through the ability to imagine, perform, and perceive "*grasping*" [Gallese and Lakoff, 2005]. Therefore, the artificial system proposed for the acquisition of abstract action words is multi-modal and the achievement of conceptualization requires the activation of multi-modal information. The actions used to ground language are multi-modal themselves [Gallese and Lakoff, 2005]; for example, the action of "CUTTING" has both a motor component (what you do in "CUTTING") and various perceptual components (what it looks like for someone to "CUT" and what it looks like an object used to "CUT") [Gallese and Lakoff, 2005]. The proposed architecture has been conceived to receive the linguistic, visual and proprioceptive input modalities and to output words, motor responses and object representations (Fig.7.1). The visual and sensorimotor inputs have been recorded from the iCub sensors while the linguistic inputs are binary vectors for which the "one-hot" encoding has been adopted. Vision, actions and language are integrated in order to ground abstract action words (e.g. "USE", "MAKE") in perceptual and sensorimotor knowledge. The general overview of the implemented software architecture is presented in figure 7.2. The iCub robot is connected with the rest of the software architecture through the "**iCub Module**" that sends the proprioceptive input read from the iCub encoders to the "**Neural Network Controller**" and transmits the control signal in output from the "**Neural Network Controller**" to the real robot. The exchange of information between the robot and all the other software modules is done through the YARP middle-ware [Metta et al., 2006] which, supplying ports for reading/writing information, provides a useful interface between the user code and the iCub robot. The "**Object Detector**" module reads a visual stream from the iCub cameras and, classifying objects according to their features, produces the

visual input for the "**Neural Network Controller**". Additionally, the "**Object Detector**" module extracts the position of the segmented objects and send this information to the "**Head Tracker**" module that moves the head of the iCub robot to the position received on-line from the "**Object Detector**" module.

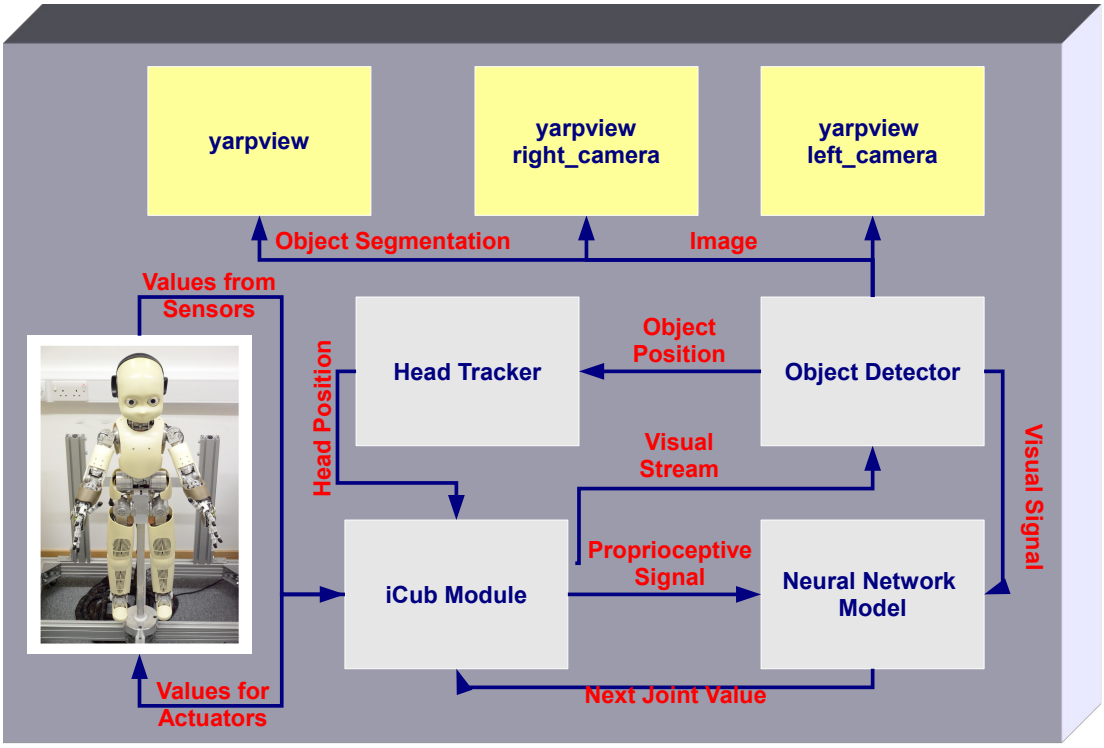


Figure 7.2: Illustration of the implemented software architecture

The visual stream read from the iCub cameras and the segmented objects are displayed through the "**yarpview**" devices, which are the image viewers provided by the YARP middle-ware [Metta et al., 2006].

7.3.1 Input and Output Coding

The input layer of the neural network model presented in this chapter (Fig.7.1) consists of five units: action's words (14 neurons), proprioceptive input (7 neurons), object's words (12 neurons), visual input (16 neurons) and the state units (7 neurons). Further details about the input layer of the network are provided below (Fig.7.1):

- **Language**: The linguistic input consists of sequences of words (i.e. verbs and nouns). The network has two units for the linguistic input; one is related to action words encoding, while the second one is for the naming of objects [Cangelosi and Parisi, 2004]. Experiments on the neural processing of verbs and nouns have shown that the left temporal neocortex plays a crucial role for nouns processing, while action's words processing involves additional regions of the left dorsolateral prefrontal cortex [Perani et al., 1999]. This is why the model was conceived with different input units for the two different word's categories (a-priori knowledge of word's classes).

- **Proprioception**: The proprioceptive signal was recorded from the iCub humanoid robot while performing the desired action primitives. The joint angles of the robot right arm were recorded and used during the sensorimotor training of the model. Additional details about the sensorimotor encoding are provided in Section 7.3.1.1.

- **Vision**: From the visual stream captured by the robot's cameras, object's features (i.e. dimension, colour and shape) were extracted. Additional details about the visual encoding are described in Section 7.3.1.1.

- **State Units**: The state units contain the activation values of the proprioceptive output units of the network at time $t-1$ that become available to the input units at time t via connections which can be modified during the training. The feedback of the proprioceptive output neurons allows the network's input units to see its own previous output, and hence the subsequent behaviour can be shaped by previous responses.

The hidden units of the model, by integrating perceptual, sensorimotor and linguistic knowledge, encode the meanings of words. The number of neurons in the hidden layer has been tuned according to the specific training stage of the network. The selected number of hidden neurons was large enough to ensure a sufficient number of degrees of freedom for the network function and small enough

to minimize the risk of loss of generalization of the network. The output layer of the network produces words associated to actions and objects, motor responses and the representation of object features.

7.3.1.1 Proprioceptive and Visual Data Set

The hierarchical structure of motor primitives is explicit and defined a-priori in order to train the robot to perform specific action sequences and acquire the desired words and action categories. The learning process in which an experimenter teaches to the robot different word categories through the physical interaction with the environment is targeted. The network is trained through a supervised learning algorithm (i.e. back-propagation); therefore, before the training can be performed, it is necessary to collect the data set for the input/output mapping. For the sensorimotor training of the iCub humanoid robot, motor primitives were planned by determining the desired end effector position in the 3D Cartesian space and then finding the joint configuration that can produce the required movements [Oztop and Arbib, 2002]. The desired task space behaviour was mapped into the appropriate joint trajectories by solving the inverse kinematics problem. The seven joint values of the iCub right arm (Shoulder Pitch, Shoulder Roll, Shoulder Yaw, Elbow, Wrist pronosupination, Wrist Pitch and Wrist Yaw) were taken into account. The inverse kinematics problem was solved using the Cartesian interface available in the iCub software repository [Pattacini et al., 2010]. The Cartesian interface determines the joint's vector $q \in \mathbb{R}^7$ of the iCub right arm in order to perform the desired movements described in terms of position $x_d \in \mathbb{R}^3$ and orientation $\alpha_d \in \mathbb{R}^4$ of the end effector. Positions and orientation refer to the root frame attached to the waist of the iCub; the orientation α_d, is represented in axis/angle notation (three components for the rotation axis and a fourth component for the rotation angle expressed in radians).

Given the position $x_d \in \mathbb{R}^3$ and orientation $\alpha_d \in \mathbb{R}^4$ of the iCub end effector:

$$
\begin{aligned}
x_d &= \begin{bmatrix} x & y & z \end{bmatrix}^T \in \mathbb{R}^3 \\
\alpha_d &= \left(\begin{bmatrix} \alpha_x & \alpha_y & \alpha_z \end{bmatrix}^T, \ \theta \right) \in \mathbb{R}^4
\end{aligned}
\tag{7.1}
$$

the joint space vector $q \in \mathbb{R}^7$ for different motor primitives is determined:

$$
q = \begin{bmatrix} \theta_{sp} & \theta_{sr} & \theta_{sy} & \theta_e & \theta_{wpr} & \theta_{wp} & \theta_{wy} \end{bmatrix}^T \in \mathbb{R}^7 \tag{7.2}
$$

By using the Cartesian interface, the encoders values of the seven joints of the iCub arm, which permitted to perform twelve different actions, were recorded. For each action, the robot's task started and ended from the same home position. Six of the twelve action primitives were iterative, while the remaining ones were non-iterative. The six iterative actions served to ground the meaning of the word "USE", while the non-iterative actions were employed to ground the meaning of "MAKE". Poses (position and orientation) associated to the twelve actions, from which the iCub arm joint values were recorded, are shown in (Tab.7.1 and Tab.7.2).

Action Name			**Position**			**Orientation**				**Object** Name
			x	**y**	**z**	α_x	α_y	α_z	θ	
HOME			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
Actions related to USE	ITERATIVE ACTIONS	CHOP	−0.24	0.16	0.0	0.12	0.76	−0.64	3.0	KNIFE
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		CUT	−0.21	0.16	0.0	0.12	0.76	−0.64	3.0	SAW
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		HIT	−0.29	0.16	0.05	0.12	0.76	−0.64	3.0	HAMMER
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		POUND	−0.29	0.16	0.08	0.12	0.76	−0.64	3.0	STONE
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		DRAW	−0.29	0.21	0.0	0.12	0.76	−0.64	3.0	PENCIL
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		PAINT	−0.29	0.24	0.0	0.12	0.76	−0.64	3.0	BRUSH
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	

Table 7.1: Poses associated to the six iterative actions from which the iCub arm joint values were recorded. The last column of the table contains the name of objects used to perform actions

Action Name			Position			Orientation				Object
			x	y	z	α_x	α_y	α_z	θ	Name
HOME			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
Actions related to MAKE	NON-ITERATIVE ACTIONS	SLICE	−0.24	0.13	0.0	0.12	0.76	−0.64	3.0	SLICER
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		SLIT	−0.21	0.11	0.0	0.12	0.76	−0.64	3.0	BLADE
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		HOLE	−0.29	0.1	0.05	0.12	0.76	−0.64	3.0	NAIL
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		HOLLOW	−0.29	0.22	0.08	0.12	0.76	−0.64	3.0	PIN
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		SCRIBBLE	−0.22	0.21	0.05	0.12	0.76	−0.64	3.0	PEN
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	
		SCRAWL	−0.24	0.24	0.02	0.12	0.76	−0.64	3.0	CRAYON
			−0.29	0.16	0.0	0.12	0.76	−0.64	3.0	

Table 7.2: Poses associated to the six non-iterative actions from which the iCub arm joint values were recorded. The last column of the table contains the name of objects used to perform actions

In determining the robot's sensorimotor trajectories and to improve the learning capacity of the model, overlapping between sensorimotor sequences was avoided.

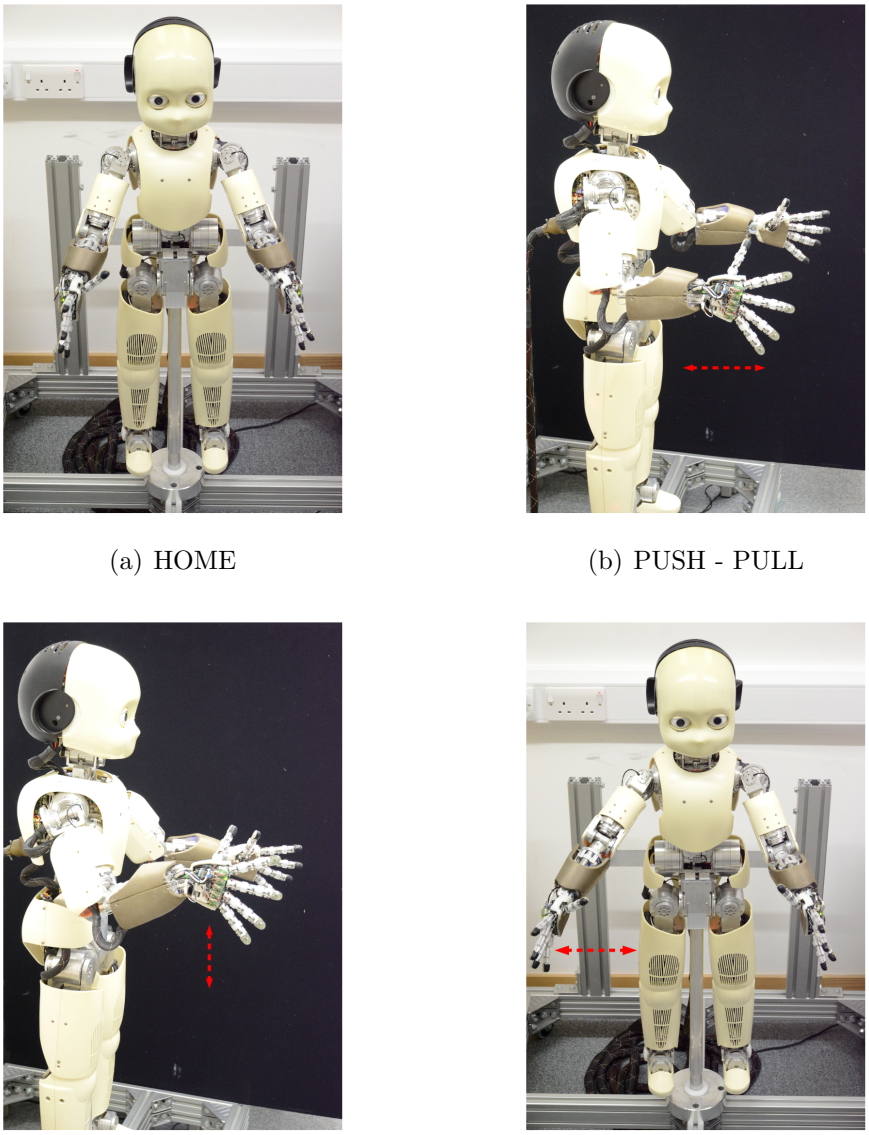
The robot could perform each of the actions with different hand configurations (e.g. precision or power grasp) which were pre-programmed. The selected hand configuration depended on the dimension of the tool employed during each task. Objects of big dimensions required a power grasp, while for small objects a precision grasp was used.

Each action was performed by changing the joint angle values from the initial configuration to the target configuration. By solving the inverse kinematics problem, joint values expressed in degrees were recorded. Before using the joint values as the training set of the network, the recorded values were scaled in the interval $[0, 1]$ using the following formula:

$$norm(j_i) = \frac{j_i - J_{min}}{J_{max} - J_{min}} \tag{7.3}$$

where J_{min} and J_{max} represent the minimum and maximum values for the joint j_i to be normalized. The recorded sensorimotor trajectories, after the normalization

in the interval $[0, 1]$, were used as teaching sequences of the model.



(a) HOME (b) PUSH - PULL

(c) LIFT - LOWER (d) MOVE LEFT - RIGHT

Figure 7.3: Illustration of some of the motor primitives taught to the iCub robot: HOME POSITION (a), PUSH - PULL (b), LIFT - LOWER (c), MOVE_LEFT - MOVE_RIGHT (d))

Each training sequence consisted of six elements which corresponded to three iterations of the same action. Each element of the action's sequences is a motor primitive (Fig.7.3). The control flow for the proprioceptive input is shown in figure 7.4 from which it is possible to observe that, after the initialization of the robot's encoders to the desired home position, the neural network model computes the new values for encoders to be sent to the robot (Algorithm 7.3.1).

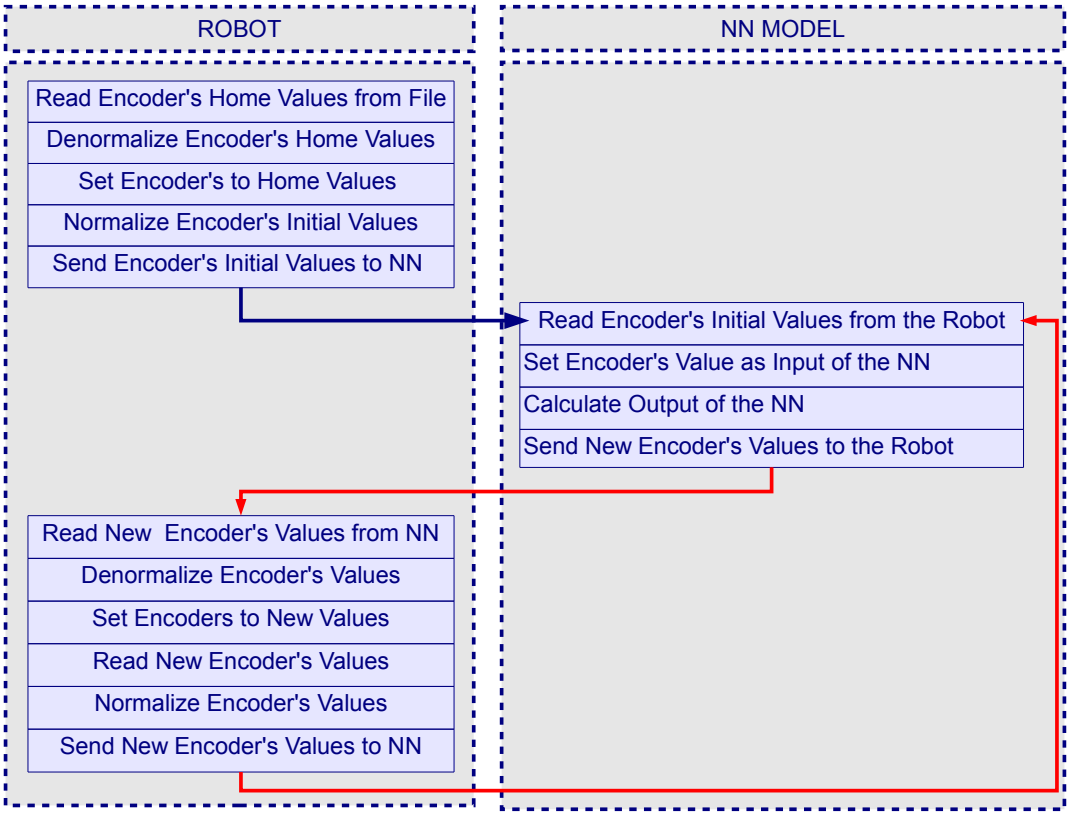


Figure 7.4: Illustration of the control flow for the proprioceptive input

The control of the proprioceptive input is described in (Algorithm 7.3.1).

Algorithm 7.3.1: CONTROL FLOW PROPRIOCEPTIVE INPUT(AN, JV)

$GIVEN$: *The encoding of action words and joint values*{AN, JV}

$OUTPUT$: *The appropriate lexical and action categories*{VC, AC}

– *Load encoding of words from file*

– *Read proprioceptive input from iCub encoders*

– *According to the current state of sensors, the linguistic input triggers the production of the appropriate output signal*

– *Calculate error (through BP)*

– *Send the output of the network (control signal) to the iCub*

return ({AN, JV})

Before sending the new encoder values to the iCub robot, the joint values were

denormalized in the original interval according to the following formula:

$$denorm(j_i) = J_{min} + norm(j_i) \times (J_{max} - J_{min}) \tag{7.4}$$

Actions were executed in presence of different objects classified using simple visual routines. The iCub robot categorized the presented objects not only according to visual features, but also in terms of the possible actions that can be carried out upon them (e.g. "CUTTING", "HITTING", etc.). In the field of neural processing of vision, according to the two-streams hypothesis [Goodale and Milner, 1992] the neural substrates of visual perception (ventral pathway) are distinct from those underlying the visual control of actions (dorsal pathway). In the proposed model the visual input is intended in terms of neural processing of vision involved with objects identification and recognition, and form representations (ventral stream). Figure 7.5 and 7.6 show the visual representations of features extracted from the objects used to perform the desired actions.

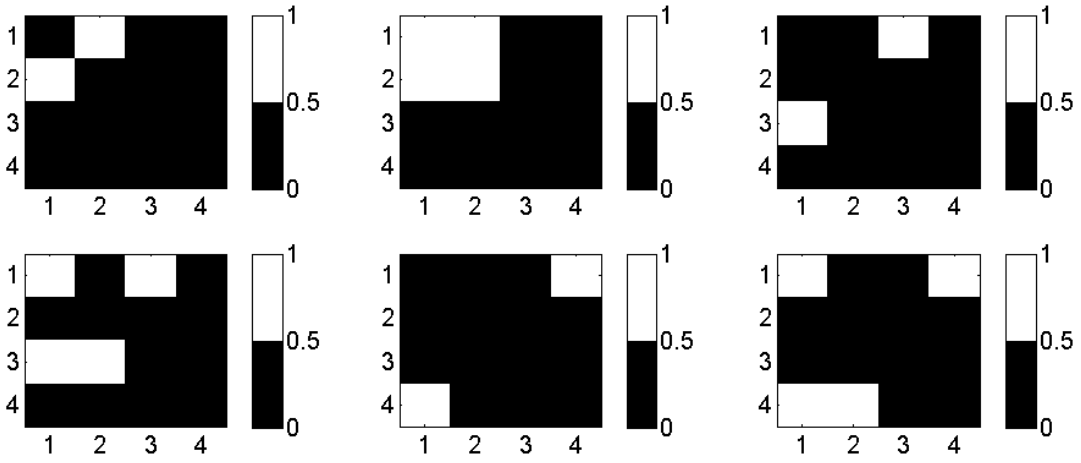


Figure 7.5: Binary matrices representing the six objects used to perform the iterative actions

Objects features are represented in a 4×4 matrix in which each value can be ether 0 or 1. The features extracted from the perceived objects were dimension, colour and shape. The first element of the matrix is related to the dimension of the object (0 for small, 1 for big objects). The second, third and forth elements of the

matrix encode the colour of the object in RGB values, while the remaining twelve elements are related to the shape of the object.

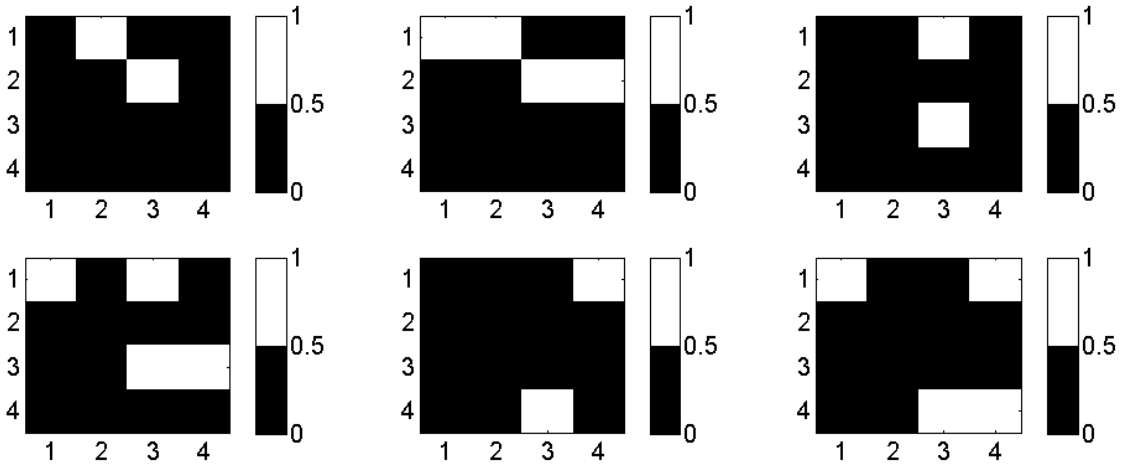


Figure 7.6: Binary matrices representing the six objects used to perform the non-iterative actions

For example, the first binary matrix in figure 7.5 corresponds to a "KNIFE" with the following features: its dimension is small (encoded as 0), its colour is red (encoded as 100) and its shape is similar to the predefined shape category 1 (encoded as 100000000000).

7.4 Robotic Task and Training Strategy

The iCub humanoid robot has been adopted as the robotic platform for this study [Metta et al., 2008]. The proposed neural network model is used to control the robot's behaviour by following commands organized in linguistic sequences. More specifically, the experiment enabled the robot to learn a set of behaviours by acting with specific tools and the associated two-words sentences consisting of a verb and a noun (Fig.7.7). Indeed, as formulated in [Arbib, 2002] the "verb-argument structure" expressing an action-object frame is a basic component of modern human languages. Additionally, according to the "Verb Island hypothesis" the child's earliest grammatical organization is verb-item specific [Tomasello, 1992]. Initially children use grammatical constructions centred on separated, individual verb items

reflecting specific core meanings. Gradually, children acquire a general construct of verb through the merging of verb islands with similar meanings and syntactic constructs [Cangelosi, 2010].

The task for the iCub robot consisted of learning to recognize a set of tools characterized by different colour, size and shape (e.g. "KNIFE", "HAMMER", "BRUSH", etc.) presented to it and perform object related actions (e.g. "CUT", "HIT", "PAINT", etc.). Subsequently the robot learned to name the objects and actions. Finally, the robot was trained to learn abstract action words guided by new linguistic sequences that the robot interpreted in terms of its own internal motor and language repertoire (Fig.7.7).

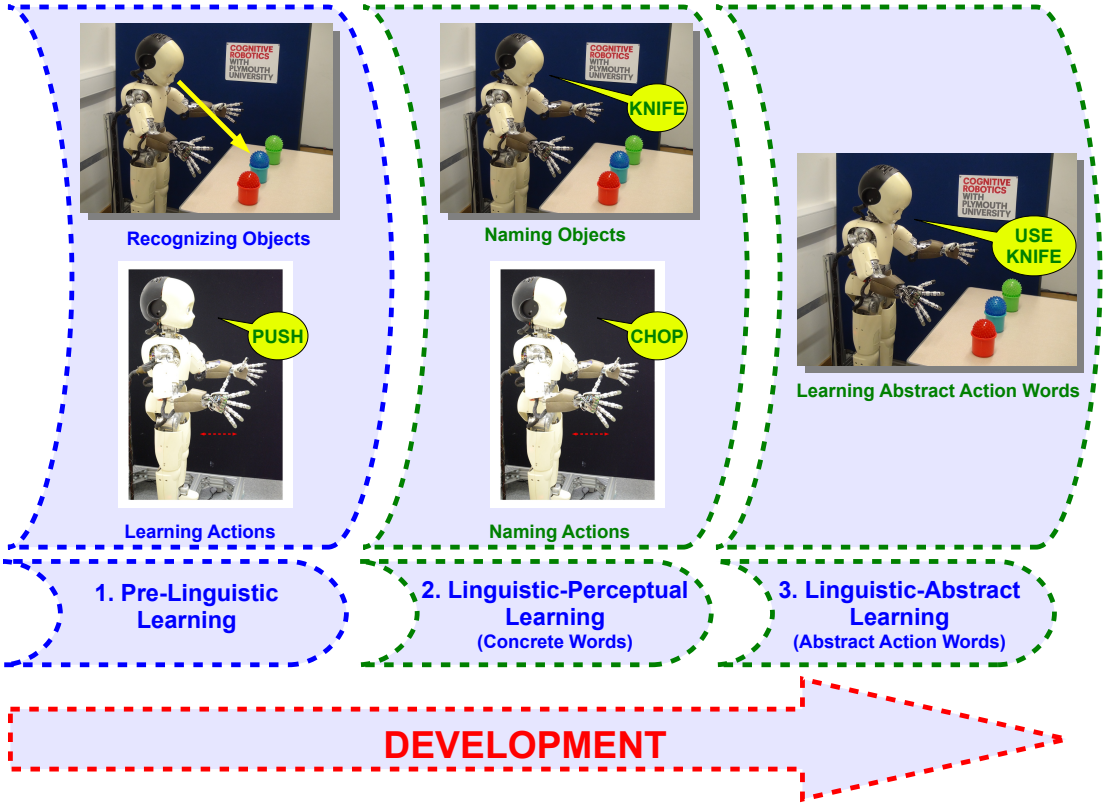


Figure 7.7: The task for the robot consists of: 1. recognizing tools and learning object related actions, 2. naming of objects and actions, 3. learning abstract action words by hierarchically organizing the knowledge directly grounded in perception and sensorimotor experience during the stages 1. and 2.

The implemented training strategy takes inspiration from developmental learning. Studies conducted in developmental psychology and neurophysiology have

revealed that perception and sensorimotor learning are pre-linguistic [Jeannerod, 1997]. That is, children acquire some motor behaviour and the capability to perceive objects before they learn to name them. Taking inspiration from these studies, the training of the architecture has been organized in three incremental stages (Fig.7.7):

1. **Pre-Linguistic Learning**: The model is trained to recognize a set of tools (e.g. "KNIFE", "HAMMER", "BRUSH", etc.) and learn object-related actions (e.g. "CUT", "HIT", "PAINT", etc.), both iterative and non-iterative, obtained by the integration of motor primitives (e.g. "PUSH", "PULL", etc.). The behaviours learned by acting with objects permit to ground the meaning of symbols in perceptual and sensorimotor experience (perceptual/sensorimotor stage). During this training stage, the neural network model learns to control the iCub arm in the joint space. The robot receives the proprioceptive input in form of target joint angles, which act as motor commands for the iCub in generating movements and interacting with the environment. Through the training process, the model learns to predict the next element in the joint sequence that permits to perform the desired behaviour.

2. **Linguistic-Perceptual Learning**: The model is trained to acquire some lexical terms through the naming of objects and actions directly grounded in perception and sensorimotor experience. This is the first stage of lexicon acquisition, when it is possible to directly link lexical terms to perceptual and sensorimotor experience. The first two stages of the training enabled the direct grounding of words into perceptual and sensorimotor inputs.

3. **Linguistic-Abstract Learning**: New words, which refer to abstract action concepts, are grounded by integrating and recalling the visual and sensorimotor knowledge that has previously been directly linked to basic concepts. In response to linguistic inputs, the model computes the corresponding behavioural patterns. Indeed, the robot learns abstract action words by receiving linguistic commands that are interpreted in terms of the robot internal motor

and linguistic repertoire. This phase of the training represents the abstract stage of language acquisition when new concepts are formed by integrating the meaning of lexical terms acquired at the previous stage of the training. At this stage the robot, guided by linguistic instructions, can organize the knowledge directly grounded in perception and sensorimotor knowledge to derive more abstract concepts. The symbol manipulation capabilities acquired by the robot permit to drive action and perceptual knowledge in order to form new concepts.

At the end of the training, semantic meanings can be gathered via lexicon organization that recalls the perceptual knowledge and motor sequences in which lexicon is grounded. In particular, the successful training of the model enables the robot to ground the meaning of words like "USE" and "MAKE" in the perceptual (e.g. "KNIFE", "HAMMER", "BRUSH", etc.) and sensorimotor experience (e.g. "CUT", "HIT", "DRAW") previously grounded. Words like "KNIFE", "HAMMER", "CUT", "HIT", etc., in the proposed hierarchical organization of lexical categories, representing basic words, are directly grounded in perceptual and sensorimotor experience through a one-to-one mapping. Words like "USE" and "MAKE", being superordinate words and referring to different events and situations, are characterized by a one-to-many mapping, that is, a single linguistic label is associated to different basic and subordinate words [Borghi et al., 2011]. Through the described training strategy the iCub robot is enabled to interpret new linguistic instructions in terms of its own internal motor and language repertoire. The hierarchical organization of concepts that the model creates can represent a useful mechanism for the acquisition and the comprehension of higher-level concepts.

The training of the neural network model has to produce an efficient classification of the inputs into different categories (Algorithm 7.4.1). Through the tuning of the neural network parameters (connection weights) the model learns to correctly classify the input signals into lexical, sensorimotor and perceptual categories. After collecting the input/target pattern sets, before proceeding with the training of the

model, it is necessary to define the topology of the network, the number of neurons in the input, hidden and output layers, the training parameters (i.e. learning rate and momentum) and to select the activation function. The training of the model was performed through back-propagation and the network performance was analysed in terms of its mean-squared error (MSE), which is the square of the average difference between the actual and the desirable output. By finding the optimal values of the network weights that minimize the difference between teaching sequences and the actual outputs, through the back-propagation algorithm, the network learned the mapping between input and output values that permitted to perform the desired tasks. In the proposed study, the back-propagation algorithm is not used for mimicking the learning process of biological neural systems, but rather as a general learning rule. Results obtained reflect characteristic features of the proposed network architecture, rather than the learning algorithm. Similar results could be obtained using other biologically more plausible learning algorithm [Yamashita and Tani, 2008]. The maximum number of iterations of the learning algorithm is 10000. In order to avoid over-training of the network, the back-propagation algorithm was terminated as soon as the error reached the threshold value of 0.001 (stopping criterion of the learning algorithm). Indeed, the back-propagation learning as possible stopping criteria includes that the total error of the network falls below a predetermined threshold value or that a certain number of epochs are completed; here a combination of the two (i.e. whichever of the two occurs first) is used. The threshold value of 0.001 was predetermined training several networks and testing the performance of each network trained. The activation function of neurons in the hidden and output layers is a logistic function defined in the interval $[0, 1]$ that permits to introduce non-linearity to the training in order to improve the convergence of the back-propagation algorithm.

The implemented model has a *simulation mode* that permits to run the algorithm either in *training* or *testing* mode. In case of training, the network's initial weights were drawn randomly from a uniform distribution $[-0.1, 0.1]$. The training of an

artificial neural network can be implemented in incremental mode or batch mode. When the incremental mode is selected, the gradient descent is computed and the weights are updated after each input is applied to the network. Through the training in batch mode, all the inputs in the training set are applied to the network before the weights are updated. For the task addressed in this Chapter, batch training demonstrated to be significantly faster and produces smaller errors than incremental training. Indeed, through the back-propagation batch learning algorithm, all weight updates were summed over the presentation of the whole training sequences and subsequently, the accumulated weight updates were performed. During each iteration of the algorithm, the accumulation of the variation of the weights were reset to zero and for each pattern set the inputs were set to zero and the state units initialised to 0.5. Hence, the new weight updates for the whole pattern set were computed until all sequences were correctly classified or the stopping criterion was satisfied (threshold on the error value). A description of the learning algorithm is given in (Algorithm 7.4.1). Given the linguistic, proprioceptive and visual inputs, the training of the model produces the categorization of the inputs into different categories (i.e. lexical, sensorimotor, and object categories).

Carrying out different simulations, it has been possible to find the network's parameters that ensured an expected training and test error as small as possible and hence a network that performed best the robotic task described in Section 7.4. Results of the performed simulations are presented in the next sections.

Algorithm 7.4.1: CLASSIFICATION OF INPUTS(AN, JV, ON, OF)

$GIVEN$: *The input pattern set* {AN, JV, ON, OF}

$OUTPUT$: *The lexical, action, and object categories*{(VC,NC), AC, OC}

– *Load network topology, training parameters and dataset*

– *Generate random seed*

if *simulation mode* is training

  then *Randomize network's initial weights* [-0.1, 0.1]

for $i \leftarrow 0$ **to** $maxCycles$

do
- *Reset delta accumulation*
- **for** $p \leftarrow 0$ **to** $patternSetSequenceSize$
  - **do**
    - *Reset all inputs of the network to* 0
    - *Initialize state units to* 0.5
    - *Learn the I/O mapping (connection weights)*
- *Update network's weights*
- $ComputeMSE$
- **if** $MSE \leq$ threshold
  - **then** *Terminate the algorithm*

return ({AN, JV, ON, OF})

7.5 Simulation Results

In order to evaluate the performance of the neural network model described in Section 7.3, different experimental scenarios were devised. Before presenting the performance and results of the implemented neural network model in the different experimental conditions, the evaluation settings are presented.

7.5.1 Evaluation Setting

The experiment was run in different training and testing conditions. Through the performance of all the training stages, the model learned the associations between words and the corresponding behavioural sensorimotor sequences and visual knowledge. During the performance of the devised tests, the capacity of the model to output the appropriate behaviours corresponding to the given linguistic instructions was verified. The implemented training strategy consisted of three incremental stages, each of which corresponded to training the model in response to different configurations of the input signals. At the end of the second stage of the training (i.e. direct naming of objects and actions), the ability of the model to generalize abstract action words was verified. Furthermore, in order to understand how the model responded to the variation of the stimuli in input and further investigate how internal representations of objects are related to action representations, the performance of the model was evaluated in response to an "incompatible condition" test. During this test condition, the provided linguistic input was either inconsistent with the objects perceived by the robot or with the actions typically associated to the objects. Through this experimental condition it was possible to verify how the robot reacted when the received linguistic command was in contrast with the perceived context. The results of this test can be helpful in understanding the mechanisms underlying positive as well as negative compatibility effects observed in behavioural experiments [Borghi et al., 2004, Tucker and Ellis, 2004]. The "incompatible condition" test was performed at the end of the second stage of the training as well as at the end of the third stage.

The collected dataset consisted of 24 sequences, half of which served for the direct grounding of basic concepts, while the rest twelve sequences were used to ground abstract action word meanings. In order to assess the performance of the model in response to different conditions, the obtained dataset was divided as described below:

- **Perceptual and Sensorimotor Mapping**. The 24 sequences were split

in two groups: 12 sequences were used for the perceptual and sensorimotor training of the model and the remaining 12 sequences were silent.

- **Direct Naming of Objects and Actions**. The 24 sequences were split in two groups: 12 sequences were used for the training (direct naming of objects and actions) and the remaining 12 sequences were used for the test of the model in order to assess the generalization capabilities of the network.

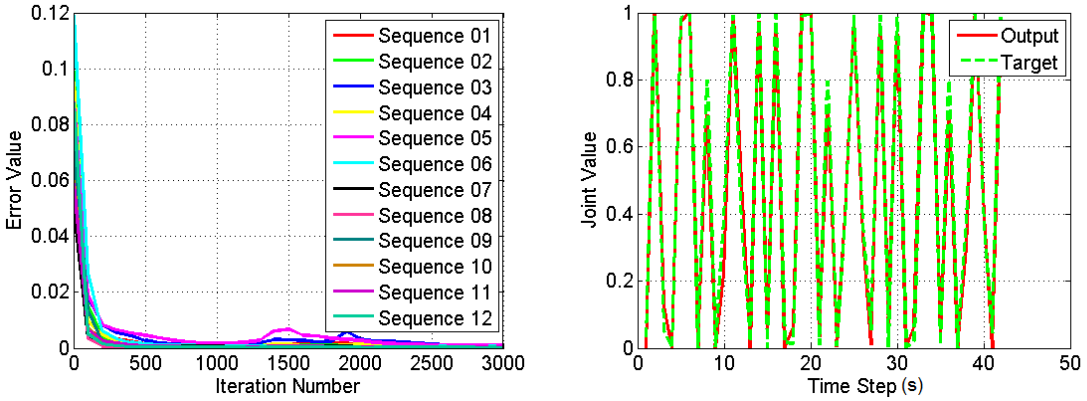
- **Abstract Action Words Learning**. The whole data set was used for the training of the model, and the performance of the network were assessed in response to the perturbation of the inputs of the model.

The performance of the generalization test at the end of the second stage of the training aimed to verify the capacity of the model to generalize superordinate words from basic words directly grounded in perception and sensorimotor experience.

7.6 Training Phase I

The first training stage of the model aimed to endow the robot with basic perceptual and sensorimotor skills necessary for scaffolding higher-order capabilities. During this phase of the training the robot acquired the knowledge related to visual properties of objects and learned to perform some motor behaviours. In particular, the model was trained in order to recognize twelve tools and perform twelve actions obtained by the integration of low level motor primitives. The model was trained with the perceptual features of all the twelve object categories and with the sensorimotor sequences of all the twelve behavioural categories in a supervised manner. In this stage, the network hidden layer consisted of 13 neurons and the training was performed for 25 random seeds by activating the visual and proprioceptive inputs only, while the linguistic inputs were silent. The network received in input twelve sequences of six elements each. The training was successfully completed and objects and actions were correctly categorized. The Mean Square Error (MSE) calculated

at the end of this training stage is shown in figure 7.8(a), while figure 7.8(b) presents the output and target joint values for one of the actions taught to the iCub.



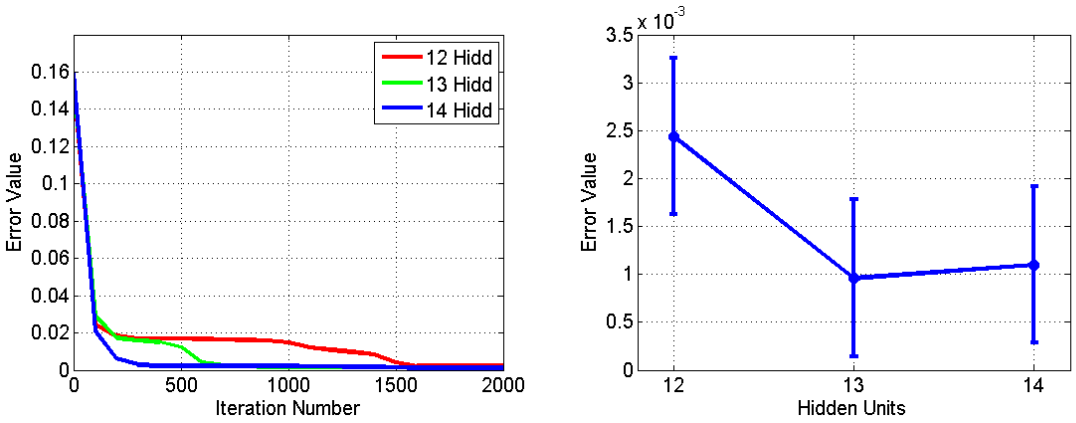
(a) (b)

Figure 7.8: Training stage I. Mean Square Error (MSE) (a). Output and target joint values for one of the actions taught to the iCub (b)

The successful performance of the Training Phase I permitted to acquire the basic perceptual and sensorimotor knowledge to be used in the next stages of training for the grounding of basic and abstract action words.

7.7 Training Phase II

The second stage of the training enabled the model to acquire linguistic capabilities through the naming of objects and actions. During this stage of the training the network created the connections between the sensorimotor/proprioceptive inputs and the linguistic labels. Therefore, the four inputs of the model were all activated. The network received in input twelve sequences of six elements each. The training of the network has been performed for 25 random seeds. This stage of the training has been performed on a network consisting of 13 neurons in the hidden layer; nevertheless, the performance of the model as a function of the neurons in the hidden layer were evaluated. In figure 7.9 the training error as a function of the neurons in the hidden layer is shown.



(a) (b)

Figure 7.9: Training stage II. (a) Mean Square Error (MSE) as a function of the hidden layer size. (b) RMSE at iteration 2000

In particular, in figure 7.9(a) the MSE is compared for the hidden layer consisting of 12, 13 and 14 hidden neurons. In figure 7.9(b) the MSE values recorded at the iteration 2000 are shown. The network with 13 neurons in the hidden layer, having the lowest MSE value, was selected to perform further analysis and tests. As it is possible to observe from figure 7.10 the mean square error value of the network with 13 neurons in the hidden layer for all the twelve input sequences, after 2000 iterations only, is smaller than 0.001 (stopping criterion of the learning algorithm).

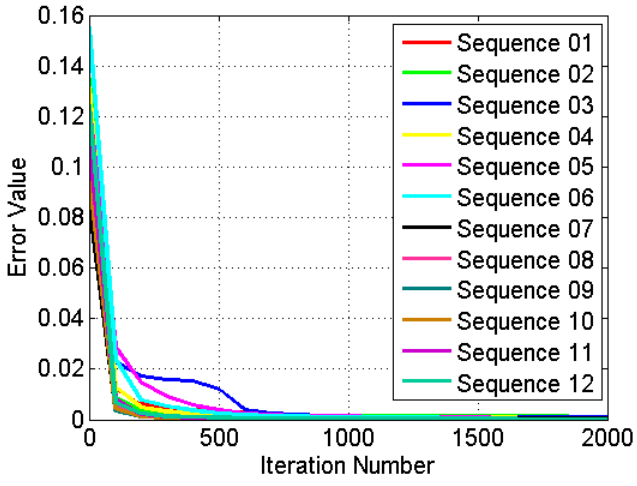


Figure 7.10: Training stage II. Mean Square Error (MSE) for the model with 13 hidden neurons

In figure 7.11 activation values of hidden units show that during the time steps $[0, 36]$, the hidden units were alternatively activated, while during the time steps

[37, 72] activation values followed a more stable and continuous pattern.

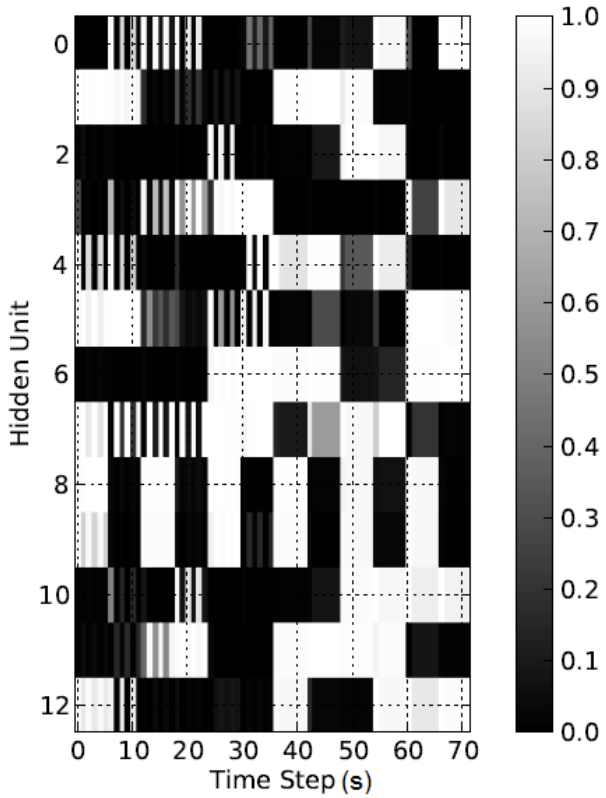
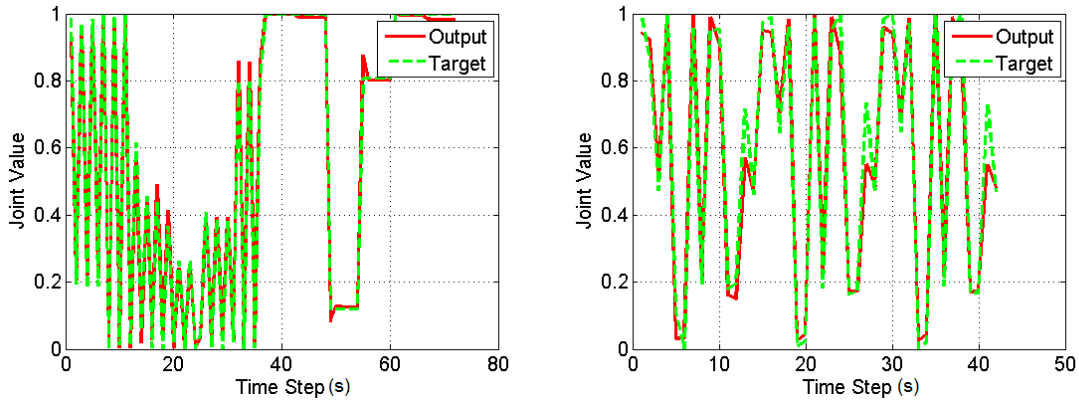


Figure 7.11: Training stage II. Raster plot of hidden units activation values

The different activation patterns of hidden units recorded during the time steps [0, 36] and [37, 72] are due to the differences in the structure of the training sequences. Indeed, half of the sequences are related to the learning of iterative actions, while the remaining half of the training set is related to the learning of non-iterative actions.

The selected network successfully learned the input/output mappings for joint values (Fig.7.12). In figure 7.12(a) the output and target values for one of the seven joints of the iCub arm controlled by the network is shown. As it is possible to observe from figure 7.12(a), the network after the training can output the appropriate joint values for the iCub arm. During the time steps [0, 36] the plot shows the trend of the joint values during the execution of iterative actions, while the time steps [37, 72] are related to the joint values associated to the non-iterative actions. In figure 7.12(b) output and target joint values for one of the twelve actions taught to the iCub are shown. During the time steps [0, 13], [14, 27] and [28, 42] the trend of the plot is

repeated. These repetitions correspond to the three iterations of the same action.



(a) (b)

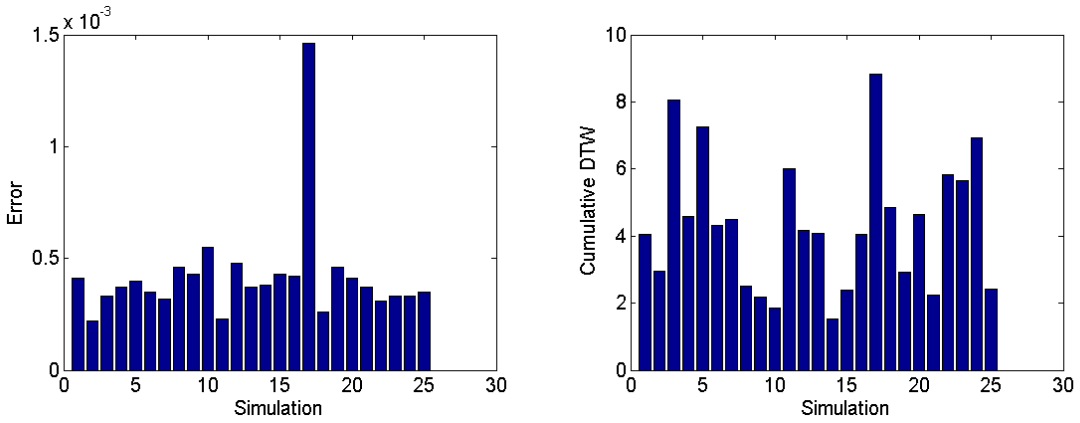
Figure 7.12: Training Stage II. Output and target values for one of the seven joints of the iCub arm controlled by the network (a). Output and target joint values for one of the actions taught to the iCub (b)

In order to have a quantitative measure of the similarity between the output and target joint values over time, the Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978] on joint sequences was computed. The DTW, differently from the Euclidean distance (or warping) that cannot compensate for small distortions in time axis, permits to calculate the similarity between behaviour (classification of behaviour) over time. Indeed, the DTW is a time series alignment algorithm developed originally for speech recognition [Sakoe and Chiba, 1978]. The aim of DTW is to align two sequences by warping the time axis iteratively until an optimal match between the two sequences is found. Herein a formal definition of the DTW is provided. Let $X(x_1, x_2, \ldots, x_n)$ and $Y(y_1, y_2, \ldots, y_m)$ be two series of length n and m, respectively [Li et al., 2010]. The point-to-point correspondence relationship between X and Y can be defined in a matrix M of dimension $n \times m$; each element M_{ij} indicates the distance $d(x_i, y_j)$ between x_i and y_j. Then the point-to-point alignment and matching relationship between X and Y can be represented by a time warping path $W(w_1, w_2, \ldots, w_K)$, $\max(m, n) \leq K < m + n - 1$, where the element $w_k = (i, j)$ indicates the alignment and matching relationship between x_i and y_j. Hence, the dynamic time warping distance between the two series X and Y is defined as:

$$DTW(X,Y) = \min_{W} \left\{ \sum_{k=1}^{K} d_k, W = \langle w_1, w_2, \ldots, w_K \rangle \right\}$$

where $d_k = d(x_i, y_j)$ indicates the distance represented as $w_k = (i, j)$ on the path W.

The result of DTW confirmed that the output joint values over time are very similar to the target values (DTW = 1.5286 un-normalized distance between sequences). Learning error and DTW for the 25 simulations, performed for different random seeds and initial synaptic weights, are shown in figure 7.13 from which it is possible to observe that the best results in terms of MSE and DTW are given by the network trained during the simulation 14 which was used as a controller of the iCub robot during the performance of the tests presented in Section 7.7.1.



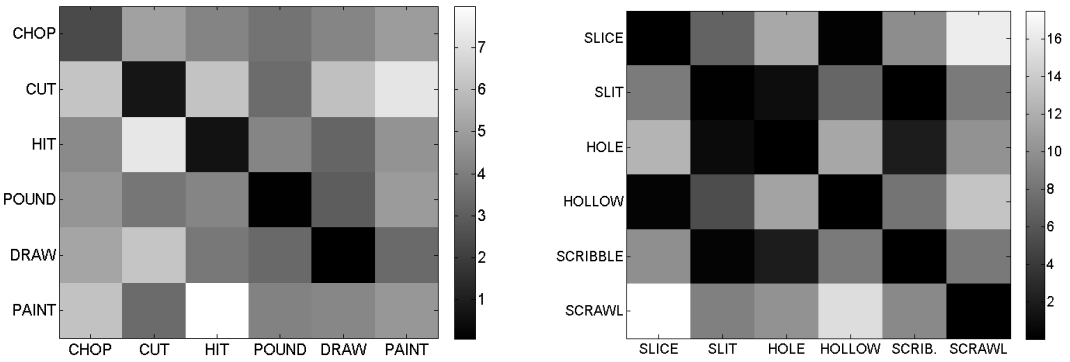
(a) (b)

Figure 7.13: Training stage II. Comparison of MSE (a) and cumulative DTW (b) computed during the 25 simulations performed for different random seeds and initial synaptic weights

7.7.1 Robot Performance

After the off-line training, the model with the best performance has been used to control the iCub robot. In order to enable the model to better adjust its internal dynamics for reaching a specific target, each action was performed for twelve time steps (instead of six as for the training of the model). The joint values recorded after the performance of each action, were compared to the corresponding target values

by performing the DTW (Fig.7.14). To better understand the capacity of the model to categorize the proprioceptive input, the DTW of the actual joint values related to each action reproduced by the model has been computed with respect to the target joint values related to all the possible actions taught to the robot. Results of the DTW are presented in the gray-maps in figure 7.14. Each row of the gray-map represents the actual joint values produced by the model, while columns represent the target joint values related to the different actions. By displaying the results of the DTW in the proposed gray-map layout, it is easier to visualise the capacity of the model to categorize the proprioceptive inputs and analyse the performance of the robot in executing the desired behaviour.



(a) (b)

Figure 7.14: Training Stage II. Gray-map of the results of the Dynamic Time Warping performed on joint values: iterative actions (a), non-iterative actions (b)

From figure 7.14(a) it is possible to observe that five out of the six iterative actions (i.e. "CHOP", "CUT", "HIT", "POUND", "DRAW") have the lowest DTW values (corresponding to cell of the gray-map of darker gray) when compared to their corresponding target values, while in case of the "PAINT" action, the lowest DTW value is obtained when compared to the target joint values related to "CUT". In other words, this means that the robot when asked to "PAINT" it performs an action that, in terms of joint values, is closer to "CUT" than "PAINT". From figure 7.14(b) it is possible to notice that all the six non-iterative actions were very well performed and classified. Given the similarity among the six non-iterative actions,

the DTW has low values in correspondence of more than one target. Nevertheless the lowest DTW for the non-iterative actions is registered in correspondence of the comparison with the appropriate targets.

A visual representation of the similarity of joint sequences in output from the model is presented in the star plots in figure 7.16. Each action consisted of twelve observations of seven variables (12-by-7 matrix). In each star plot observations are represented as stars whose *i-th* spoke is proportional in length to the *i-th* coordinate of the particular observation. Before creating the star plot, the matrix associated to joint values was standardized (centred and scaled). For example, the start plot in figure 7.15 represents joint values recorded during the "CHOP" action. Given that the execution of each action requires the update of joint values from the home to the target position, the matrix representing joint values (12×7) was rearranged (6×14) in order to show the variation of joint values during two consecutive time steps (necessary to update joint values from the home to the target position). From the figure 7.15 it is possible to observe that from the centre of the star depart 14 spokes, each of which corresponds to one of the 14 observed variables.

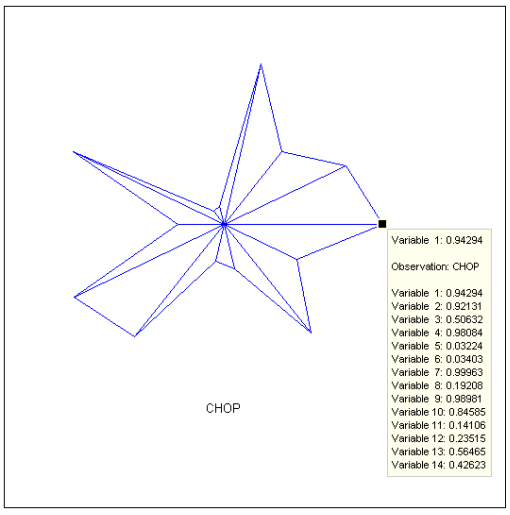
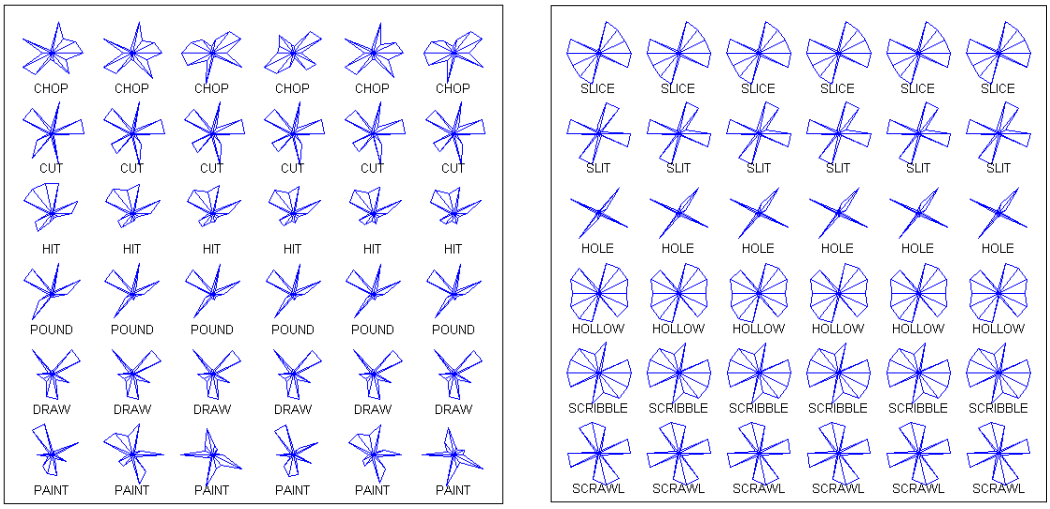


Figure 7.15: Training stage II. Star plot for joint values recorded during the CHOP action

In figure 7.16(a) each row contains six star plots, each of which corresponds to the joint values recorded during two consecutive time steps (necessary to perform

one action and to update the joint values of the robot from the home to the target position). For example, the star plots from 1 to 6 correspond to joint values related to the "CHOP" action. The star plots of the iterative actions in figure 7.16(a) provide a qualitative measure that confirms results obtained performing the DTW.

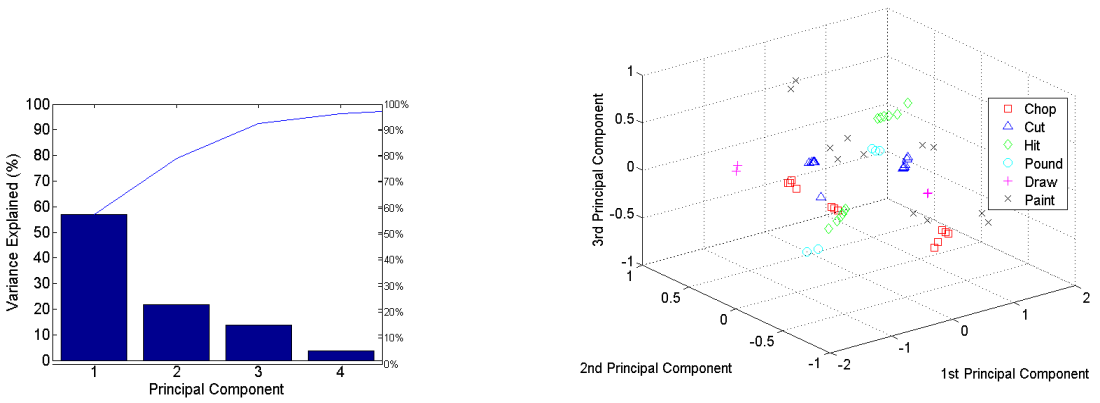


(a) (b)

Figure 7.16: Training stage II. Star plots for joint values: iterative actions (a), non-iterative actions (b)

Indeed, as shown in figure 7.14(a), for the iterative actions the highest values of the DTW are related to "PAINT" and "CHOP" (corresponding to cells of the gray-map of lighter gray), that in case of the star plots (Fig.7.16(a)) correspond to stars with different shapes along the six repetitions of the same action (stars in the first and sixth rows). The start plots for the non-iterative actions are shown in figure 7.16(b). The high similarity among the star plots during the six repetitions of each action shown in figure 7.16(b), confirms that all the six non-iterative actions are very well categorized. Considering that each action is represented by twelve observations of seven variables (12-by-7 matrix), to visualize these multivariate data and analyse the relationship between variables, it is necessary to simplify the problem by replacing correlated variables with a single new variable. Principal Component Analysis (PCA) is a quantitatively rigorous method for achieving this simplification. The method generates a new set of variables, called principal components each of

which is a linear combination of the original variables. All the principal components are orthogonal to each other, so that there is no redundant information. On the matrix associated to the joint values recorded during the execution of the twelve actions, the PCA was performed. The PCA was first executed on the matrix of the joint values related to the iterative actions (matrix 72-by-7). From figure 7.17(a) it is possible to observe that the percent variance explained by the first three principal components is roughly equal to 93%, while the plot in figure 7.17(b) shows the data projected into the space defined by the first three principal components.

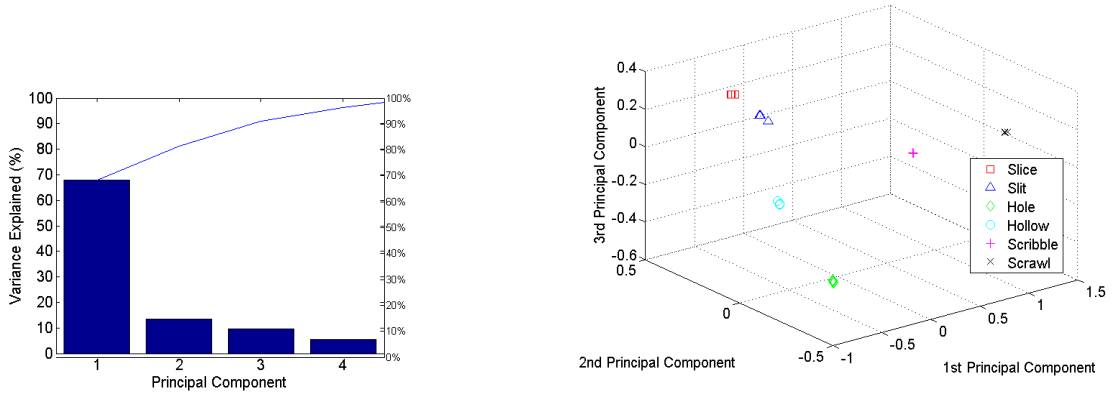


(a) (b)

Figure 7.17: Training Stage II. Principal Components Analysis of joint values recorded during the iterative actions: percent variability explained by each principal component (a). Data projected onto the first three principal components (b)

The points in the 3-D plot in figure 7.17(b) represent the observations of the seven joints values, with coordinates indicating the score of each observation for the three principal components. Markers of different colours and shapes correspond to the observations related to different actions. From figure 7.17(b) it is possible to observe that the joint values related to the six iterative actions form twelve clusters corresponding to the joint values of each action recorded during two consecutive time steps. The observations displayed with the black cross markers correspond to the joint values recorded during the execution of the "PAINT" action which, as it has been shown in the gray-map of the DTW in figure 7.14(a) and in the start plots in figure 7.16(a), is the action that has not been correctly categorized (i.e. highest

DTW value) by the model. The PCA was then performed on the matrix of the joint values related to the non-iterative actions (matrix 72-by-7). From figure 7.18(a) it is possible to observe that the percent variance explained by the first three principal components is around 91%. The plot in figure 7.18(b) shows the data projected onto the first three principal components.

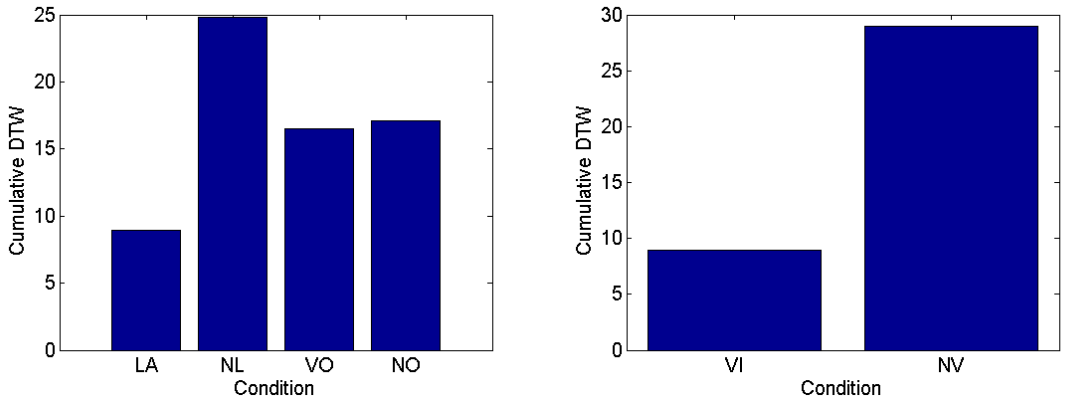


(a) (b)

Figure 7.18: Training Stage II. Principal Components Analysis of joint values recorded during the non-iterative actions: percent variability explained by each principal component (a). Data projected onto the first three principal components (b)

The points in the 3-D plot in figure 7.18(b) represent the observations of the seven joint values, with coordinates indicating the score of each observation for the three principal components. Markers of different colours and shapes correspond to the observations related to different actions. From figure 7.18(b) it is possible to observe that the joint values related to the six non-iterative actions form six clusters, each of which corresponds to the joint values recorded during the execution of each non-iterative action. Hence, the PCA confirms that the performance of the robot in terms of action execution for the non-iterative actions is better than the performance during the iterative ones. Indeed, the joint values related to the six non-iterative actions are very well clustered. However, the mapping of the joint values associated to the non-iterative actions was easier than learning the mapping of the joint values associated to the iterative ones, which required to repetitively alternate the values of the robot's encoders from the home to the target values.

The robot's performance in terms of action's execution has been evaluated in different conditions. In figure 7.19(a) the cumulative DTW (for all the twelve actions) of the joint values is compared in different experimental conditions: with language (LA), without language (NL), with verb only (VO) and with noun only (NO) in input to the model. From figure 7.19(a) it is possible to observe that the robot performance deteriorates when the linguistic input is not provided (NL condition).



(a) (b)

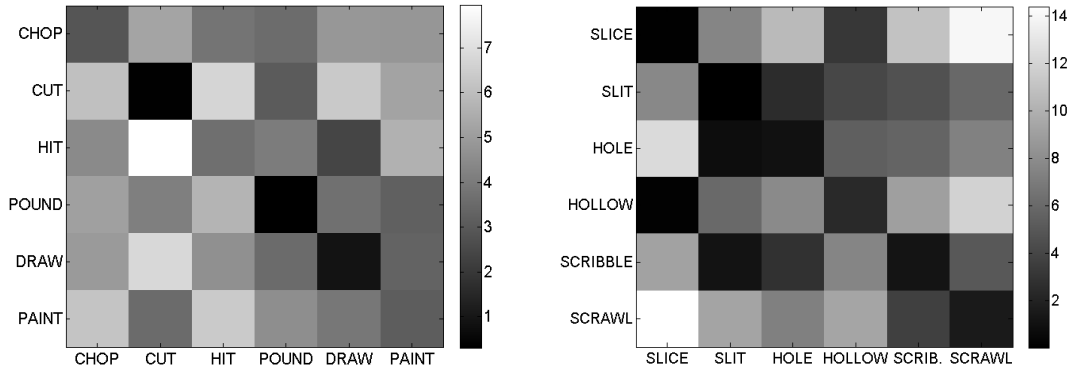
Figure 7.19: Cumulative DTW of the joint values compared in different experimental conditions: with language (LA), without language (NL), with verb only (VO) and with noun only (NO) (a). Cumulative DTW of the joint values compared in presence (VI) and absence of the visual input (NV) (b)

Furthermore, the robot's performance has been evaluated in absence of the visual input. The cumulative DTW (for all the twelve actions) of the joint values is compared in presence (VI) and absence of the visual input (NV) (Fig.7.19(b)). From figure 7.19(b) it is possible to observe that the robot performance, when the visual input is not provided, is even worst than in absence of linguistic input (Fig.7.19(a)). Furthermore, when the perceptual input is deactivated, the hidden units of the model follow a less structured and more chaotic pattern.

7.7.2 Generalization

After the training phase II, the capability of the model to generalize the meanings of new words has been tested. In particular, the performance of the model in response to new linguistic inputs, for which the network has never been trained on before,

has been analysed. The new linguistic inputs, at these stage, are the abstract action words that the network will learn during the third stage of training. During this test, the model reads the visual and proprioceptive inputs related to objects and actions from the robot's sensors, while for the linguistic inputs, labels associated to new words are read from text files. The DTW for joint values in output from the model has been computed. Results are presented in the gray-maps in figure 7.20(a) and 7.20(b).

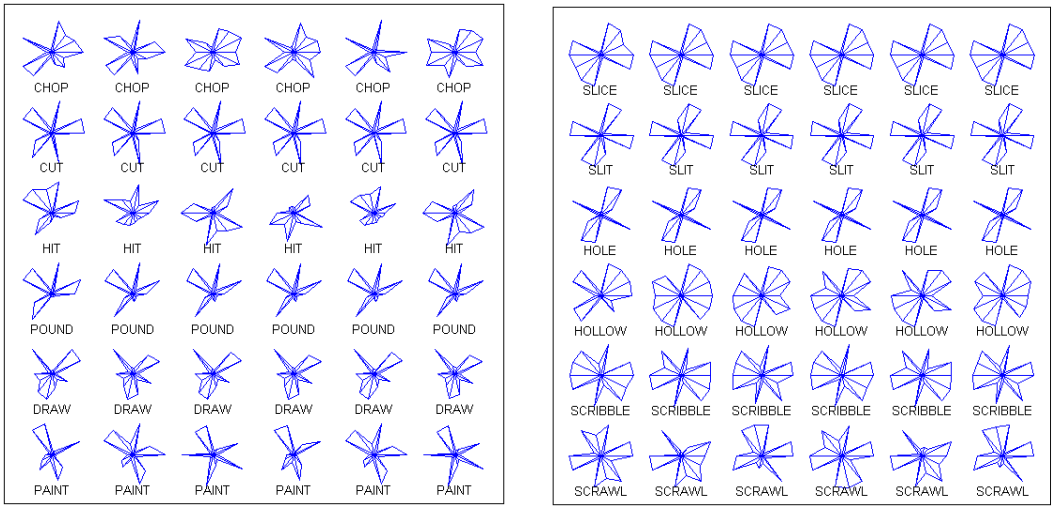


(a) (b)

Figure 7.20: Gray-map for the results of the DTW performed on joint values recorded during the generalization test: iterative actions (a), non-iterative actions (b)

During the generalization test, the robot is still capable of performing the appropriate actions, although the DTW has higher values in comparison with the DTW computed in the previous training stage (Fig.7.14). For the iterative actions, from the gray-map (Fig.7.20(a)) it is possible to observe that the "HIT" action has the lowest DTW when compared to the target joint values related to "DRAW". In case of non-iterative actions, from the gray-map (Fig.7.20(b)) it is possible to observe that the "HOLLOW" action has the lowest DTW when compared to target joint values related to "SLICE". The start plots in figure 7.21(a) and 7.21(b)) provide a confirmation for the results obtained calculating the DTW. Indeed, in figure 7.21(a) the stars in the third row (stars from 13 to 18) have different shapes during the six repetition of the "HIT" action, while in figure 7.21(b) the stars in the fourth

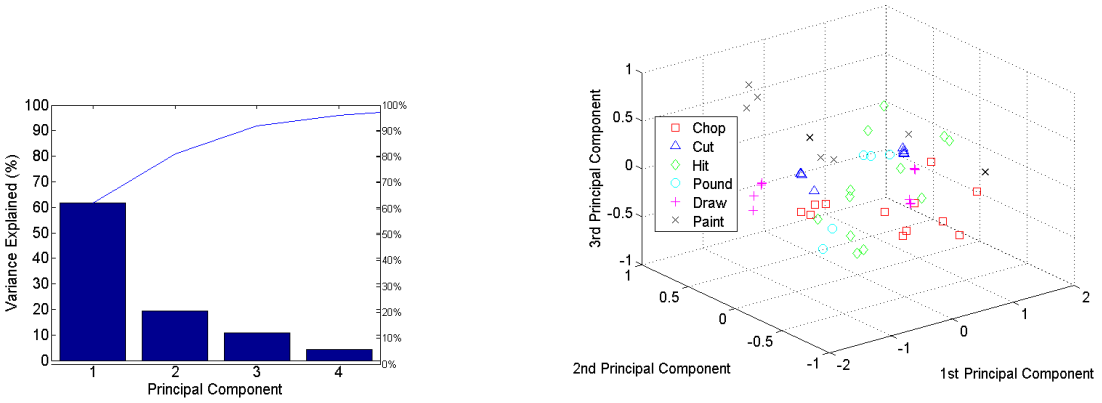
row (stars from 19 to 24) have different shapes during the six repetition of the "HOLLOW" action.



(a) (b)

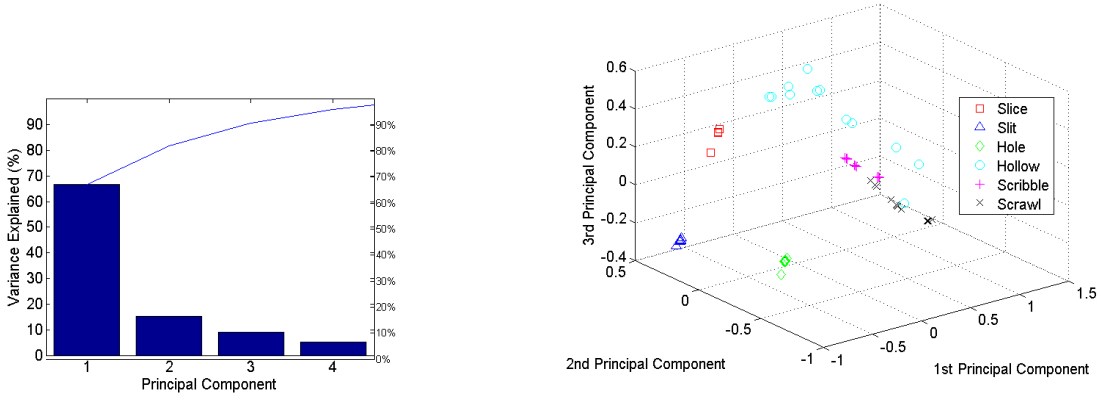
Figure 7.21: Star plots for the joint values recorded during the generalization test: iterative actions (a), non-iterative actions (b)

For the joint values recorded at the end of the generalization test, a PCA has been applied. From figure 7.22(a) it is possible to observe that for iterative actions the percentage variance explained by the first three principal components is equal to 91.88%, while for the non-iterative actions the percent variance explained is 90.73% (Fig.7.23(a)). From figure 7.22(b) and 7.23(b) it is possible to notice that during the generalization test the joint values associated to iterative and non-iterative actions form clusters that are less structured with respect to joint values recorded during the previous experimental scenario (Fig.7.17, Fig.7.18).



(a) (b)

Figure 7.22: Generalization. Principal Components Analysis of joint values recorded during the iterative actions: percent variability explained by each principal component (a). Data projected onto the first three principal components (b)



(a) (b)

Figure 7.23: Generalization. Principal Components Analysis of joint values recorded during the non-iterative actions: percent variability explained by each principal component (a). Data projected onto the first three principal components (b)

7.7.3 Incompatible Condition Test

Before proceeding with the third stage of the training, the "Incompatible Condition" test was performed. The test consisted in analysing the response of the model in case of inconsistency between the linguistic and visual inputs. During this test, objects and actions that the robot has previously learned to name, were referred using incompatible linguistic labels. In particular, two different incompatible condition tests were performed:

- **Incompatible Noun Condition**: to analyse the response of the model when the name of the object is incompatible with the object seen by the robot
- **Incompatible Verb Condition**: to analyse the response of the model when the name of the action is incompatible with the behaviour that the robot has previously performed with the presented object

At the end of the simulations related to these two tests, activation values of hidden units were analysed. In particular, the temporal hierarchical cluster analysis on hidden units has been performed in order to compare the hidden activation values recorded during the compatible and incompatible conditions. The cluster analysis has been performed on activation values of hidden units recorded at each time step (matrix of 12 observations by 13 variables for each action). For the formation of clusters, as measure of dissimilarity between pairs of observations, the Euclidean distance ($\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$) has been used. The results of hierarchical clustering are presented in the dendrograms in figure 7.24 and 7.25, from which it is possible to observe that over time the hidden units during the incompatible condition follow an activation pattern that is similar to the activation values recorded during the compatible condition.

- **Results of the Incompatible *Noun* Condition Test**.

In figure 7.24 the results of the hierarchical clustering of activation values of hidden units at the time steps $T = 0$, $T = 5$ and $T = 11$ are presented. The dendrograms in figure 7.24 compare the hidden activation values recorded during the compatible condition "**CHOP** [with] **KNIFE**" to the hidden activation values recorded during the incompatible condition "**CHOP** [with] **HAMMER**". In this particular case, the incompatibility is related to the KNIFE/HAMMER nouns. Despite that the robot sees a KNIFE, the word HAMMER is used to refer to the object. The dendrograms in figure 7.24 show that the observations are organized in three main clusters that pair the inputs related to the six iterative actions. The

presence of clusters in the hidden units suggests the formation of concepts from the multi-modal data received as input to the model.

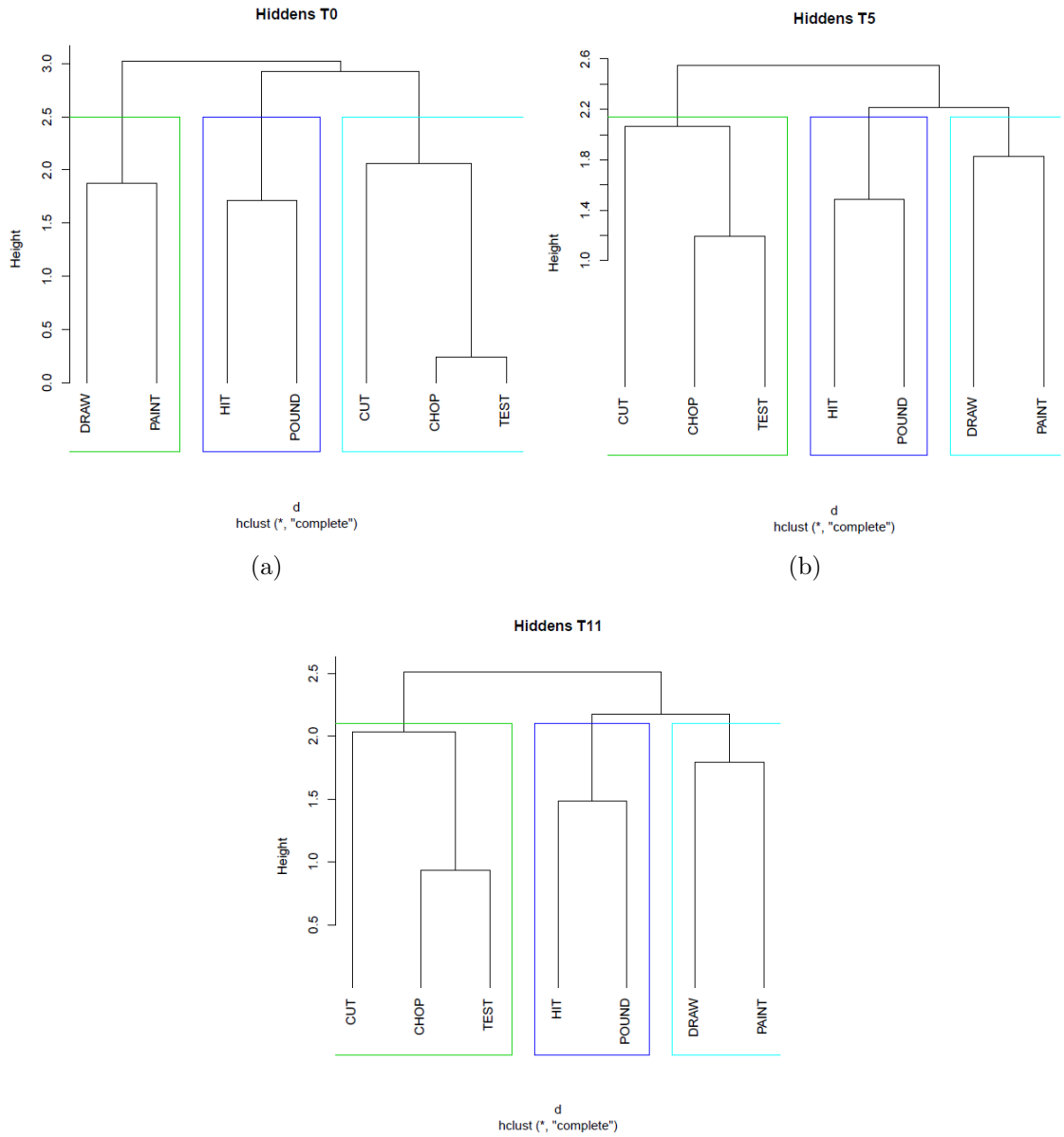
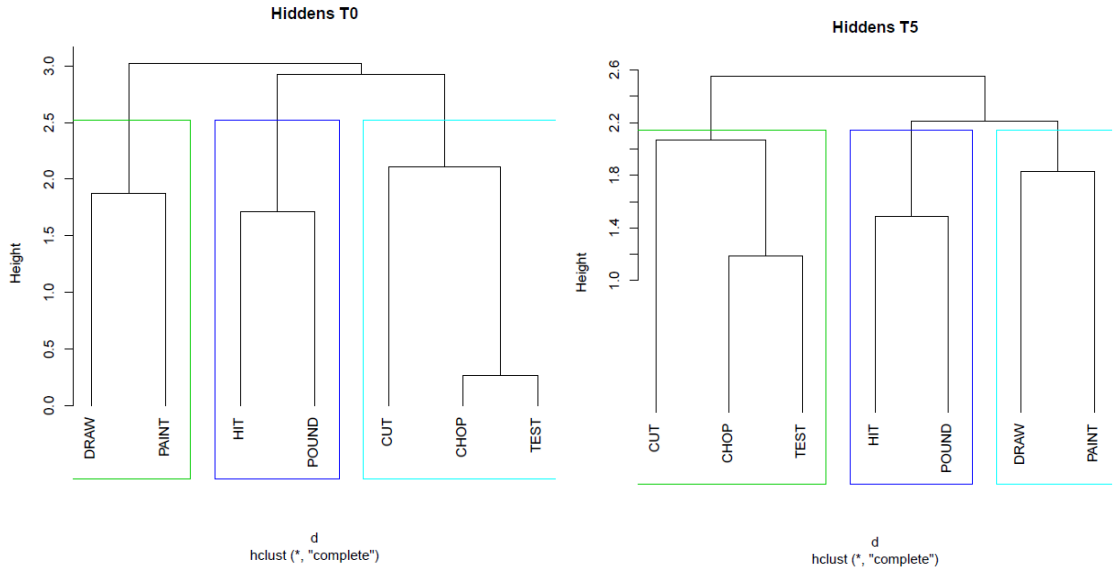


Figure 7.24: Incompatible *Noun* Condition (e.g. "**CHOP** [with] **KNIFE**" *became* "**CHOP** [with] **HAMMER**"). Results of the hierarchical clustering of hidden units activation values at the time steps $T = 0$ (a), $T = 5$ (b) and $T = 11$ (c)

The hidden activation values related to the incompatible condition "**CHOP** [with] **HAMMER**" (that in the dendrograms are labelled "TEST") are clustered together with "CHOP". This means that the activation values of hidden units during this incompatible condition test are similar to the activation values of hidden units

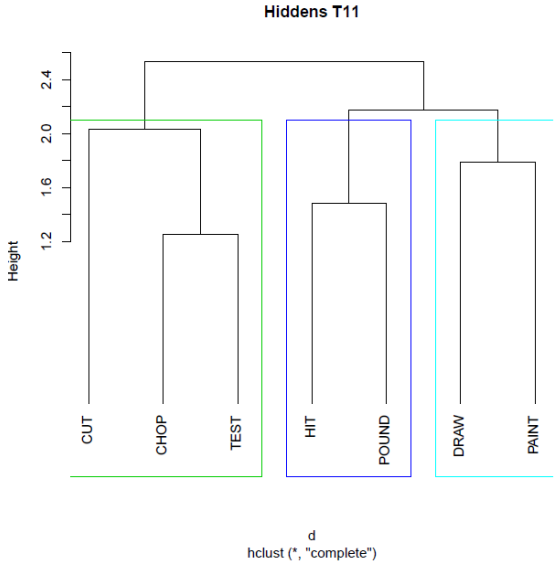
recorded during the compatible condition.

- **Results of the Incompatible *Verb* Condition Test.**



(a)

(b)

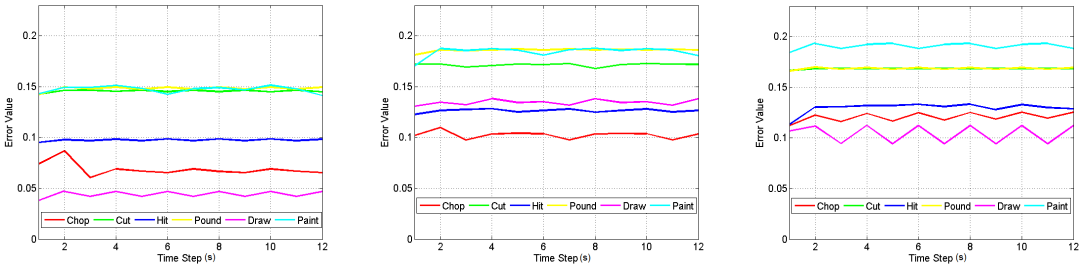


(c)

Figure 7.25: Incompatible *Verb* Condition (e.g. "**CHOP** [with] **KNIFE**" *became* "**DRAW** [with] **KNIFE**"). Results of the hierarchical clustering of hidden units activation values at the time steps $T = 0$ (a), $T = 5$ (b) and $T = 11$ (c)

In figure 7.25 the results of the hierarchical clustering of hidden units activation values at the time steps $T = 0$, $T = 5$ and $T = 11$ are presented. The dendrograms in figure 7.25 compare the hidden activation values recorded during the compatible condition "**CHOP** [with] **KNIFE**" to the hidden activation values recorded

during the incompatible condition "**DRAW** [with] **KNIFE**". In this case, the incompatibility is related to the CHOP/DRAW verbs. Despite that the robot sees a KNIFE, the verb DRAW is used to refer to the action to be performed with the presented object. The dendrograms in figure 7.25 show that the observations are organized in three main clusters that pair the inputs related to the six iterative actions. The hidden activation values related to the incompatible condition "**DRAW** [with] **KNIFE**" (that in the dendrograms are labelled "TEST") are clustered together with "CHOP". This means that the activation values of hidden units during this incompatible condition test are similar to those recorded during the compatible condition. Furthermore, the error recorded at the end of each action execution in the compatible condition has been compared to the error recorded at the end of each action execution during the incompatible condition (Fig.7.26).



(a) (b) (c)

Figure 7.26: MSE recorded during the execution of the iterative actions: compatible condition (a), incompatible NOUN condition (b) and incompatible VERB conditions (c)

In figure 7.26 the MSE recorded during the execution of the six iterative actions in the compatible condition, is compared to the MSE recorded during the incompatible noun and verb conditions. The higher error rates in incompatible trials than in the compatible ones, suggest that referring to objects with the appropriate words facilitate the perceptual and sensorimotor categorization of the input signals.

Additionally, on the matrices associated to the activation values of hidden units recorded during the two incompatible conditions tests, the PCA has been performed. The trajectories of the activation values of hidden units in time, recorded during the

incompatible condition tests, in the space of the first three principal components, have been compared to the trajectories of activation values recorded during the compatible condition (Fig.7.27, Fig.7.28). From figure 7.27 it is possible to observe that the trajectories of activation values of hidden units recorded during the incompatible noun condition "**CHOP** [with] **HAMMER**" follow trajectories that are similar to the trajectories of hidden units recorded during the compatible condition "**CHOP** [with] **KNIFE**"; and this is the case for all the six iterative actions.

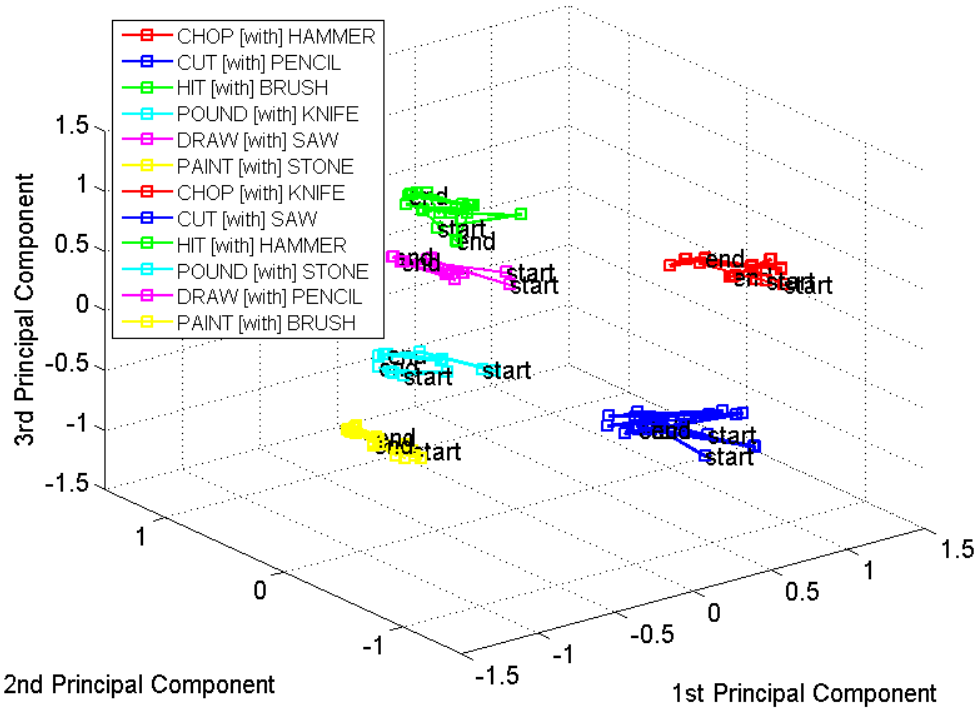


Figure 7.27: Trajectories of the activation values of hidden units recorded during the incompatible NOUN condition test compared to the trajectories of activation values recorded during the compatible condition

From figure 7.28 it is possible to observe that the trajectories of activation values of hidden units recorded during the incompatible verb condition "**DRAW** [with] **KNIFE**" follow trajectories similar to the ones recorded during the compatible condition"**CHOP** [with] **KNIFE**".

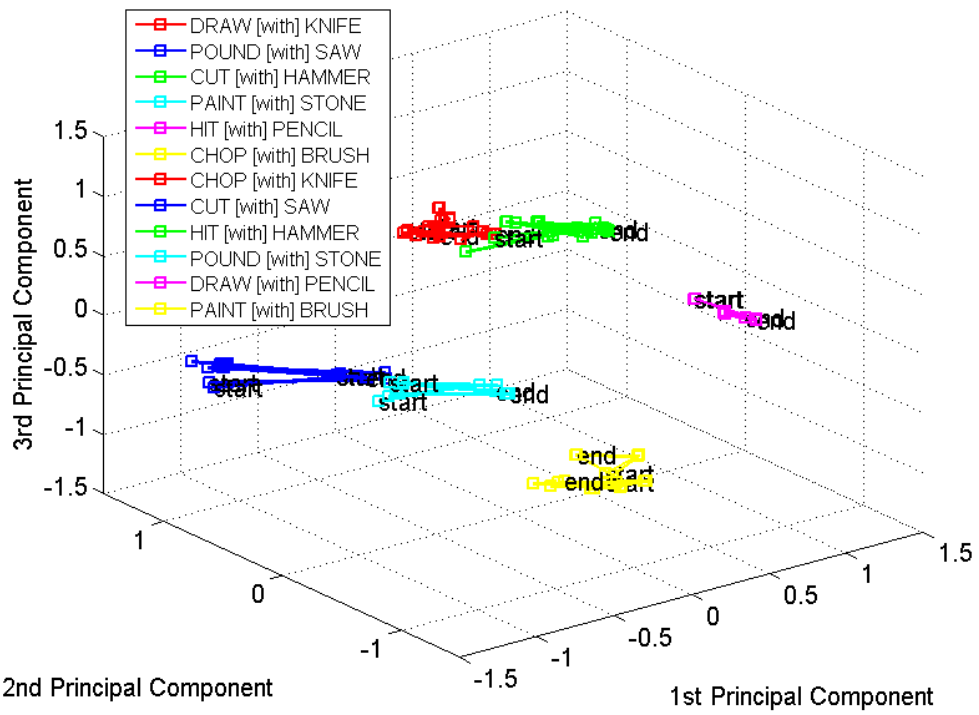


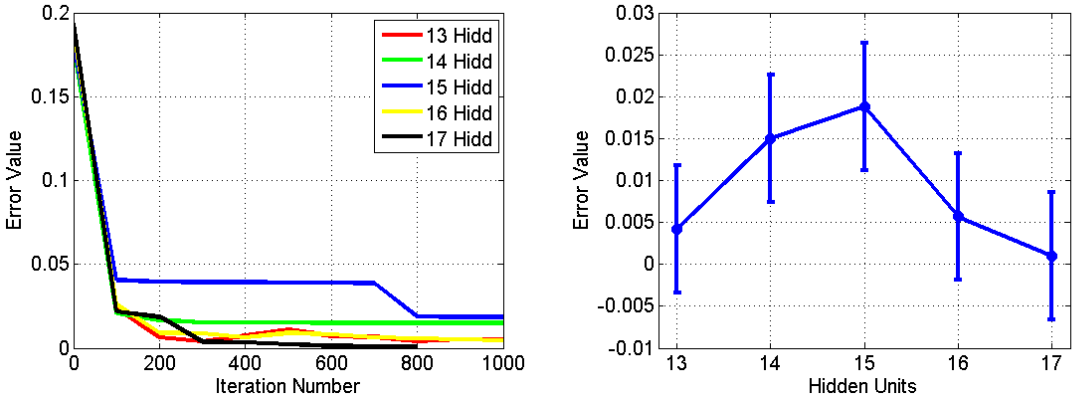
Figure 7.28: Trajectories of the activation values of hidden units recorded during the incompatible VERB condition test compared to the trajectories of activation values recorded during the compatible condition

Results obtained in the incompatible condition tests showed that in case of inconsistency between the perceptual and linguistic input, the robot executed the actions elicited by the seen object. Recent evidence in neuroscience and behavioural sciences has shown that visually perceived objects activate motor information [Jeannerod, 1994, Arbib, 1997]. That is, seeing objects elicits the actions that tend to be performed on/with objects [Jeannerod, 1994, Arbib, 1997]. Additionally, studies conducted on monkeys [Gallese et al., 1996], have suggested that the brain stores a *vocabulary* of actions that can be applied to objects and that the fixation of a given object activates potential motor acts [Cangelosi et al., 2010].

7.8 Training Phase III

The last stage of the training enabled the model to learn abstract action words and acquire higher-order categories. During this part of the training new concepts are

formed by integrating the lexical terms acquired during the previous stage of the training. Since such lexical terms are directly connected to perceptual and sensorimotor experience, they recall the grounded perceptual and sensorimotor knowledge (multi-modal symbols). Through this stage of the training, the model can learn novel meanings by integrating the perceptual and sensorimotor knowledge previously grounded. For the third stage of the training, the number of neurons in the hidden unit was increased from 13 to 17. The training of the network was done using 50 random seeds. The network received in input 24 sequences of six elements each. In figure 7.29 the training error as a function of the neurons in the hidden layer is shown.



(a) (b)

Figure 7.29: Training stage III: Mean Square Error (MSE) as a function of the hidden layer size

In particular, in figure 7.29(a) the MSE is compared for the hidden layer consisting of 13, 14, 15, 16 and 17 hidden neurons. In figure 7.29(b) the MSE values recorded at iteration 1000 are shown. After analysing the performed simulations, the network with 17 neurons in the hidden layer, that exhibited the best performance in terms of training error, was selected to perform additional tests and analysis. From figure 7.30 it is possible to observe that the mean square error value for all the 24 sequences in input, after 800 iterations only, is smaller than 0.001.

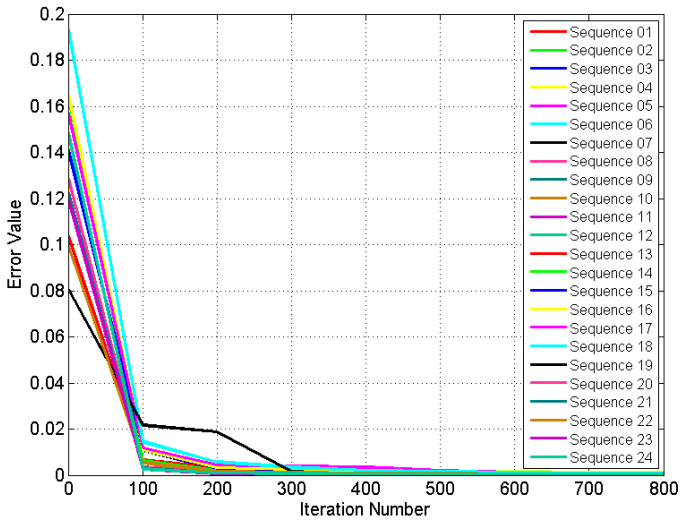
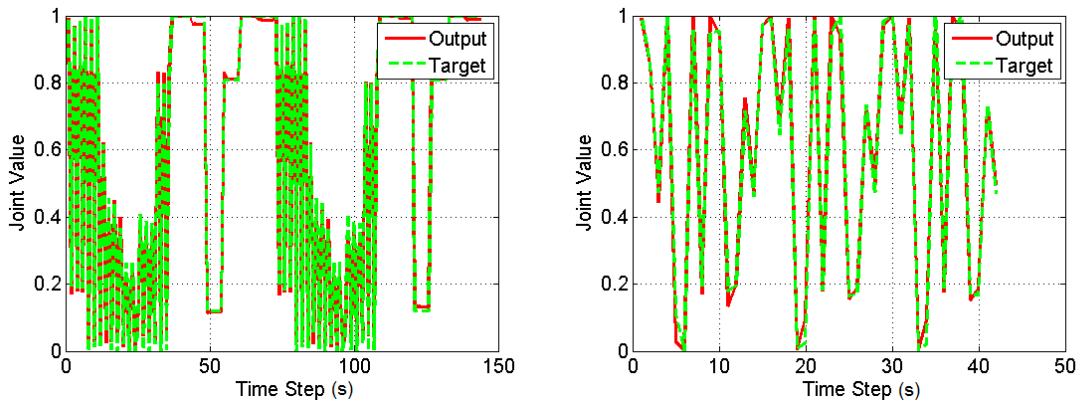


Figure 7.30: Training stage III. Mean Square Error (MSE)

The selected network successfully learned the input/output mapping for the joint values (Fig.7.31). In figure 7.31(a) the output and target values for one of the seven joints of the iCub arm controlled by the network is shown. As observed in figure 7.31(a), the network is able to output the appropriate joint values for the iCub arm.

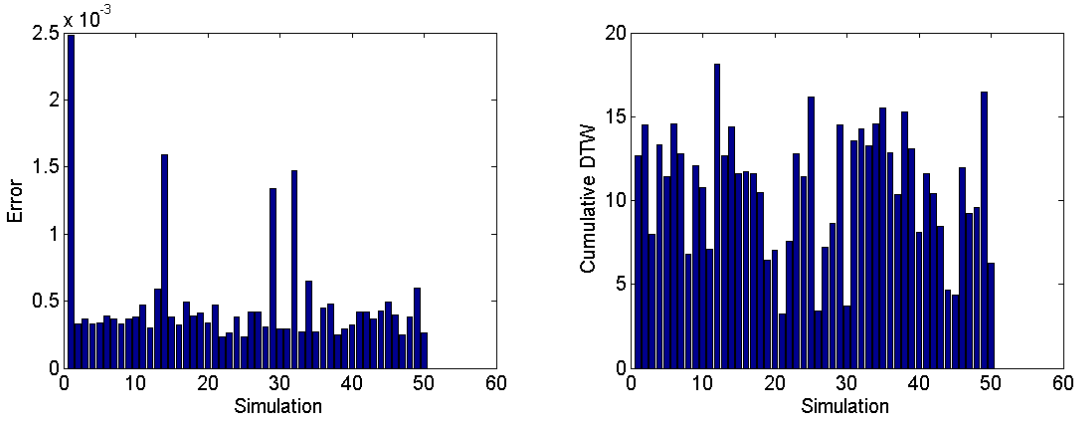


(a) (b)

Figure 7.31: Training stage III. Output and target values for one of the seven joints of the iCub arm controlled by the network (a). Output and target joint values for one of the actions taught to the iCub (b)

The similarity between the output and target joint values over time has been calculated by performing the DTW on joint sequences. The result of DTW confirmed that the output joint values over time are very similar to the target values (DTW

= 3.2341 un-normalized distance between sequences). The DTW calculated at the end of the third training stage is higher than the DTW calculated at the end of the second one; this is due to the fact that the training set during the third stage of the training is larger. Learning errors and DTW values related to the 50 simulations, performed for different random seeds and initial synaptic weights, are shown in figure 7.32. From this figure it is possible to observe that the best results in terms of MSE and DTW is given by the network trained during the simulation 21.

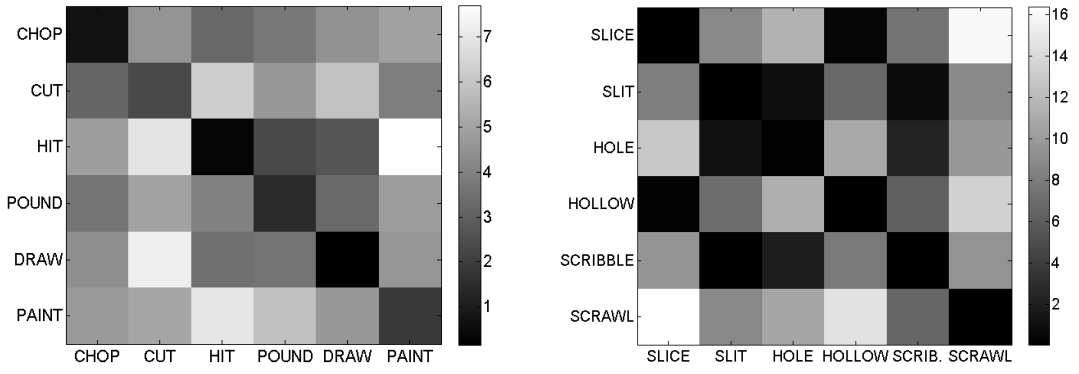


(a) (b)

Figure 7.32: Training stage III. MSE (a) and cumulative DTW (b) computed during the 50 simulations performed

7.8.1 Robot Performance

After all the three stages of the training were successfully accomplished, the network trained during simulation 21, which exhibited the best performance in terms of training error and DTW, was selected to control the real iCub robot. The joint values recorded after the performance of each actions, were compared to the corresponding target values by performing the DTW (Fig.7.33). For both iterative and non-iterative actions it is possible to observe that the lowest DTW is obtained when the actual output joint values are compared to their corresponding targets (Fig. 7.33(a), Fig.7.33(b)).



(a) (b)

Figure 7.33: Training stage III. Gray-map for the results of the DTW performed on joint values: iterative actions (a), non-iterative actions (b)

The robot performance in terms of action execution, at the end of the second and third stage of the training, are compared and displayed in Table 7.3.

Training Stage	Action Type	Robot Performance (%)
Training II	Iterative	83.3
	Non-Iterative	83.3
Training III	Iterative	100
	Non-Iterative	100

Table 7.3: Comparison of the robot performance, in terms of action execution, at the end of the second and third stage of the training

After the second stage of training, five out of six actions (for both iterative and non-iterative) are correctly categorized. The performance of the robot improves after the third stage of the training when all the six actions (both iterative and non-iterative) are correctly categorized (Tab. 7.3).

7.8.2 Incompatible Condition Test

The last test performed, consisted in analysing the response of the model in case of inconsistency between the linguistic and visual inputs. In particular, the incompatible noun condition was tested, in order to analyse the response of the model when the name of the object is incompatible with the object perceived by the robot (e.g. "**USE** [a] **KNIFE**" *became* "**USE** [a] **HAMMER**"). Activation values of hidden

units recorded during the compatible and incompatible conditions were analysed by performing the temporal hierarchical cluster analysis. The results of hierarchical clustering are presented in the dendrograms in figure 7.34, from which it is possible to observe that over time the hidden units recorded during the incompatible condition follow an activation pattern that is similar to the activation values recorded during the compatible condition (Fig.7.34).

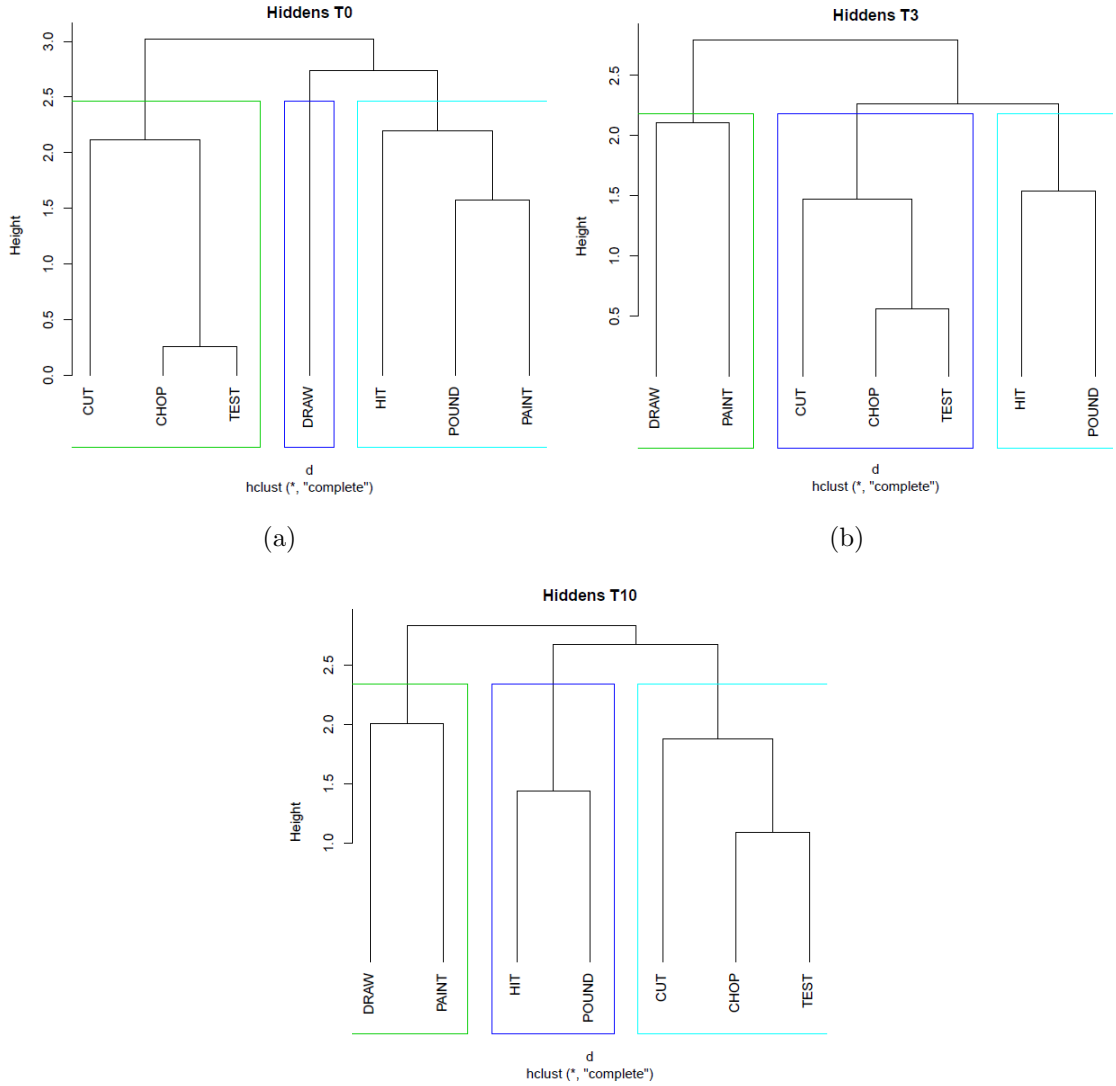


Figure 7.34: Incompatible Noun Condition (e.g. "**USE** [a] **KNIFE**" *became* "**USE** [a] **HAMMER**")). Results of the hierarchical clustering of hidden units activation values at the time steps T = 0 (a), T = 3 (b) and T = 10 (c)

In figure 7.34 results of the hierarchical clustering of activation values of hidden units at the time steps T = 0, T = 3 and T = 10 are presented. The dendrograms

in figure 7.34 compare the hidden activation values recorded during the compatible condition "**USE** [a] **KNIFE**" to the hidden activation values recorded during the incompatible condition "**USE** [a] **HAMMER**". In this particular case, the incompatibility is related to the KNIFE/HAMMER nouns. Despite that the robot sees a KNIFE, the word HAMMER is used to refer to the object. The hidden activation values related to the incompatible condition "**USE** [a] **HAMMER**" (that in the dendrograms are labelled "TEST") are clustered together with "**USE** [a] **KNIFE**". This means that the activation values of hidden units during this incompatible condition test are very close to the activation values of hidden units during the compatible condition.

Additionally, on the matrix associated to the activation values of hidden units recorded during the incompatible noun condition, the PCA has been performed (Fig.7.35).

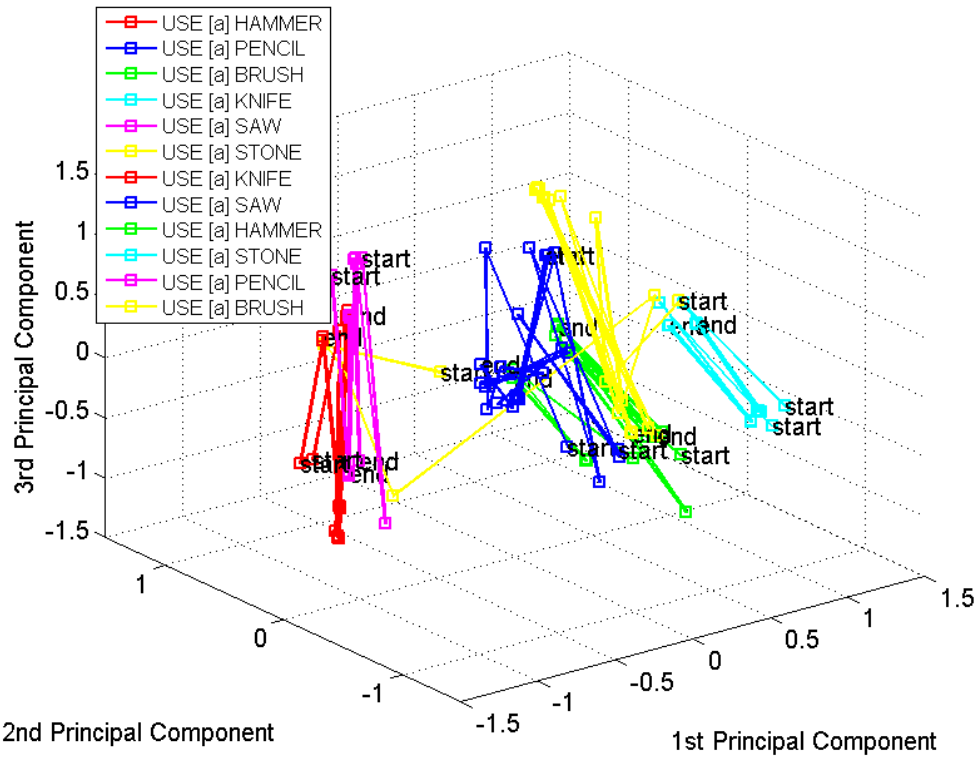
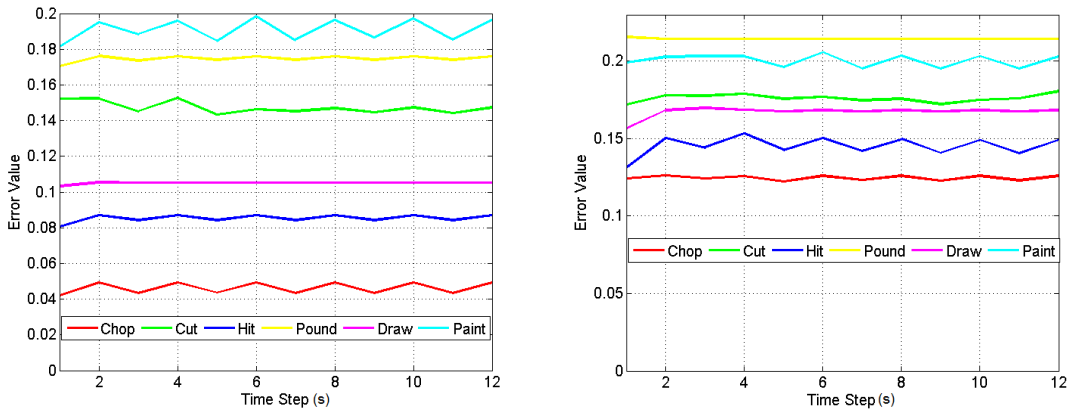


Figure 7.35: Trajectories of the activation values of hidden units recorded during the incompatible NOUN condition test compared to the trajectories of activation values recorded during the compatible condition

The trajectories of the activation values of hidden units in time, recorded during the incompatible condition test, in the space of the three principal components, have been compared to the trajectories of the activation values recorded during the compatible condition. From figure 7.35 it is possible to observe that the trajectories of activation values of hidden units recorded during the incompatible noun condition "**USE** [a] **HAMMER**" follow trajectories that are very similar to the trajectories recorded during the compatible condition "**USE** [a] **KNIFE**". The results obtained in the incompatible noun condition test has confirmed that in case of inconsistency between the perceptual and linguistic input, the robot executes the actions elicited by the seen objects. Furthermore, the error recorded at the end of each action execution in the compatible condition has been compared to the error recorded at the end of each action execution in the incompatible condition (Fig.7.36).



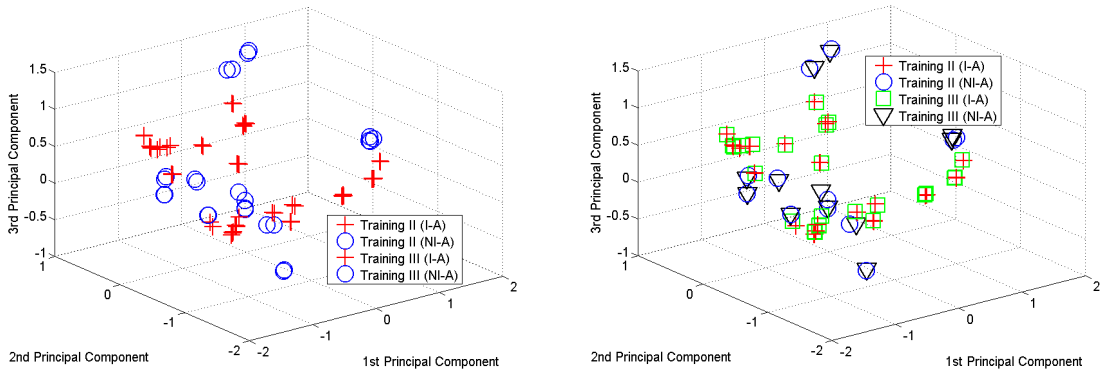
(a) (b)

Figure 7.36: Training stage III: MSE recorded during the execution of the iterative actions for: compatible condition (a) and incompatible NOUN condition (b)

In figure 7.36 the MSE recorded during the execution of the iterative actions for the compatible condition after the third training stage is compared to the MSE recorded during the incompatible noun condition. The higher error rates in the incompatible trial than in the compatible ones, suggest the proper naming of objects and actions support action categorization and that seeing objects automatically elicits the representations of their affordances.

7.8.3 Representations of Abstract Action Words

To better understand the internal dynamics of the model, after all the stages of the training were successfully completed, the activation values of hidden units were analysed. The Principal Component Analysis was performed on the hidden activation values. In figure 7.37 the hidden units activation values are represented in the space of the first three principal components. In particular, in figure 7.37(a) the observations plotted with the red markers are related to the activation values in the space of the three principal components recorded during the second and third stage of the training for the iterative actions (Training II (I-A) and Training III (I-A)), while the blue markers identify the activation values recorded during the second and third stage of the training for non-iterative actions (Training II (NI-A) and Training III (NI-A)). In figure 7.37(b) observations are displayed in four groups representing respectively the Training II for Iterative-Actions, Training II for Non-Iterative-Actions, Training III for Iterative-Actions and Training III for Non-Iterative-Actions.

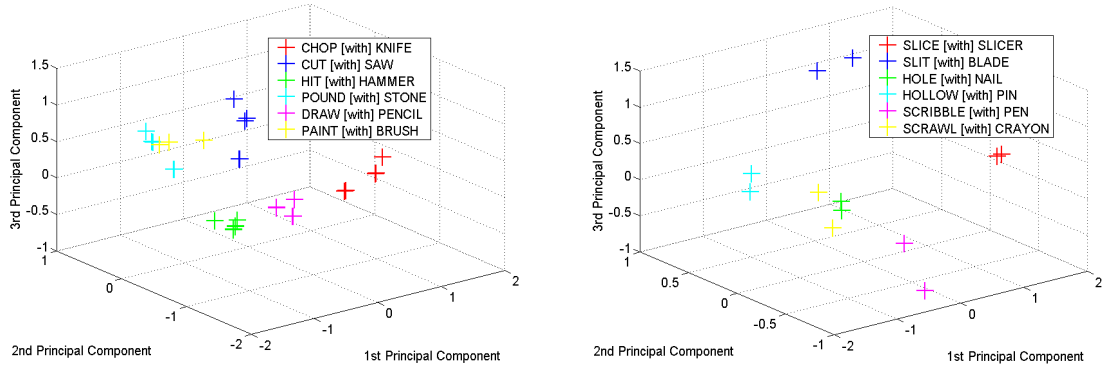


(a) (b)

Figure 7.37: Hidden units activation values in the space of the three principal components. Data displayed in two groups: Training II and Training III Iterative-Actions, Training II and Training III Non-Iterative-Actions (a). Data displayed in four groups: Training II Iterative-Actions, Training II Non-IterativeActions, Training III Iterative-Actions and Training III Non-Iterative-Actions (b)

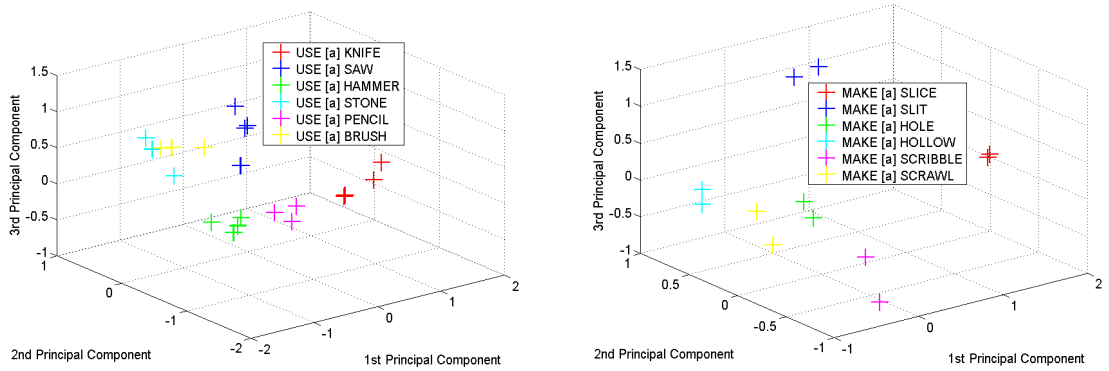
From figure 7.37(b) it is possible to notice that the observations related to the iterative actions recorded during the second and third stage of the training almost

fully overlap (data displayed by red and green markers). The same consideration can be done for non-iterative actions (data displayed by blue and black markers are overlapped). Hence, from figure 7.37(b) it is possible to conclude that hidden units during the second and third stage of the training follow a similar activation pattern. To better visualize the activation values of hidden units over time in the space of the three principal components, the hidden units activation values, recorded during different stages of the training, are shown in separate plots (Fig.7.38, Fig.7.39).



(a) (b)

Figure 7.38: Hidden units activation values in the space of the three principal components: Training II Iterative-Actions (a), Training II Non-Iterative-Actions (b)



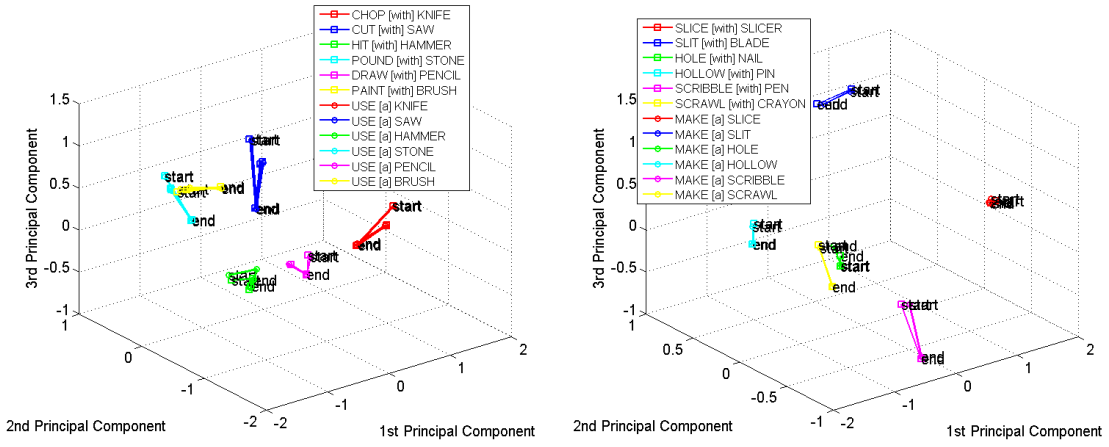
(a) (b)

Figure 7.39: Hidden units activation values in the space of the three principal components: Training III Iterative-Actions (a), Training III Non-Iterative-Actions (b)

In particular, in figure 7.38 the hidden units activation values in the space of the three principal components, recorded after the second stage of the training, for

the Iterative-Actions and Non-Iterative-Actions are shown. In figure 7.39 the hidden units activation values in the space of the three principal components, recorded after the third stage of the training, for the Iterative-Actions and Non-Iterative-Actions are shown. By comparing figure 7.38(a) to figure 7.39(a) and figure 7.38(b) to figure 7.39(b), it is possible to better visualise that the hidden units activation values during the second and third stage of the training follow a similar activation pattern.

In figure 7.40 the trajectories of hidden units activation values in the space of the three principal components are shown. Figure 7.40(a) presents the trajectories of hidden units during the Training II and Training III for Iterative-Actions, while figure 7.40(b) shows the trajectories of hidden units during the Training II and Training III for Non-Iterative-Actions.



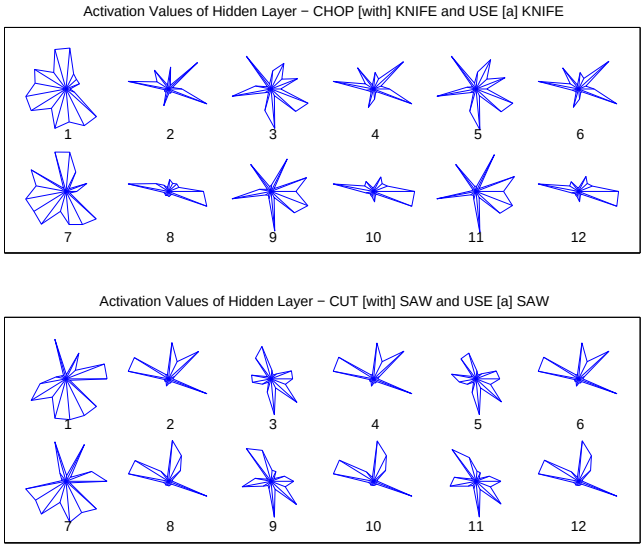
(a) (b)

Figure 7.40: Trajectories of hidden units activation values in the space of the three principal components: Training II and Training III Iterative-Actions (a), Training II and Training III Non-Iterative-Actions (b)

The results presented in figure 7.40 suggest that the acquisition of concepts related to abstract action words (e.g. "USE" and "MAKE") requires the reactivation of similar internal representations activated during the acquisition of the basic concepts contained in the hierarchical structure of words used to ground the abstract action words. In other words, the hidden units of the model, during the acquisition of abstract action words, follow similar activation patterns recorded during the acquisition of the basic concepts that are hierarchically organized to ground a par-

ticular abstract action word. The processing of abstract action words requires the same internal activation needed for the processing of basic concepts. This seems to suggest that even the semantic/conceptual representation of abstract action words consists of reusing sensorimotor and perceptual representational capabilities [Barsalou, 1999].

The start plots in figure 7.41 provide a visual representation of activation values of hidden units that permits the comparison of the the internal representations of the model in response to words directly linked to perceptual and sensorimotor experience to the internal representations of the model in response to abstract action words. Each plot in figure 7.41 consists of twelve stars arranged in two rows. The six stars of the first row visualise the hidden activation values recorded during the direct grounding of words (for different time steps) in perception and sensorimotor experience, while the six stars of the second row visualize the hidden activation values recorded during the grounding of abstract action words. The star plots of the first row compared to the star plots arranged in the second row permit to visually compare activation values of hidden units during different stages of the training. For example, the first plot permit to compare activation values of hidden units recorded for the inputs "**CHOP [with] KNIFE**" and "**USE [a] KNIFE**".



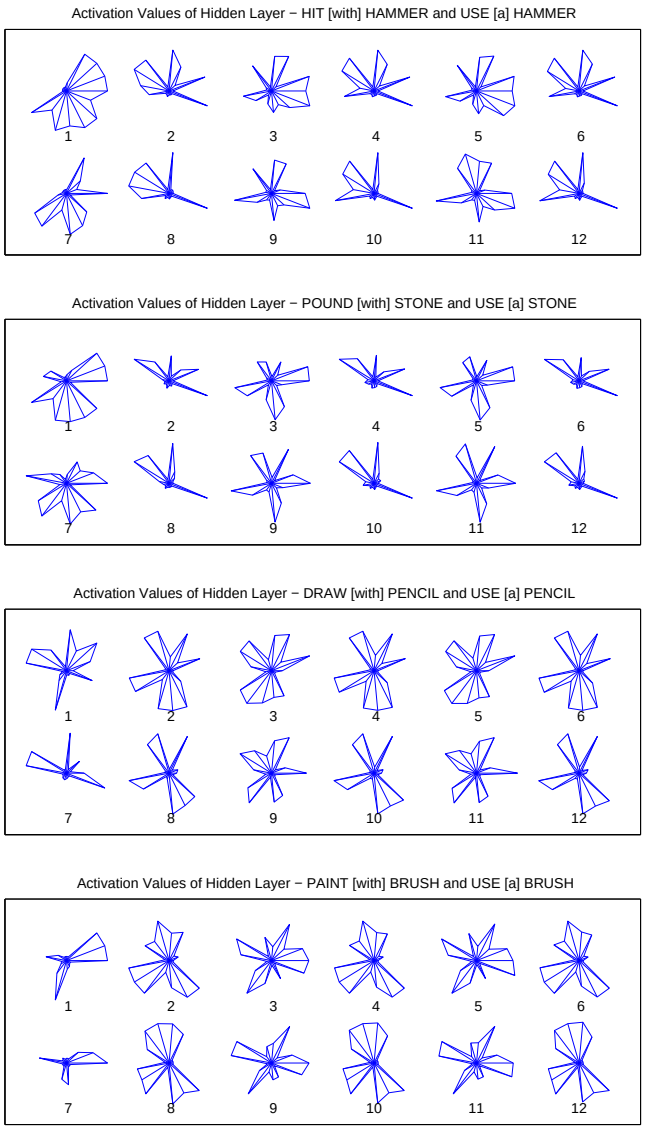


Figure 7.41: Star plots to visually compare activation values of hidden units in response to words directly linked to perceptual and sensorimotor experience and abstract action words

The visualization of the internal representation of the model in figure 7.41 confirms the high similarity between the activation values recorded during the second and third stage of the training.

7.9 Discussion

The presented study, through the implementation of an embodied multi-modal cognitive architecture, enabled the iCub to ground the meaning of abstract action

words in perceptual and sensorimotor knowledge and permitted the investigation of the relation between the development of conceptual knowledge and perceptual and sensorimotor categories in the iCub humanoid robot. The starting points considered in the implementation of the proposed cognitive model were the compositionality and embodiment of language, according to which higher-order concepts (i.e. abstract action words) can be grounded through the hierarchical organization of motor primitives and perceptual knowledge. The implemented architecture was based on simple recurrent neural networks which enabled the modelling of the mechanisms underlying motor and linguistic sequence processing. The training of the model was incremental and consisted of three stages that permitted to acquire perceptual and sensorimotor knowledge first, to learn words directly grounded in perceptual and sensorimotor knowledge then, and to acquire the meaning of abstract action words through the hierarchical organization of the words directly linked in perceptual and sensorimotor knowledge at the end. Simulation results showed that, at the end of the training, the robot was able to correctly categorize the perceptual, proprioceptive and linguistic inputs by performing the appropriate behaviour triggered by the linguistic input and the perceived object. The presence of clusters in the hidden units of the model suggest the formation of concepts from the multi-modal data received in input by the network. Additional tests have shown that the performance of the robot decreased in case the linguistic or visual inputs were not provided to the model. The robot showed the ability to generalize new concepts by receiving unlearned sentences and generating the appropriate corresponding behaviour. Results obtained in the incompatible condition tests showed that in case of inconsistency between the perceptual and linguistic input, the robot executed the actions elicited by the seen object. These results are consistent with recent evidence in neuroscience and behavioural sciences that has shown that visually perceived objects activate motor information [Jeannerod, 1994, Arbib, 1997]. Hence, the knowledge associated to objects relies not only on objects perceptual features but also on the actions (i.e. affordances) that can be performed on them. These results have suggested that

perceptual and sensorimotor inputs have a central role in reasoning and language understanding. Indeed, the performed simulations suggested that the acquisition of concepts related to abstract action words (e.g. "USE" and "MAKE") requires the reactivation of similar internal representations activated during the acquisition of the basic concepts, directly grounded in perceptual and sensorimotor knowledge, contained in the hierarchical structure of the words used to ground the abstract action words. This finding seems to suggest that the semantic/conceptual representation of abstract action words consists of reusing sensorimotor and perceptual representational capabilities [Barsalou, 1999] (embodied understanding of abstract language). Along this line of research, different theories proposed in psychology have claimed that embodiment plays an important role even in representing abstract concepts. These theories are based on "metaphors" [Lakoff and Johnson, 1980], "simulations" [Barsalou, 1999] and "actions" [Glenberg and Kaschak, 2002]. These different approaches to the embodiment of abstract language are not mutually exclusive and they might emphasize different aspects of the same phenomenon [Glenberg et al., 2008]. Recently, in [Glenberg et al., 2008] neurophysiological evidence for the modulation of the motor system activity during the comprehension of both concrete and abstract language has been provided. Results of this neurophysiological study have shown that the processing of words both concrete and abstract involves the modulation of the motor system. This means that the comprehension of words is likely to involve or require the simulation of the meaning represented by the corresponding concept. These results represent an important step forward in providing evidence for the embodied understanding of abstract language.

Chapter 8

Conclusion

This thesis addressed the problem of **Grounding Abstract Categories in Cognitive Robots**; the implementation of developmental neuro-robotics models permitted to investigate the relations between the development of abstract symbolic representations (e.g. *language*) and sensorimotor knowledge (e.g. *action* and *vision*). Three experimental studies on the grounding of abstract categories in cognitive robots were presented.

The first experiment, based on a feed-forward artificial neural network, permitted to test the training methodology adopted for the grounding of language in humanoid robots. This model teaches the robot the meaning of words that lack of a direct concrete referent such as "'ACCEPT" and "REJECT". The training of this model was effective although some limitations of its implementation were evident.

In the second experiment, the architecture adopted for carrying out the first study was reimplemented by using recurrent artificial neural networks that permitted to specify temporally the action primitives to be executed by the robot. This permitted to increase the combinations of actions that can be taught to the robot for the generation of more complex movements. The neural network controller implemented for this study enabled the learning of higher-order concepts based on sequences of low-level primitives. Simulation results showed that higher-order symbolic representations can be indirectly grounded in action primitives, which are

themselves directly grounded in sensorimotor experience. Through the analysis of the network dynamics, it has been observed that motor primitives show different activation patterns according to the action's sequence in which they are embedded. These simulation results are consistent with empirical neuroscience and computational neuroscience studies on action representation that showed that the goal of an action changes the substrate of neurons involved in the action processing.

In the third experiment, an embodied multi-modal cognitive architecture enabled the iCub to ground the meaning of abstract action words in perceptual and sensorimotor knowledge and permitted the investigation of the relation between the development of conceptual knowledge and perceptual and sensorimotor categories in the iCub humanoid robot. Simulation results showed that the ability of the robot to correctly categorize the perceptual, proprioceptive and linguistic inputs by performing the appropriate behaviour triggered by the linguistic input and the perceived object decreased in case the linguistic or visual inputs were not provided. The robot showed the ability to generalize new concepts by receiving un-learned sentences and generating the appropriate corresponding behaviour. Moreover, results obtained in the incompatible condition tests showed that in case of inconsistency between the perceptual and linguistic input, the robot executed the actions elicited by the seen object. These results are consistent with recent evidence in neuroscience and behavioural sciences that has shown that visually perceived objects activate motor information [Jeannerod, 1994, Arbib, 1997]. Hence, the knowledge associated to objects relies not only on objects perceptual features but also on the actions (i.e. affordances) that can be performed on them.

The performed simulations suggested that the acquisition of concepts related to abstract action words (e.g. "USE" and "MAKE") requires the reactivation of similar internal representations activated during the acquisition of the basic concepts, directly grounded in perceptual and sensorimotor knowledge, contained in the hierarchical structure of the words used to ground the abstract action words. This finding seems to suggest that the semantic/conceptual representation of abstract

action words consists of reusing sensorimotor and perceptual representational capabilities [Barsalou, 1999] (embodied understanding of abstract language). Indeed, in [Glenberg et al., 2008] neurophysiological evidence for the modulation of the motor system activity during the comprehension of both concrete and abstract language has been provided. Results of this neurophysiological study have shown that the processing of words both concrete and abstract involves the modulation of the motor system. This means that the comprehension of words is likely to involve or require the simulation of the meaning represented by the corresponding concept. These results represent an important step forward in providing evidence for the embodied understanding of abstract language.

8.1 Future Work

Future research, following the developmental and neural paradigm applied to robotics, will consider the gradual development observed in human beings as a potential road-map for artificial systems; this will permit the implementation of new neuro-robotics models that can account for other aspects of cognitive development observed in humans. Indeed, some aspects of the presented research, that can be further addressed and investigated, are listed below:

- Developmental and ecological model to ground the meaning of language in tool affordances discovered via Statistical Inference. Despite it is clear that language has to be grounded in sensorimotor experience, it is also important to go beyond simple sensorimotor grounding. To this end, statistical inference will be adopted as an original and innovative methodology that can serve in grounded theories of meaning. Embodied theories of meanings in a probabilistic framework can lead to "hybrid models" in which some concepts are directly grounded in a robot's sensorimotor experience while, for other concepts, statistical inference will permit to go beyond the available data and acquire new concepts.

- Human-Robot Interaction (HRI) study to understand whether the iCub cognitive architecture exhibits a number of brain compatible features and to investigate if the robot participates in a shared task as a plausible interaction partner. The study will permit to analyse how people experience interaction through language with a humanoid robot. In particular, the proposed study will permit to validate the implemented cognitive architecture on a number of benchmarks, such as affordance understanding, participation into a shared task and language to action mapping.

The presented future research directions will aim at finding a general mechanism that permits to ground the meaning of action words and simple sentences (e.g. phrases composed by an action verb and an object name) in tool affordances. The implementation of this architecture will enable the learning of action words by discovering new affordances related to objects and the environment. Indeed, inspired by the Indexical Hypothesis [Glenberg and Robertson, 2000], the acquisition of action words in the iCub humanoid robot will be achieved through the following steps: (i) direct grounding of object and action names to their referents; (ii) learning the "*stable affordances*" [Borghi and Riggio, 2009] for the grounded words through exploratory behaviour; (iii) discovering new affordances (including "*variable affordances*" [Borghi and Riggio, 2009]) and word meanings through statistical inference. The step (i) will permit to gather the representation of object and action names as perceptual symbols, which will endow the robot with some basic perceptual and motor skills to be reused for bootstrapping the learning of more complex behaviours. The step (ii), through exploratory behaviours, will enable the learning of stable affordances by performing several trials for each $<object, action>$ pairs and observing the consequent effects; exploratory behaviours also permit to discover new more efficient ways of interaction while performing actions (e.g. different grasp types). Affordances can be modelled by perceiving the effects of actions executed on objects and then, by categorizing them according to the obtained effect. The effect of actions can be modelled either in terms of changes produced on the object fea-

tures (e.g. visual segmentation) after performing the action or in terms of changes perceived in the state of the robot sensors by creating an association between one action and the resulting perceptual consequences (e.g. direction of motion, tactile activation, force applied and felt). After learning affordances, every time a specific effect is required (i.e. goal), the appropriate action that matches the affordances of the object present in the environment will be executed. In other words, the learned effect of actions on objects can be used to drive goal-directed behaviour. During step (iii), perceptual symbols embedded in a probabilistic framework will produce new knowledge in response to novel data collected from the environment.

Artificial neural networks (e.g. Recurrent Neural Networks) will serve for the integration of temporal sequences of linguistic and motor primitives and to model the underlying mechanisms. Probabilistic graphical models (e.g. Bayesian Networks) will serve to model affordances by learning the casual relations between object features, actions and effects. As observed in [Barsalou, 2008], Bayesian statistics provide a powerful tool that can be viewed as statistical accounts of the multi-modal information stored in the dynamic systems that generate simulations and guide situated action. Behavioural experiments related to object's affordances evoked by linguistic stimuli, can be replicated to generate new predictions. For example, in [Borghi and Riggio, 2009] the effects of sentences comprehension (i.e. phrases composed by a verb and an object name) on different kind of affordances (i.e. precision and power grip) have been investigated. Results of the study indicated that sentences comprehension activated a mental simulation that led to the formation of a "motor prototype" which reflects stable affordances of the object (i.e. the typical way the object is acted upon) and the "canonical" aspects of variable affordances (i.e. the canonical object orientation). Moreover, in [Tucker and Ellis, 2004] compatibility effects between object size (small and large) and the kind of grip (i.e. precision and power grip) used to respond whether seen objects were artefacts or natural objects have been found. The model can be used to formulate and test new compatibility effects between language processing and object's affordances.

Bibliography

A. Alishahi. Computational modeling of human language acquisition. *Synthesis Lectures on Human Language Technologies*, 3(1):1–107, 2010.

J. Altarriba, L.M. Bauer, and C. Benvenuto. Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods*, 31(4):578–602, 1999.

M. Andrews, G. Vigliocco, and D. Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463, 2009.

M. A. Arbib. From visual affordances in monkey parietal cortex to hippocampo–parietal interactions underlying rat navigation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1360):1429–1436, 1997.

M. A. Arbib. fo the mirror system, imitation, and the evolution of language. *Imitation in animals and artifacts*, page 229, 2002.

M. A. Arbib, P. Érdi, and J. Szentágothai. *Neural organization: Structure, function, and dynamics*. The MIT Press, 1998.

M. A. Arbib et al. From grasp to language: embodied concepts and the challenge of abstraction. *Journal of physiology, Paris*, 102(1-3):4–20, 2008.

M. Asada, K.F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37(2-3):185–193, 2001.

M.D. Barrett. *The development of language*. Psychology Pr, 1999.

L. W. Barsalou, A. Santos, W. K. Simmons, and C. D. Wilson. Language and simulation in conceptual processing. *Symbols, embodiment, and meaning*, pages 245–283, 2008.

L.W. Barsalou. Perceptual symbol systems. *Behavioral and brain sciences*, 22(04): 577–660, 1999.

L.W. Barsalou. Situated simulation in the human conceptual system. *Conceptual Representation*, pages 513–562, 2003.

L.W. Barsalou. Grounded cognition. *Annual Review of Psychology*, 59:617–645, 2008.

L.W. Barsalou. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1281–1289, 2009.

L.W. Barsalou and K. Wiemer-Hastings. Situating abstract concepts. *Grounding cognition: The role of perception and action in memory, language, and thinking*, pages 129–163, 2005.

L.W. Barsalou, W. Kyle Simmons, A.K. Barbey, and C.D. Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, 7(2):84–91, 2003.

A. Blondin-Massé, S. Harnad, O. Picard, and B. St-Louis. Symbol grounding and the origin of language: From show to tell. 2010.

P. Bloom. *How children learn the meanings of words: Learning, development and conceptual change*. Cambridge, MA: MIT Press, 2002.

A. M. Borghi and L. Riggio. Sentence comprehension and simulation of object temporary, canonical and stable affordances. *Brain Research*, 1253:117–128, 2009.

A. M. Borghi, N. Caramelli, and A. Setti. Conceptual information on objects locations. *Brain and language*, 93(2):140–151, 2005.

A. M. Borghi, A. Flumini, F. Cimatti, D. Marocco, and C. Scorolli. Manipulating objects and telling words: a study on concrete and abstract words acquisition. *Frontiers in Psychology*, 2011.

A.M. Borghi and F. Cimatti. Words as tools and the problem of abstract words meanings. In *Proceedings of the 31st annual conference of the cognitive science society*, volume 31, pages 2304–2309, 2009.

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

M. M. Botvinick and D. C. Plaut. Short-term memory for serial order: A recurrent neural network model. *Psychological Review*, 113(2):201–233, 2006.

J. S. Bruner and G. A. Austin. *A study of thinking*. Transaction Books, 1986.

G. Buccino, L. Riggio, G. Melli, F. Binkofski, V. Gallese, and G. Rizzolatti. Listening to action-related sentences modulates the activity of the motor system: a combined tms and behavioral study. *Cognitive Brain Research*, 24(3):355–363, 2005.

J. Bullinaria. Analyzing the internal representations of trained neural networks. *Neural Network Analysis, Architectures and Algorithms*, pages 3–26, 1997.

D. Caligiore and M. H. Fischer. Vision, action and language unified through embodiment. *Psychological Research*, 77(1):1–6, 2013.

D. Caligiore, A. M. Borghi, D. Parisi, and G. Baldassarre. Affordances and compatibility effects: a neural-network computational model. In *Connectionist models of behaviour and cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop*, pages 15–26, 2009.

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

A. Cangelosi. Symbol grounding in connectionist and adaptive agent models. *New Computational Paradigms*, pages 69–74, 2005.

A. Cangelosi. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of life reviews*, 7(2):139–151, 2010.

A. Cangelosi and D. Parisi. *Simulating the evolution of language*. Springer London, 2002.

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

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

A. Cangelosi and M. Schlesinger. *Developmental Robotics: From Babies to Robots*. Cambridge MA: MIT Press, 2014.

A. Cangelosi, E. Hourdakis, and V. Tikhanoff. Language acquisition and symbol grounding transfer with neural networks and cognitive robots. *International Joint Conference on Neural Networks*, pages 1576–1582, 2006.

A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, et al. Integration of action and language knowledge: A roadmap for developmental robotics. *Autonomous Mental Development, IEEE Transactions on*, 2(3):167–195, 2010.

N. Caramelli, A. Setti, and D. D. Maurizzi. Concrete and abstract concepts in school age children. *Psychology of Language and Communication*, 8(2):17–32, 2004.

F. Chersi, A. Mukovskiy, L. Fogassi, P. F. Ferrari, and W. Erlhagen. A model of intention understanding based on learned chains of motor acts in the parietal lobe. *Computational Neuroscience*, 69:48, 2006.

F. Chersi, S. Thill, T. Ziemke, and A.M. Borghi. Sentence processing: linking language to motor chains. *Frontiers in Neurorobotics*, 4, 2010.

N. Chomsky. *Principles and parameters in syntactic theory*. 1979.

P. F. Dominey, A. Mallet, and E. Yoshida. Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *International Journal of Humanoid Robotics*, 6(02):147–171, 2009.

J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

T. Eynon. Cognitive linguistics. *Advances in Psychiatric Treatment*, 8(6):399–407, 2002.

I. Farkaš, T. Malík, and K. Rebrová. Grounding the meanings in sensorimotor behavior using reinforcement learning. *Frontiers in Neurorobotics*, 6:1, 2012.

I. K. Fodor. A survey of dimension reduction techniques, 2002.

J. A. Fodor. *The language and thought*. Harvard University Press, 1975.

J. A. Fodor. *Concepts: Where cognitive science went wrong*. Oxford University Press, USA, 1998.

J. A. Fodor and E. Lepore. *The compositionality papers*. Oxford University Press, 2002.

L. Fogassi, P. F. Ferrari, B. Gesierich, S. Rozzi, F. Chersi, and G. Rizzolatti. Parietal lobe: from action organization to intention understanding. *Science*, 308(5722): 662, 2005.

V. Gallese and G. Lakoff. The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, 22(3-4):455–479, 2005.

V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119(2):593–609, 1996.

D. Gentner. *Why nouns are learned before verbs: Linguistic relativity versus natural partitioning*. University of Illinois at Urbana-Champaign, 1982.

J. J. Gibson. The concept of affordances. *Perceiving, acting, and knowing*, pages 67–82, 1977.

A. M. Glenberg. What memory is for. *Behavioral and Brain Sciences*, 20(1):1–19, 1997.

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

A. M. Glenberg and D. A. Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401, 2000.

A. M. Glenberg, M. Sato, L. Cattaneo, L. Riggio, D. Palumbo, and G. Buccino. Processing abstract language modulates motor system activity. *The Quarterly Journal of Experimental Psychology*, 61(6):905–919, 2008.

A. I. Goldman. *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford University Press, 2006.

M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

A. Greco, T. Riga, and A. Cangelosi. The acquisition of new categories through grounded symbols: An extended connectionist model. In *Artificial Neural Networks and Neural Information ProcessingICANN/ICONIP 2003*, pages 763–770. Springer, 2003.

S. Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42 (1-3):335–346, 1990.

S. Harnad. From sensorimotor categories and pantomime to grounded symbols and propositions. 2010.

S. Harnad, S. J. Hanson, and J. Lubin. Categorical perception and the evolution of supervised learning in neural nets. *Cognition and Brain Theory*, 5:29–47, 1991.

O. Hauk, I. Johnsrude, and F. Pulvermüller. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307, 2004.

D. O. Hebb. *The organization of behavior: A neuropsychological approach*. John Wiley & Sons, 1949.

G. E. Hinton and T. Shallice. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological review*, 98(1):74–95, 1991.

J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

IFR. Executive summary world robotics 2012, industrial robots and service robots. *International Federation of Robotics (IFR)*, 2012.

M. Jeannerod. The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain sciences*, 17(2):187–201, 1994.

M. Jeannerod. *The cognitive neuroscience of action*. Blackwell Oxford, 1997.

M. I. Jordan. Serial order: A parallel distributed processing approach. *Institute for Cognitive Science Report 8604, University of California, San Diego*, 1986.

S. Kalkan, N. Dag, O. Yürüten, A. M. Borghi, and E. Sahin. Verb concepts from affordances. *to appear*.

M. P. Kaschak and A. M. Glenberg. Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of memory and language*, 43(3):508–529, 2000.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

S. T. Kousta, G. Vigliocco, D. P. Vinson, M. Andrews, and E. Del Campo. The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, 140(1):14, 2011.

G. Lakoff and M. Johnson. *Metaphors we live by*, volume 111. Chicago London, 1980.

T. K. Landauer and S. T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

Y. LeCun. Learning process in an asymmetric threshold network. In *Disordered systems and biological organization*, pages 233–240. Springer, 1986.

Y. Li, H. Chen, and Z. Wu. Dynamic time warping distance method for similarity test of multipoint ground motion field. *Mathematical Problems in Engineering*, 2010, 2010.

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological review*, 74(6):431, 1967.

P. Lieberman. On the nature and evolution of the neural bases of human language. *American Journal of Physical Anthropology*, 119(S35):36–62, 2002.

M. M. Louwerse and P. Jeuniaux. The linguistic and embodied nature of conceptual processing. *Cognition*, 114(1):96–104, 2010.

C. D. Mah and F. A. Mussa-Ivaldi. Evidence for a specific internal representation of motion–force relationships during object manipulation. *Biological cybernetics*, 88(1):60–72, 2003.

D. Marocco, A. Cangelosi, K. Fischer, and T. Belpaeme. Grounding action words in the sensorimotor interaction with the world: experiments with a simulated icub humanoid robot. *Frontiers in Neurorobotics*, 4, 2010.

D. Marr. Vision: A computational investigation into the human representation and processing of visual information, henry holt and co. *Inc., New York, NY*, 1982.

J. L. McClelland, D.E. Rumelhart, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 2: Psychological and Biological Models, Cambridge, MA: MIT PressMass, 1986.

W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, 1943.

B. McGhee-Bidlack et al. The development of noun definitions: A metalinguistic analysis. *Journal of Child Language*, 18(02):417–434, 1991.

C. B. Mervis and E. Rosch. Categorization of natural objects. *Annual review of psychology*, 32(1):89–115, 1981.

G. Metta, G. Sandini, L. Natale, R. Manzotti, and F. Panerai. Development in artificial systems. In *Proceedings of the EDEC Symposium at the International Conference on Cognitive Science*, 2001.

G. Metta, P. Fitzpatrick, and L. Natale. Yarp: yet another robot platform. *International Journal on Advanced Robotics Systems*, 3(1):43–48, 2006.

G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori. The icub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pages 50–56. ACM, 2008.

O. Michel. Webotstm: Professional mobile robot simulation. *arXiv preprint cs/0412052*, 2004.

M. Minsky and S. Papert. Perceptrons. 1969.

A. F. Morse, T. Belpaeme, A. Cangelosi, and L. B. Smith. Thinking with your body: Modelling spatial biases in categorization using a real humanoid robot. In *Proc. of 2010 annual meeting of the Cognitive Science Society. Portland, USA*, pages 1362–1368, 2010.

F. A. Mussa-Ivaldi and E. Bizzi. Motor learning through the combination of primitives. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 355(1404):1755–1769, 2000.

N. J. Nilsson. Shakey the robot. Technical report, DTIC Document, 1984.

R. M. Nosofsky, J. K. Kruschke, and S. C. McKinley. Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):211, 1992.

P. Y. Oudeyer and F. Kaplan. Discovering communication. *Connection Science*, 18 (2):189–206, 2006.

E. Oztop and M. A. Arbib. Schema design and implementation of the grasp-related mirror neuron system. *Biological cybernetics*, 87(2):116–140, 2002.

E. Oztop, N. S. Bradley, and M. A. Arbib. Infant grasp learning: a computational model. *Experimental Brain Research*, 158(4):480–503, 2004.

A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(12):1–25, 1968.

D. B. Parker. Learning logic. 1985.

A. Parmiggiani, M. Maggiali, L. Natale, F. Nori, A. Schmitz, N. Tsagarakis, J. S. Victor, F. Becchi, G. Sandini, and G. Metta. The design of the icub humanoid robot. *International Journal of Humanoid Robotics*, 9(04), 2012.

U. Pattacini, F. Nori, L. Natale, G. Metta, and G. Sandini. 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)*, pages 1668–1674. IEEE, 2010.

B. Pearlmutter. Dynamic recurrent neural networks. 1990.

D. Perani, S. F. Cappa, T. Schnur, M. Tettamanti, S. Collina, M. M. Rosa, and F. Fazio. The neural correlates of verb and noun processing. *Brain*, 122(12): 2337–2344, 1999.

R. Pfeifer, J. Bongard, and S. Grand. *How the body shapes the way we think: a new view of intelligence*. The MIT Press, 2007.

J. Piaget and M. T. Cook. The origins of intelligence in children. 1952.

F. J. Pineda. Generalization of back-propagation to recurrent neural networks. *Physical review letters*, 59(19):2229–2232, 1987.

S. Pinker. *The language instinct: How the mind creates language*. HarperCollins e-books, 2010.

S. Pinker and A. Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1):73–193, 1988.

F. Pulvermüller. Words in the brain's language. *Behavioral and brain sciences*, 22 (2):253–279, 1999.

F. Pulvermüller. *The neuroscience of language: on brain circuits of words and serial order*. Cambridge University Press, 2003.

F. Pulvermüller. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7):576–582, 2005.

F. Pulvermüller and L. Fadiga. Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, 11(5):351–360, 2010.

F. Pulvermüller, M. Härle, and F. Hummel. Walking or talking?: Behavioral and neurophysiological correlates of action verb processing* 1. *Brain and language*, 78(2):143–168, 2001.

F. Pulvermüller, O. Hauk, V. V. Nikulin, and R. J. Ilmoniemi. Functional links between motor and language systems. *European Journal of Neuroscience*, 21(3): 793–797, 2005.

D. C. Richardson, M.J. Spivey, L. W. Barsalou, and K. McRae. Spatial representations activated during real-time comprehension of verbs. *Cognitive science*, 27 (5):767–780, 2003.

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

G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141, 1996a.

G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive brain research*, 3(2):131–141, 1996b.

E. Rosch. Principles of categorization. *Concepts: core readings*, pages 189–206, 1999.

E. H. Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.

F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

D. Roy. Grounding words in perception and action: computational insights. *Trends in cognitive sciences*, 9(8):389–396, 2005a.

D. Roy. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1):170–205, 2005b.

D. E. Rumelhart, J. L. McClelland, and PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 1: Foundations, Cambridge, MA: MIT Press, 1986.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 1:213, 2002.

S. J. Russell, P. Norvig, J. F. Canny, J. M. Malik, and D. D. Edwards. *Artificial intelligence: a modern approach*, volume 74. Prentice hall Englewood Cliffs, 1995.

H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 1:43–49, 1978.

G. Sandini, G. Metta, and D. Vernon. The icub cognitive humanoid robot: An open-system research platform for enactive cognition. *50 years of artificial intelligence*, pages 358–369, 2007.

P. J. Schwanenflugel. Why are abstract concepts hard to understand? 1991.

C. Scorolli and A. M. Borghi. Sentence comprehension and action: Effector specific modulation of the motor system. *Brain research*, 1130:119–124, 2007.

C. Scorolli and A. M. Borghi. Language and embodiment. *Anthropology and Philosophy*, 9(1-2):7–23, 2008.

J. R. Searle. Minds, brains, and programs. *Behavioral and brain sciences*, 3(3): 417–424, 1980.

D. J. Simons and F. C. Keil. An abstract to concrete shift in the development of biological thought: the insides story. *Cognition*, 56(2):129–163, 1995.

L. B. Smith and L. Samuelson. Objects in space and mind: from reaching to words, 2010.

L. Steels. Language games for autonomous robots. *Intelligent Systems, IEEE*, 16 (5):16–22, 2001.

L. Steels and F. Kaplan. Aibo's first words: The social learning of language and meaning. *Evolution of Communication*, 4(1):3–32, 2002.

L. Steels, F. Kaplan, A. McIntyre, and J. Van Looveren. Crucial factors in the origins of word-meaning. *The transition to language*, 2(1):4–2, 2002.

R. J. Sternberg. *Cognitive psychology*. Cengage Learning, 2009.

F. Stramandinoli, A. Cangelosi, and D. Marocco. 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. ©2011 IEEE.

F. Stramandinoli, D. Marocco, and A. Cangelosi. The grounding of higher order concepts in action and language: A cognitive robotics model. *Neural Networks*, 32:165–173, 2012.

Y. Sugita and J. Tani. Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, 13(1):33, 2005.

M. Tajine and D. Elizondo. The recursive deterministic perceptron neural network. *Neural Networks*, 11(9):1571–1588, 1998.

T. Tardif, P. Fletcher, W. Liang, Z. Zhang, N. Kaciroti, and V. A. Marchman. Baby's first 10 words. *Developmental Psychology*, 44(4):929, 2008.

J. B. Tenenbaum, T. L. Griffiths, and C. Kemp. Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318, 2006.

M. Tettamanti, G. Buccino, M. C. Saccuman, V. Gallese, M. Danna, P. Scifo, F. Fazio, G. Rizzolatti, S. F. Cappa, and D. Perani. Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, 17(2):273–281, 2005.

V. Tikhanoff, A. Cangelosi, P. Fitzpatrick, G. Metta, L. Natale, and F. Nori. 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*, pages 57–61. ACM, 2008.

V. Tikhanoff, A. Cangelosi, and G. Metta. Integration of speech and action in humanoid robots: icub simulation experiments. *IEEE Transactions on Autonomous Mental Development*, 3(1):17–29, 2011.

M. Tomasello. *First verbs: A case study of early grammatical development*. Cambridge University Press, 1992.

M. Tomasello. *Constructing a language*. Harvard University Press, 2003.

M. Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, 2009.

M. Tomasello and P.J. Brooks. Early syntactic development: A construction grammar approach. 1999.

M. Tomasello, M. Carpenter, J. Call, T. Behne, H. Moll, et al. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and brain sciences*, 28(5):675–690, 2005.

M. Tucker and R. Ellis. Action priming by briefly presented objects. *Acta psychologica*, 116(2):185–203, 2004.

M. Tucker, R. Ellis, et al. On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology-Human Perception and Performance*, 24(3):830–846, 1998.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

F. Warneken and M. Tomasello. Helping and cooperation at 14 months of age. *Infancy*, 11(3):271–294, 2007.

L. N. Wauters, A. Tellings, W. HJ Van Bon, and A. W. Van Haaften. Mode of acquisition of word meanings: The viability of a theoretical construct. *Applied Psycholinguistics*, 24(3):385–406, 2003.

P. Werbos. Beyond regression: New tools for prediction and analysis in the behavioral sciences. 1974.

S. Wermter, E. Riloff, and G. Scheler. Connectionist, statistical and symbolic approaches to learning for natural language processing. 1996.

B. Widrow, M. E. HOFF, et al. Adaptive switching circuits. 1960.

K. Wiemer-Hastings and Xu Xu. Content differences for abstract and concrete concepts. *Cognitive Science*, 29(5):719–736, 2005.

K. Wiemer-Hastings, J. Krug, and X. Xu. Imagery, context availability, contextual constraint, and abstractness. In *Proceedings of the 23rd annual conference of the cognitive science society*, pages 1134–1139, 2001.

M. Wilson. Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4): 625–636, 2002.

Y. Yamashita and J. Tani. Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS computational biology*, 4(11):e1000220, 2008.

O. Yürüten, K. F. Uyanık, Y. Çalışkan, A. K. Bozcuoğlu, E. Şahin, and S. Kalkan. Learning adjectives and nouns from affordances on the icub humanoid robot. In *From Animals to Animats 12*, pages 330–340. Springer, 2012.

R. A. Zwaan and L. J. Taylor. Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, 135(1): 1, 2006.

References of Selected Publications

- **2012.** Stramandinoli F., Cangelosi A., Wermter S., *"Special issue on advances in developmental robotics"*, In Paladyn. Journal of Behavioral Robotics, vol.3, n.3, p.112, DOI: http://dx.doi.org/10.2478/s13230-013-0112-x, SP Versita

- **2012.** Stramandinoli F., Marocco D., Cangelosi A., *"The Grounding of Higher Order Concepts in Action and Language: a Cognitive Robotics Model"*, In Neural Networks, vol. 32, pp 165–173, DOI: http://dx.doi.org/10.1016/j.neunet.2012.02.012

- **2011.** Stramandinoli F., Cangelosi A., Marocco D., *"Towards the Grounding of Abstract Words: A Neural Network Model for Cognitive Robots"*, In Proceedings of the 2011 International Joint Conference on Neural Networks, San Jose, pp. 467–474, DOI: http://dx.doi.org/10.1109/IJCNN.2011.6033258

- **2011.** Stramandinoli F., Rucinski M., Znajdek J., Rohlfing K.J., Cangelosi A., *"From Sensorimotor Knowledge to Abstract Symbolic Representations"*, In Procedia Computer Science, FET11 - The European Future Technologies Conference and Exhibition, (**Second Prize, Best Poster**), Budapest, 4-6 May, vol. 7, pp. 269–271, DOI: http://dx.doi.org/10.1016/j.procs.2011.09.018