International Journal of Advanced Robotic Systems

OPEN ACCESS
ARTICLE

Embodied Language Learning and Cognitive Bootstrapping: Methods and Design Principles

Invited Feature Article

Caroline Lyon[1]*, Chrystopher L. Nehaniv[1], Joe Saunders[1], Tony Belpaeme[2], Ambra Bisio[3], Kerstin Fischer[4], Frank Förster[1], Hagen Lehmann[1,5], Giorgio Metta[5], Vishwanathan Mohan[6], Anthony Morse[2], Stefano Nolfi[7], Francesco Nori[5], Katharina Rohlfing[8], Alessandra Sciutti[6], Jun Tani[9], Elio Tuci[7], Britta Wrede[8], Arne Zeschel[4] and Angelo Cangelosi[2]

1 Adaptive Systems Research Group, University of Hertfordshire, UK
2 Center for Robotics and Neural Systems, Plymouth University, UK
3 Dept. of Experimental Medicine, University of Genoa, Italy
4 Dept. for Design and Communication, University of Southern Denmark, Denmark
5 Italian Institute of Technology, iCub Facility, Genoa, Italy
6 Italian Institute of Technology, Robotics, Brain and Cognitive Science, Genoa, Italy
7 Institute of Cognitive Science and Technology, National Research Council, Rome, Italy
8 Applied Computer Science Group, University of Bielefeld, Germany
9 Department of Electrical Engineering, KAIST, South Korea
*Corresponding author(s) E-mail: c.lyon@herts.ac.uk

Received 19 October 2015; Accepted 04 April 2016

DOI: 10.5772/63462

Abstract

Co-development of action, conceptualization and social interaction mutually scaffold and support each other within a virtuous feedback cycle in the development of human language in children. Within this framework, the purpose of this article is to bring together diverse but complementary accounts of research methods that jointly contribute to our understanding of cognitive development and in particular, language acquisition in robots. Thus, we include research pertaining to developmental robotics, cognitive science, psychology, linguistics and neuroscience, as well as practical computer science and engineering. The different studies are not at this stage all connected into a cohesive whole; rather, they are presented to illuminate the need for multiple different approaches that complement each other in the pursuit of understanding cognitive development in robots. Extensive experiments involving the humanoid robot iCub are reported, while human learning relevant to developmental robotics has also contributed useful results.

Disparate approaches are brought together via common underlying design principles. Without claiming to model human language acquisition directly, we are nonetheless inspired by analogous development in humans and consequently, our investigations include the parallel co-development of action, conceptualization and social interaction. Though these different approaches need to ultimately be integrated into a coherent, unified body of knowledge, progress is currently also being made by pursuing individual methods.

Keywords Robot Language, Human Robot Interaction, HRI, Developmental Robotics, Cognitive Bootstrapping, Statistical Learning

1. Introduction

This article presents a contribution to the field of robot language learning and cognitive bootstrapping. Our goals are to develop artificially embodied agents that can acquire behavioural, cognitive and linguistic skills through individual and social learning.

Co-development of action, conceptualization and social interaction mutually scaffold and support each other within a virtuous feedback cycle in the development of human language in children. Language requires the bringing together of many different processes and we draw attention to the need for an interdisciplinary approach in this context. Thus, we include work in developmental robotics, cognitive science, psychology, linguistics and neuroscience, as well as practical computer science and engineering. Much of the research described in this paper was initiated in the EU ITALK project, undertaken within six universities in Europe, with collaborators in the US and Japan [1]. Extensive experiments involving the iCub humanoid robot are reported, while research into human language learning relevant to robotics yielded useful results.

The purpose of this paper is to present different methods that complement and influence one another, despite not being fully integrated at this stage. At present, progress is being made by pursuing individual methods and introducing novel ideas, which all contribute to a common goal: to advance language learning in robots. The various approaches described in this paper are underpinned by a common set of design principles, as explained below.

1.1 Design principles

Without claiming to model human language acquisition directly, our work is inspired by analogous human development, one aspect of which is the key role of social interaction in language learning. Thus, we conducted extensive experiments in human-robot interaction (HRI) and also investigated human-human interaction (HHI) in areas relevant to developmental robotics. Following the human analogy, we subscribe to the hypothesis that the integration of multiple learning paths promotes cognitive development and in particular, that co-development of action and language enable the enhancement of language capabilities, an area that has received little attention in the past.

The focus of the HRI experimental work in this study was the embodied humanoid robot iCub; see Figure 1. Research

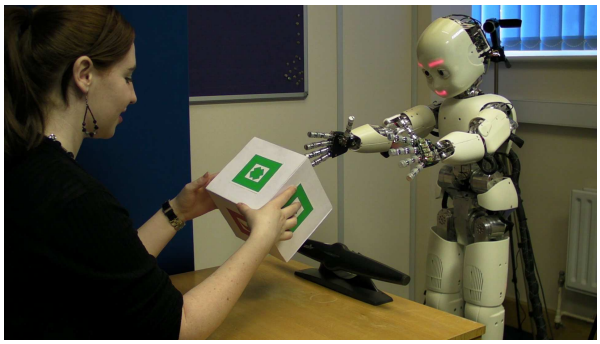

Figure 1. An experiment with the iCub robot. The participant is asked to teach the iCub words for shapes and colours on the box, speaking as if the robot were a small child. See sections 2.3, 2.4, and 3.1.

was also carried out on simulated robots and through computational modelling.

Since much of our work was inspired by child development, we investigated how robotic agents might handle objects and tools autonomously, how they might communicate with humans and how they might adapt to changing internal, environmental and social conditions. We also explored how parallel development, the integration of cognitive processes with sensorimotor experiences, behavioural learning and social interaction can promote language capabilities.

The concept of statistical learning underpins much of the work described here - that the frequency of an experience is related to learning associations, such as a speech sound and an object it names. The experiments described in the first two sections (see below) use various forms of statistical learning. Much of our work will feed in to wider concepts of statistical learning, where computational principles that operate in different modalities contribute to domain general mechanisms [2].

Thus, the following principles underpin the approach adopted in this project:

- Agents acquire skills through interaction with the physical environment, given the importance of embodiment, sensory-motor coordination and action oriented representation[1] - *physical interaction*.
- Agents acquire skills through interaction with humans in the social environment - *social interaction*.
- Learning is related to the frequency of associated experiences - *statistical learning*
- Behavioural, cognitive and linguistic skills develop together and affect each other - *co-development*.

Clearly these categories are interrelated and comprise many common challenges. For example, the concept of *symbol grounding*, where the meaning of language is grounded in sensing and experiencing the world, is

1 By "representation" we refer broadly to particular informational correlations between physical, social, linguistic or internal and sensorimotor processes.

fundamental throughout [3, 4]. A constructivist view of language underpins the work of this project [5]. Similarly, the concept of *time* and the physical experience of time is crucial both to sequential actions and to aspects of language learning, such as the order of words and the understanding of linguistic construction.

Our research is influenced, either explicitly or implicitly, by enactive and sensorimotor theories of perception and cognition [6, 7, 8, 9]. We have developed the hypothesis that embodied active perception in different modalities can be integrated to simulate human cognition and assume that language learners experience multiple modalities. However, some initial experiments, such as those described in section 3, were conducted using single mode input, prior to the development of multimodal methods.

Note that we diverge from some earlier sensorimotor principles in terms of combining initial low level analysis with the structure of internal models (for example, see section 3.3).

1.2 Structure of the report

This report is divided into three divisions, as shown in Table 1. The first section reports on work with multimodal language learning methodologies. The common thread is that in this work, different sensory streams are integrated and dependencies between action and language are modelled. However, diverse approaches are applied. Section 2.1 bases experiments with iCub on a computational model that integrates language and action with recurrent neural nets. Section 2.2 reports on the development of an architecture that models multiple cognitive and behavioural phenomena, providing structured Hebbian associations between self-organizing maps. Section 2.3 describes work based on human-robot interaction (HRI),e.g., multimodal perceptions of iCub being integrated to enable the learning of word meanings. Section 2.4 reports novel HRI experiments with iCub involving the learning of linguistic negation by integrating a minimal motivational system with sensorimotor perceptions.

The second main division in this report covers work with iCub pertaining to separate components of language learning that remain to be integrated with other approaches. Section 3.1 reports on HRI experiments that indicate how the transition from babbling to word form learning can occur, i.e., the preliminary processing of an unsegmented stream of sounds. Section 3.2 describes HRI experiments, loosely based on Steel's language game paradigm [10, 11], showing how word meanings might be learned. The third section, 3.3, addresses the issue of generating goal directed movements in robots. Using the passive motor paradigm (PMP), traditional problems with multijoint co-ordination are avoided, as experiments with iCub show. This work can play a critical role in this area of research, as the integration of action with language requires the practical implementation of goal directed movements.

The final main division in this report relates to work on human-human interaction (HHI), as well as human-robot interaction (HRI), which is relevant to developmental robotics and influences approaches to research into language acquisition by robots. Key areas are related to understanding how humans learn and enact linguistic meaning, as well as the dynamics of social interaction. The work presented here informed experiments, as described in sections 2 and 3. For example, investigations into the use of child-directed speech has a role in establishing contingent social interactions.

The sections in each part describe in detail the methods used in this research. Each approach is described under three headings: *Introduction, Experimental work* and *Outlook*. We introduce the method, providing some research background and describe the experimental work that was carried out; we also explain the techniques involved, noting advantages and disadvantages, and then conclude with a future outlook. Some of the results of the work described here can be found in [12], as well as in individual reports cited below.

2. Embodied Language Learning Methodologies Integrated with Action

Reports in this section describe work in which perceptions through multimodal sensory channels - audio, visual, tactile, proprioceptive, as well as simulated keyboard input - lead to language learning, The four methods described in this section have been investigated independently, each illuminating different aspects of language acquisition. First, we describe work showing how time-sensitive neural networks can be used to represent the integration of action and language. The second subsection describes the epigenetic robotics architecture (ERA), which enables the practical integration of sensory and motor data. The third subsection reports on experiments in which the robot learns the meaning of words from speech and visual input. Finally, using a similar scenario, a method for acquiring negation words is reported, a novel research area.

2.1 Integrating language and action with time-sensitive recurrent neural nets

Introduction: During early phases of development, the acquisition of language is strongly influenced by the development of action skills and vice versa. Dealing with the complex interactions between language and actions, as has been observed in language comprehension [13, 14] and acquisition [15, 16, 17, 18], requires the identification of computational means that are capable of representing time. The ability to deal with temporal sequences is a central feature of language and indeed, of any cognitive system.

Therefore, we opted for artificial neural networks for the investigation of grammatical aspects in language and in particular, for the capability of those systems to autonomously capture grammatical rules from examples [19, 20,

Embodied language learning methodologies integrated with action

Section	Research area	Perceptual modes	Social interaction	Work with iCub
2.1	Integrating language and action with time-sensitive recurrent neural nets	speech, vision tactile	yes	yes, also with models
2.2	ERA - epigenetic robotics architecture - SOM* neural nets combining sensory and motor data	speech, vision proprioception	yes	yes
2.3	Meaningful use of words and compositional forms	prosody and transcribed speech, vision, proprioception	yes, naïve participants	yes
2.4	Acquisition of linguistic negation in embodied interaction	speech, vision affect/ motivation	yes, naïve participants	yes

Embodied language learning methodologies developed separately

Section	Research area	Perceptual modes	Social interaction	Work with iCub
3.1	Transition from babbling to word forms in real-time learning	speech (vision)	yes, naïve participants	yes
3.2	Language game paradigm and social learning of word meanings	touch screen input	yes	yes
3.3	Passive motion paradigm (PMP) - to generate goal-directed movements in robots	simulated perceptions integrated with action	no	yes

Investigations into social interaction through HHI* and HRI*

Section	Research area	Perceptual modes	Social interaction	Work with iCub
4.1	HHI* and HRI* mediated by motor resonance	speech, vision with action	yes, naïve participants	yes
4.2	Co-development and interaction in tutoring scenarios. HHI and HRI	speech, vision with action	yes, naïve participants	partial
4.3	Analysing user expectations. HHI and HRI	speech, vision	yes, naïve participants	partial
4.4	Linguistic corpora studies to investigate child language acquisition. HHI	orthographic transcripts	yes, naïve participants	no

*SOM: Self-organizing map. HHI: human-human interaction. HRI: human-robot interaction.

See text for references

Table 1. Structure of the report

21]. More recently, several connectionist models have approached the problem of language acquisition and in particular, the co-acquisition of elements of syntax and semantics, by implementing artificial systems that acquire language through the direct behavioural experience of artificial agents [22, 23, 24, 25]. This approach has the specific aim of responding to the criticism of the symbol grounding problem [3, 4] on the one hand, which is one of the major challenges for symbolic AI-based systems and on the other, to exploit the autonomous learning capabilities of neural networks, both in terms of behaviours and elements of syntax.

The work described here was influenced by pioneering studies conducted by Jun Tani and collaborators [24, 26, 27] who investigated how a neuro-robot can co-develop action and language comprehension skills.

Experimental work: In the models cited above, the representation of time was achieved via the internal organization of specific types of neural networks, namely, recurrent neural networks (RNN), which can learn and recall temporal sequences of inputs and have been shown to be reliable models of short-term memory circuitry (see [28]). In addition to the typical implementation of RNNs, in which certain nodes show re-entrant connections, that is, they are connected to themselves, different variations have been proposed. An interesting variation is the multiple time-scales RNN [26, 27]. The MTRNN core is based on a continuous time recurrent neural network [29] that is characterized by the ability to preserve its internal state and hence, exhibit complex temporal dynamics. The neural activities of MTRNN are calculated following the classic firing rate model, where each neuron's activity is given by the average firing rate of the connected neurons. In addition, the MTRNN model implements a leaky integrator and therefore, the state of every neuron is not only defined by the current synaptic inputs, but also considers its previous activations.

Neural networks are often trained using a variation of the back-propagation methods. In particular, RNN, as well as

MTRNN, are trained using the back propagation through time algorithm (BPTT), which is typically used to train neural networks with recurrent nodes. This algorithm allows a neural network to learn the dynamic sequences of input-output patterns as they develop over time. See [30].

The MTRNN and RNN methods above were applied in experiments with the iCub robot to investigate whether the robot could develop comprehension skills analogous to those developed by children during the very early phase of their language development. More specifically, we trained the robot using a trial and error process to concurrently develop and display a set of behavioural skills, as well as an ability to associate phrases such as "reach the green object" or "move the blue object" to the corresponding actions (see Figure 2). A caretaker provided positive or negative feedback about whether the robot achieved the intended results.

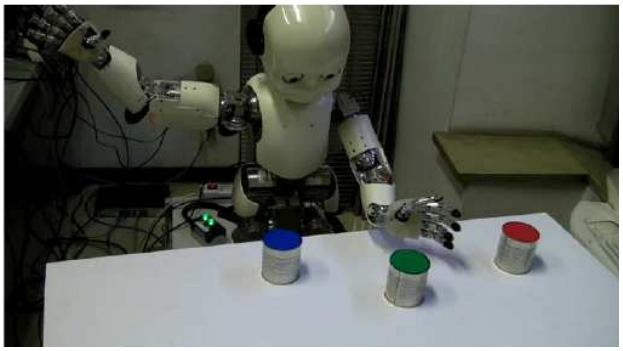

Figure 2. The set up of iCub for experiments is described in Section 2.1. The robot was trained via a trial-and-error process to respond to sentences such as "reach the green object". It then becomes able to generalize new, previously unheard sentences with new behaviours.

This method allowed the perceived sentences and the sensors encoding other (visual, tactile and proprioceptive) information to influence the robot actuators without first being transformed into an intermediate representation: that is, a representation of the meaning of the sentence. This method enabled us to study how a robot can generalize at a behavioural level – how it can respond to new, never experienced utterances with new and appropriate behaviours. At the same time, we also studied how it can "comprehend" new sentences by recombining the "meaning" of constituent words in a compositional manner to produce new utterances. Similarly, we studied how the robot can produce new actions by recombining elementary behaviours in a compositional manner [31, 27].

The BPTT of medium- to large-scale MTRNNs is computationally expensive, as the algorithm relies heavily on large matrix-vector multiplications. State-of-the-art CPU-based algorithms require a prohibitively large amount of time to train and run the network, prohibiting the real-time applications of MTRNNs. To optimize this, we instead relied on graphical processing unit (GPU) computing to speed up the training of the MTRNNs [32].

Outlook: Our approach provides an account of how linguistic information can be grounded in sub-symbolic sensory-motor states, how conceptual information is formed and initially structured and how agents can acquire compositional behaviour and display generalization capabilities. This in turn leads to the emergence of compositional organization that enables the robot to react appropriately to new utterances never experienced previously, without explicit training.

2.2 Epigenetic robotics architecture (ERA) - combining sensory and motor data

Introduction: The epigenetic robotics architecture (ERA) was developed to directly address issues of ongoing development, concept formation, transparency, scalability and the integration of a wide range of cognitive phenomena [33]. The architecture provides a structure for a model that can learn, from ongoing experience, abstract representations that combine and interact to produce and account for multiple cognitive and behavioural phenomena. It has its roots in early connectionist work on spreading activation and interactive activation and competition models. In its simplest form, ERA provides structured Hebbian associations between multiple self-organizing maps in such a manner that spreading and competing activity between and within these maps provide an analogue of priming and basic schemata.

Once embodied and connected to both sensory and motor data streams, the model has the ability to predict the sensory consequences of actions thereby providing implementation of theories pertaining to sensorimotor perception [34, 7].

Experimental work: ERA provides for structured association between multiple self-organizing maps via special "hub" maps; several "hubs" then interact via a "hub" map at the next level and so on. Here, the structure of the architecture emerges as a consequence of the statistics of the input/output signals. Activity flows up the architecture, driven by sensor and motor activity, and back down the architecture via associations to prime or predict the activity at the surface layer. See Figure 3, as well as figures in [35].

Scalability is addressed in several ways; firstly, by constructing hierarchies, large inputs can be accommodated and the gradual integration of information in ever higher regions of the hierarchy provides an analogue of abstraction. Secondly, while the model is fundamentally an associative priming model, it is able to produce analogies to a wide variety of psychological phenomena. Thirdly, the homogeneous treatment of different modalities – whether sensor- or motor-based – provides a method that can easily accommodate new and additional modalities without requiring specialized pre-processing, though we do acknowledge that appropriate pre-processing may be beneficial. Finally, in relation to sensorimotor theories, the gap between sensorimotor prediction and an interaction-based account of affordances is significantly narrowed [36].

The ERA architecture in its simplest form was successfully applied to modelling bodily biases in children's word

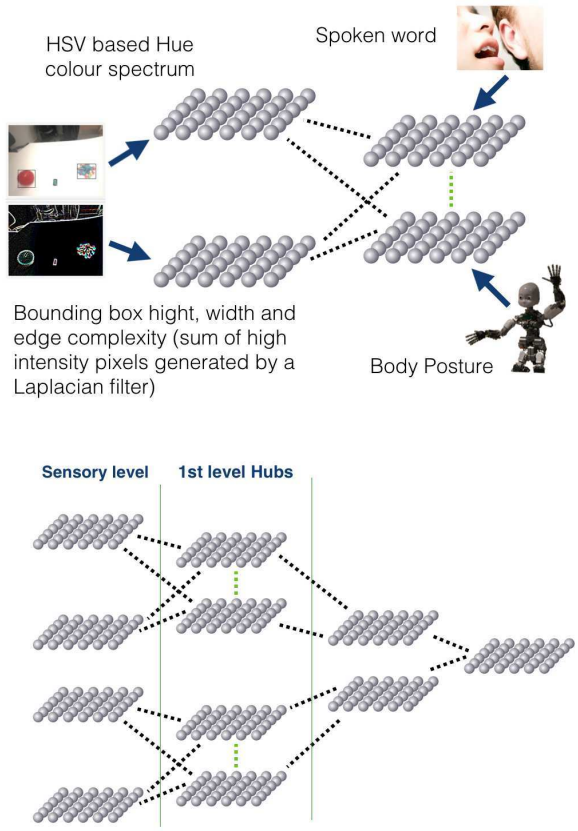

Figure 3. Top panel: The ERA model in its simplest form as structured mapping between self-organizing maps driven by sensory input. Bottom panel: the extended ERA model in which a hierarchy of self-organizing maps are driven at the sensory level by sensory input, then at the hub level by the positions of winning nodes in the connected maps at the previous layer. See section 2.2.

learning [37], the effect of grouping objects upon learning a common feature and the transformative effect of labelling and spatial arrangement on the computational or cognitive complexity of tasks [38]. Additionally, an extended version of the architecture utilizing active Hebbian links to directly influence the learning within each self-organizing map was explored in relation to modelling the "switch" task and more generally, the so called "U-shaped performance curves" during development [39, 40].

Outlook: While ERA is fundamentally an associative priming model, it is able to produce a wide variety of psychological phenomena, which have been validated against both existing child data and additional child experiments, confirming predictions of the model (see also the Conclusion to this paper regarding "research loops"). Beyond the integration of cognitive phenomena, ERA also provides a fulcrum for the technical integration of many of the modelling outputs of the project by developing structures based on simple relationships between inputs, outputs and anything else provided. The architecture can learn, from ongoing experience, abstract representations that combine and interact to produce and account for multiple cognitive and behavioural phenomena. Further work on the cognitive development of humanoid robots based on the ERA architectural system is under way [35], incorporating elements of the method described in section 3.2.

In its current form, the ERA modelling approach has a number of limitations, including problems learning sequential information and producing complex dynamic and adaptive behaviour. While dynamic behaviour can and has been generated from the model, it is motor-focused and therefore, not particularly useful for learning action affordances. Combined with pre-wired action production systems, action words and basic affordances can be learned; however, this is unsatisfactory and more plausible methods for action production are needed.

2.3 Integrating multimodal perceptions for the meaningful use of words and compositional forms

Introduction: In this section we focus on the methods employed for grounding lexical concepts in a robot's sensorimotor activity via human robot interaction [41]. Language learning is a social and interactive process, as emphasized by Tomasello [5], Kuhl [42] and Bloom [43]. The methods described here concern language learning from embodied interaction, how this is affected by feedback from the robot and how this affects the robot's learning experience. The work presented in sections 4.2 and 4.3 on tutoring situations and user expectations influenced this approach.

In this and other work (see sections 2.4 and 3.1), the human speech tutors to the robot were naïve participants, paid a token amount as a gesture of appreciation. Most of them were administrative staff from the university or students from other disciplines. They were asked to speak to the robot as if it were a small child. Note that the robot learned separately from each participant over multiple sessions so that in effect, learning occurred as if each participant had their own robot that learned only from them.

Experimental work: The methodologies employed are broken down into three parts: firstly, extracting relevant salient words from the human tutor's speech, based on research with human children and aspects of child directed speech (CDS); secondly, the learning mechanisms linking salient human speech with the robot's own perceptions, enabling it to produce similar speech during similar sensorimotor experiences; thirdly, attempting to achieve rudimentary compositionality, exhibited in simple two-word utterances made by the robot [44].

i. The first of the above methodologies focuses on extracting salient words from the human tutor's speech. This was achieved by considering what a human infant hears in a social situation with a caregiver. Typically, utterances are short, often less than five words and with many utterances consisting of a single word. Repetition is common. The caregiver tends to talk slower than would typically be expected of an adult. Most words are mono- or disyllabic. Salient

words are lengthened and prosody is used to give greater emphasis to such words. Salient words are often placed at the end of utterances directed at young infants. Initially, there are generally more nouns than other types of words (see also section 3.1.)

Two primary methods were used for extracting salient words: firstly, prosodic measures combining energy (volume), pitch and duration (the length of an uttered word); secondly, splitting utterances into two sections focusing on the high salience final word and pre-final words. Both of these techniques reflect aspects of CDS, mentioned above.

ii. The second methodology is the learning mechanism itself. In this context, we considered that the meaning of a communicatively successful utterance was grounded in its usage, based on the robot's sensorimotor history: auditory (prosody and transcribed speech), visual and proprioceptive – derived from acting and interacting in the world. These grounded meanings can then be scaffolded via regularities in the recognized word/sensorimotor stream of the robot. The first step in this process is to merge the speech stream of the human, represented as a sequence of salient words, with the robot's sensorimotor stream. This is achieved by matching the two modalities based on time and thus linking salient words with the robot's sensorimotor inputs at the time the word is "heard" by the robot. The word can later be expressed through the robot's speech synthesizer.

To achieve such associations, we faced a number of challenges. The first of these concerns associating what was said to the appropriate parts of the sensorimotor stream. Thus, the human tutor may show the robot a shape (e.g., the "sun"), but only say the word *sun* within the utterance before or after the shape has appeared/disappeared from the view of the robot (e.g., "here's a sun" and then show the sun shape, or say "that was a sun" after having shown the sun shape). Secondly, which set of sensorimotor attributes are involved and at what points in time are such attributes relevant to the speech act? We made no pre-programmed choices as to what was relevant for the robot. However, in order to manage these issues, we applied two heuristics. The first coped with the association of events by remapping each salient word uttered by the human tutor onto each element of the temporally extended sensorimotor stream of the utterance containing it. In effect, this made the chosen word potentially relevant to the robot's entire sensorimotor experience during that utterance and therefore relevant to any sensorimotor inputs that arose during this time. The second heuristic used mutual information to weigh the appropriate sensorimotor dimensions relevant to the classification of that word (effectively using an "information index" [45]). The associator mechanism employed was *k*-nearest neighbour (*k*NN). The robot may then later *utter* such a salient word when it re-experiences a sensorimotor context similar to the one it associates the word with.

iii. Thirdly, we investigated the robot's acquisition and production of two-word (or longer) utterances, the lexical component items of which had been learned through experience. This was again based on experiments and analysis dealing with the acquisition of lexical meaning, in which prosodic analysis and the extraction of salient words are associated with a robot's sensorimotor perceptions as an attempt to ground these words in the robot's own embodied sensorimotor experience. An in-depth analysis of the relationship between the characteristics of the robot tutor's speech and the robot's sensorimotor perceptions was conducted.

Following the extraction of salient words, we investigated the learning of word order. Two *k*NN memory files were employed to capture the combination of salient words occurring in an utterance. The first held all salient words uttered before the final salient word in the utterance. The second held the final salient word in the utterance. Note that these were *salient* words; thus, the final salient word may not necessarily be the final actual word in an utterance. The robot matched these memory files against its current sensorimotor perceptions and in this way tried to find the most similar experience (if any) when it "heard" a word previously compared to what it now experienced. This had the effect of making the robot utter words that reflected both what it was taught (about objects and colours) and the order in which the words originally occurred. That is, by successively uttering any of the best matching words for each of the two memory files, upon seeing a new coloured shape, even if in a novel combination, the robot should express the correct attribute within a proto-grammatic compositional form, thereby reflecting usage by the human it learned from, possibly as part of a completely novel utterance.

Outlook: The approaches outlined above have advantages and drawbacks. A positive factor is that the human tutor is able to use natural unconstrained speech. However, speech topics are limited to the simple environment of the robot talking about blocks, shapes and colours, and is therefore naturally constrained. In terms of prosodic salience the mapping of sensory embodiment to words automatically allows the robot to associate simple lexical meaning with them, based on its own perceptions. However, the assignment of salient words within the temporal utterance in which they occurred may have competing solutions.

One problem with the method outlined above is the non-real-time nature of the association of words and sensorimotor experiences. In current implementations, a limiting factor has been the inability to apply phonetic or phonemic word recognition in real-time without extensive training.

Extensions to these methods include further analysis of the prosodic nature of the interaction and investigations into how the robot might use prosodic clues to support the capacity for learning to use words meaningfully beyond its

mere sensorimotor associations attached to particular words. More specifically, we might ask, how well can the robot attach an attribute to a word, and distinguish between a set of attributes such as "colour" and a member of that set such as "red"? This distinction would be a step towards deriving linguistic constructions, combined with perceived word order or inflectional markings. This method could contribute to grammar induction as a way of forming templates for word types or thematic constructions and their appropriate contexts of use (i.e., meanings in a Wittgensteinian sense of *language games*).

2.4 Acquisition of linguistic negation in embodied interactions

Introduction: Linguistic negation is a fundamental phenomenon of human language and a simple "no" is often one of the first words to be uttered by English-speaking children. Research concerned with the ontogeny of human linguistic negation indicates that the first uses of a rejective "no" are linked to affect [46]. We therefore introduced a minimal motivational model into our cognitive architecture as proposed by Förster et al. [47] in order to support the grounding of early types of negation. We employed methods as in the acquisition of lexical usage work in an interactive scenario (like those discussed in section 2.3 and in Saunders et al. [41]) to support the enactive acquisition and use of single lexical items.

Experimental work: The purpose of the experimental work was to investigate a robot's capacity to learn to use negative utterances in an appropriate manner. The resulting architecture, designed to elicit the linguistic interpretation of robot behaviour from naïve participants, was used in human-robot interaction (HRI) studies with iCub. This architecture consists of the following parts:

- **i.** A perceptual system that provides the other parts of the behavioural architecture with high-level percepts of particular objects and human faces (loosely based on the saliency filters described by Ruesch et al. [48]).
- **ii.** A minimal motivational system that can be triggered by other sub-systems.
- **iii.** A behavioural system that controls the robot's physical behaviour, based on both the output of the perceptual system and the motivational system.
- **iv.** A speech extraction system that extracts words from a recorded dialogue and which operates offline.
- **v.** Sensorimotor-motivational data originating from the systems described above and recorded during an interaction session are subsequently associated with the extracted words, using the same heuristics as described in section 2.3, which discusses learning to use words in iterated language games with naïve participants.
- **vi.** A language system that receives inputs from the systems outlined above. Subsequently, it maps perceptions, motivation and behavioural states to an embodied dictionary provided by the speech extraction system. Mapping is performed using a memory-based learning algorithm [49] similar to the one described in section 2.3. This system controls the speech actions of iCub.

We employed what we refer to as an active vocabulary: in order to enrich the dialogue and anticipating ties in the mapping algorithm, two consecutively-uttered words were enforced to be different from each other in the very same experiential situation, i.e., when the sensorimotor-motivational data are exactly the same. This was achieved by enforcing that a subsequent potentially uttered word associated to the experience would be different.

We constructively investigated the hypothesis that rejective negation is linked to motivation, rather than simply to perceptual entities. Affective response to objects is valenced as positive, neutral or negative and can therefore shape motivation and volition for actions in response to the said actions. This important psychological insight appears also in the related enactive model of the embodied mind as detailed in Varela et al. [6, Ch. 6]. The cognitive architecture used here is the first to implement this principle on a humanoid, albeit in a simple way; this serves as an essential element in grounding language learning by the robot in a way that expands beyond mere sensorimotor associations, by including "feeling", i.e, a valenced stance towards objects. The constructed motivational system leads to the avoidance of certain objects, i.e., (non-verbal) rejection of these objects via facial expressions and matching body language, or the opposite for objects towards which valence is positive.

However, the described architecture was also constructed for a second purpose: to support or weaken the hypothesis that the very root of negation lies in the prohibitive action of parents. In language, rejective negation is used when one rejects an object or action, while prohibitive negation is used to prohibit the action of someone else. It may be that exposure to prohibitive negation promotes the development of negation in children. To this purpose, we performed an HRI study that compares the performance of systems learning in a combined prohibitive plus rejective scenario against a purely rejective negation scenario. In the rejective scenario, participants were asked to teach humanoids different shapes printed on small boxes that are placed in front of them. They are told that the humanoid has different preferences for these objects: it may like, dislike or be neutral about them. In the prohibitive scenario, participants are told to teach the robot the names of the shapes, but also that some of the shapes are not allowed to be touched. Participants are in neither instance aware of the true purpose of the experiment: to investigate the robot's acquisition of the capacity to use negative utterances in an appropriate manner.

Outlook: The system described here is the first grounded language learning system to include motivational aspects

in addition to sensorimotor data in language grounding. In developmental trajectories with different naïve participants, the humanoid is able to acquire in only a few sessions the capability for using negation. Its speech and behaviour appears to humans to express an array of negative functions or types similar to the ones engaged in by infants [8, 50]. The elicitation of linguistic negation in the interactions of a humanoid with naïve participants and the comparative efficacy of negation acquisition with and without prohibition can help assess the notion that internal states, such as affect and motivation, can be as important as sensorimotor experience in the development of language; for detailed results to date, see [50].

3. Embodied Language Learning Methodolgies Developed Separately

In the previous primary section of this paper, we reported experimental work in which action and language were connected. In this section, we describe three pieces of work relevant to language learning robots, which have been developed separately. The first subsection implements a method for processing an unsegmented stream of speech sounds in order to produce word forms. It shows how prior word form learning precedes the learning of meaning and how written transcripts differ from the actual audio stream. Using a language game paradigm, the second subsection describes how word meanings are learned through social interaction. Thirdly, we report on methods for producing goal-directed movements in robots using the passive motion paradigm (PMP), which replaces computationally intractable earlier methods. As in the previous section, these research experiments were carried out independently and in parallel.

3.1 The transition from babbling to word forms in real-time learning

Introduction: The experiments described here have the initial purpose of showing how an unsegmented audio stream might be processed and thereby model the transition from babbling to salient word form acquisition. This is approached through real time proto-conversations between human participants and an iCub robot. The work on human-human interaction described in section 4 influenced the experimental design, especially the need for *contingent* social interaction.

The processes implemented here are partially analogous to some of the developments in human infants aged from 6-14 months. For additional details see Lyon (2012) [51]. The scenario is shown in Figure 1.

The learning of word forms is a prerequisite to learning word meanings [52]. Before a child can begin to understand the meanings of words, he or she must be able to represent word forms, which then come to be associated with particular objects or events [53]. The acquisition of word forms also facilitates the segmentation of an acoustic stream: learned word forms act as anchor points, dividing the stream of sounds into segments and thereby supporting segmentation by various other routes.

There is a close connection between the perception and production of speech sounds in human infants [54, 55]. Children practice what they hear; there is an auditory-articulatory loop and children deaf from birth, although they can understand signed and written language, cannot learn to talk. An underlying assumption is that the robot, like a human infant, is sensitive to the statistical distribution of sounds, as demonstrated by Saffran [56] and other subsequent researchers.

Most of the salient words in our scenario were in practice single syllable words (*red, green, black* etc. *box, square, ring* etc.). The more frequent syllables produced by the teacher were often salient word forms and iCub's productions were influenced by what it heard. When iCub produced a proper word form, the teacher was asked to make a favourable comment, which acted as reinforcement.

Experimental work: A critical component of early human language learning is contingent interaction with carers [57, 42, 58, 59, 60]. Therefore, we conducted experiments in which human participants, using their own spontaneous speech, interacted with an iCub robot with the aim of teaching it word forms.

The human tutors comprised 34 naïve participants who were asked to speak to the robot as if it were a small child. After the experiment, they answered a short questionnaire pertaining to their attitude toward the iCub. Most had the impression that iCub acted independently. On a scale of 1-5, where 1 represented dependent, and 5 fully independent, 16 out of 19 respondents gave a score of 4 or 5.

The following assumptions about iCub's capabilities were made:

i. It practices turn-taking in a proto-conversation.

ii. It can perceive phonemes in a manner analogous to human infants.

iii. It is sensitive to the statistical distribution of phonemes in a manner analogous to human infants [56, 61].

iv. It can produce syllabic babble, but without the articulatory constraints of human infants; thus, unlike humans, it can produce consonant clusters.

v. It has the intention to communicate and therefore reacts positively to reinforcement, such as comments of approval.

The scenario for the experiments (Figure 1) sees the teacher sitting at a table opposite iCub, which can change its facial expression and move its hands and arms. The robot's lower body is immobile. There is a set of blocks and the participant is asked to teach iCub the names of the shapes and colours on the sides of the blocks. Initially, iCub

produces random syllabic babble, but this changes to quasi-random syllabic babble that is biased towards speech heard from the teacher. When the teacher hears a proper word form, they are asked to reinforce this with an approving comment.

The teacher's speech is represented as a stream of phonemes. As no assumption is made about how this phonemic stream might be segmented into words or syllables, iCub perceives the phonemic input as a set of all possible syllables. For example, using letters as pseudo-phonemes, the string *i s a b o x* generates *i is sa sab a ab bo box o ox*. A frequency table for each of these syllables is incremented in iCub's language processor as they are perceived.

Influenced by what it has heard, iCub's initial random syllabic babble becomes biased towards the speech of the teacher.

Each participant had 2*4 minute proto-conversations with iCub. For the conversion of the teacher's speech to a string of phonemes, an adapted version of the SAPI 5.4 speech recognizer was used. The iCub's output was converted using the eSpeak speech synthesizer. The CMU phonemic alphabet was used [62, 63].

Since our participants were asked to talk to iCub as if it were a small child, the user's expectation was influenced in advance. Participants used their own spontaneous words and we observed child-directed speech being extensively used, particularly by individuals that had experience caring for human infants. A wide range of interactive styles was observed: some participants were very talkative, while others said very little.

A video clip that provides an example of a "conversation" can be viewed at http://youtu.be/eLQnTrX0hDM (note that '0' is zero).

Outlook: The results indicate that phonetic learning, based on a sensitivity to the frequency of sounds occurring, can contribute to the emergence of salient words. This result also supports other methods, for example, through prosody and actions, as described in section 2.3 and in Saunders (2011) [64].

To understand why this method works, we need to distinguish between speech sounds and the orthographic transcripts of words, between which there is not a 1-to-1 correspondence. Orthographic transcripts of speech do not represent exactly what the listener actually hears. Salient content words (nouns, verbs, adjectives) are more likely to have a consistent canonical phonemic representation than function words, where variation in prosody and pronunciation is often pronounced. For example, in four hours of spontaneous speech annotated phonemically, Greenberg reported that the word "and" had been recorded in 80 different forms [65]. A consequence of this is that, as perceived phonemically, the frequency of function words is less than their frequency in orthographic transcripts. In contrast, the frequency of salient content words accumulates and so does their influence on the learner.

Our current approach accords with recent neuroscientific research showing that dual streams contribute to speech processing [66, 67]. The experiments described here investigate dorsal stream factors by modelling the transition from babbling to speech.

Future work should investigate other methods of representing speech sounds, as well as, or instead of, phonemes. Advances have been made in using articulatory features such as place and manner of articulation and voicing; their acoustic manifestations can be captured as a basis for the representation of speech. See for example [68, page 294].

3.2 The language game paradigm and the social learning of word meanings

Introduction: Social learning relies on the interplay between learning strategies, social interaction and the willingness of a tutor and learner to engage in a learning exchange. We studied how social learning can be used by a robot to acquire the meaning of words [35]

Experimental work: We implemented a social learning algorithm based on the language game paradigm of Steels [10, 11] (a concept resonant with Wittgenstein's language games). The algorithm differed from classic machine learning approaches in that it allowed for relatively unstructured data and actively solicited appropriate learning data from a human teacher. As an example of the latter, when the agent noticed a novel stimulus in the environment, it would enquire from the human the name of that stimulus. Alternatively, when its internal knowledge model was ambiguous, it would ask for clarification. The algorithm, after validation via simulation [69], was integrated in a humanoid robot that displayed appropriate social cues for engaging with the human teacher (see Figure 4). The robot was placed opposite a human subject, with a touch screen between the robot and the human to display visual stimuli and to allow the human to provide input to the robot, thereby avoiding the need for speech recognition and visual perception in the robot, which may have introduced noise in the experiment.

In the experiment, two conditions were used – one in which the robot used social learning and respective social cues to learn (social condition) and another in which the robot did not provide social cues (non-social condition). The social condition resulted in both faster and better learning by the robot, which – given the fact that the robot has access to more learning data in the social condition through the additional feedback given by the human tutor – is perhaps not surprising. However, we did notice that people formed a "mental model" of the robot's learning and tailored their tutoring behaviour to the needs of the robot. We also noticed a clear gender effect, where female tutors were markedly more responsive to the robot's social bids than male tutors [35].

Figure 4. Setup for social learning of word-meaning pairs by a humanoid robot. See section 3.2.

Outlook: These experiments showed how the design of the learning algorithm and the social behaviour of the robot can be leveraged to enhance the learning performance of embodied robots when interacting with people. Further work is demonstrating how additional social cues can result in tutors offering better quality teaching to artificial agents, leading to improved learning performance. These experiments have been incorporated into an ERA architecture, as described in section 2.2

3.3 The passive motor paradigm (PMP): generating goal directed movements in robots

Introduction: This section addresses robotic movements that are essential to research about the integration of action and language learning.

A movement on its own has no connection with language unless it is associated with a goal; this usually requires the recruitment of a number of motor variables (or degrees of freedom) in the context of an action. Even the simple task of trying to reach point B in space, starting from a point A, in a given time T can in principle be carried out in an indefinitely large number of ways, with regards to spatial aspects (hand path), timing aspects (speed profile of the hand) and the recruitment patterns of the available joints in the body (final posture achieved). How does the brain choose one pattern from numerous other possible ones? Recognizing the crucial importance of multi-joint coordination was a true paradigm shift away from the classic Sherringtonian viewpoint [70] (typically focused on single-joint movements) and toward the Bernsteinian [71] quest for principles of coordination or synergy formation. Since then, the process by which the central nervous system (CNS) coordinates the action of a high-dimensional (redundant) set of motor variables for carrying out the tasks of everyday life – the "degrees of freedom problem" – has been recognized as a central issue in the scientific study of the neural control of movement. Techniques that quantify task goals as cost functions and use sophisticated formal tools of optimization have recently emerged as a leading approach for solving this ill-posed action generation problem. [72, 73].

However, questions arise regarding the massive amount of computations that need to be performed to compute an optimal solution. We need to know how distributed neural networks in the brain implement these formal methods, how cost functions can be identified/formulated in contexts that cannot be specified *a priori*, how we can learn to generate optimal motor actions, as well as the related issue of sub-optimality. All of these topics are still widely debated [74, 75]. Recent extensions [76] provide novel insights into issues related to a reduction in computational cost and learning.

An alternative theory of synergy formation is the passive motion paradigm (PMP) [77, 78, 79], an extension of the equilibrium point hypothesis (EPH) [80, 81, 82] and based on the theory of impedance control [83]. In PMP, the focus of attention shifts from "cost functions" to "force fields". In general, the hypothesis here is that the "force field" metaphor is closer to the biomechanics and the cybernetics of action than the "cost function" metaphor. Our aim was to capture the variability and adaptability of human movement in a continuously changing environment in a way that was computationally "inexpensive", allowing for compositionality and run-time exploitation of redundancy in a task specific fashion, together with fast learning and robustness.

Experimental work: The hypothesis was investigated by implementing the model on the iCub and conducting a number of experiments related to upper body coordination and motor skills learning [78]. The basic idea in PMP is that actions are the consequences of an internal simulation process that "animates" body schema with the attractor dynamics of force fields induced by the goal and task specific constraints. Instead of explicitly computing cost functions, in PMP, the controller has to simply switch on task relevant "force fields" and let the body schema evolve in the resulting attractor dynamics. The force fields, which define/feed the PMP network, can be modified at run time as a consequence of cognitively relevant events such as the success/failure of the current action/sub-action [84, 85]. Further experimental work has been carried out showing how the robot can learn about objects and perform actions on them, such as pushing a cube of a certain colour [79].

Outlook: An important property of PMP networks is that they operate only through well-posed computations. This feature makes PMP a computationally inexpensive technique for synergy formation. The property of always operating through well-posed computations further implies that PMP mechanisms do not suffer from the "curse of dimensionality" and can be scaled up to any number of degrees of freedom [78, 86]. In the framework of PMP, the issue of learning relates to learning the appropriate elastic (impedances), temporal (time base generator) and geometric (Jacobian) parameters related to a specific task. Some work

has been done in this area, for example, [87] deals with the learning of elastic and temporal parameters and [88] deals with the issue of learning geometric parameters. However, a general and systematic framework that applies to a wide range of scenarios remains an open question and work is ongoing in this area.

The local and distributed nature of computations in PMP ensures that the model can be implemented using neural networks [88, 89]. At the same time, the brain- basis of PMP is an issue that remains underexplored at present and requires more comprehensive investigation. A justification can still be made that highlights the central difference between EPH and PMP. In the classic view of EPH, the attractor dynamics that underlies production of movement is based on the elastic properties of the skeletal neuromuscular system and its ability to store/release mechanical energy [90]. Taking into account results from motor imagery [91, 92] PMP posits that cortical, subcortical and cerebellar circuits can also be characterized by similar attractor dynamics. This might explain the similarity of effects of real and imagined movements, because although in the latter case the attractor dynamics associated with the neuromuscular system is not operant, the dynamics due to the interaction among other brain areas are still at play. In other words, considering the mounting evidence from neuroscience in support of common neural substrates being activated during both real and imagined movements, we posit that real, overt actions are also the results of an "internal simulation" as in PMP. Even though results exist from behavioural studies [86], a more comprehensive programme for investigating the neurobiological basis of PMP may be needed to substantiate this viewpoint.

It remains open to question whether or not the motor system represents equilibrium trajectories [93]. Many motor adaptation studies demonstrate that equilibrium points or equilibrium trajectories per se are not sufficient to account for adaptive motor behaviour; however, these findings do not rule out the existence of neural mechanisms or internal models capable of generating equilibrium trajectories. Rather, as suggested by Karniel [93], such findings should induce the research to shift from the lower level analysis of reflex loops and muscle properties to the level of internal representations and the structure of internal models.

4. Investigations into social interaction through human-human interaction (HHI) and human-robot interaction (HRI)

The reports in this section focus on analysing social interactions, particularly between a teacher and a learner. The results from these experiments feed into work on robotic language learning, as described in sections 2 and 3, where iCub learns from a human teacher (on the other hand, a human may learn an action from a robotic demonstrator). The first subsection describes research into communication, possibly not intentional, through gaze, which is realized through motor resonance. The second subsection reports on experiments involving tutoring situations, based on adult-child scenarios that can be compared to human-robot interactions. The third subsection covers work on user expectations, showing how such expectations can affect the human teacher's approach to the robotic learner. The final subsection reports on work with corpora of recorded child language, while longitudinal experiments throw light on the process of language learning.

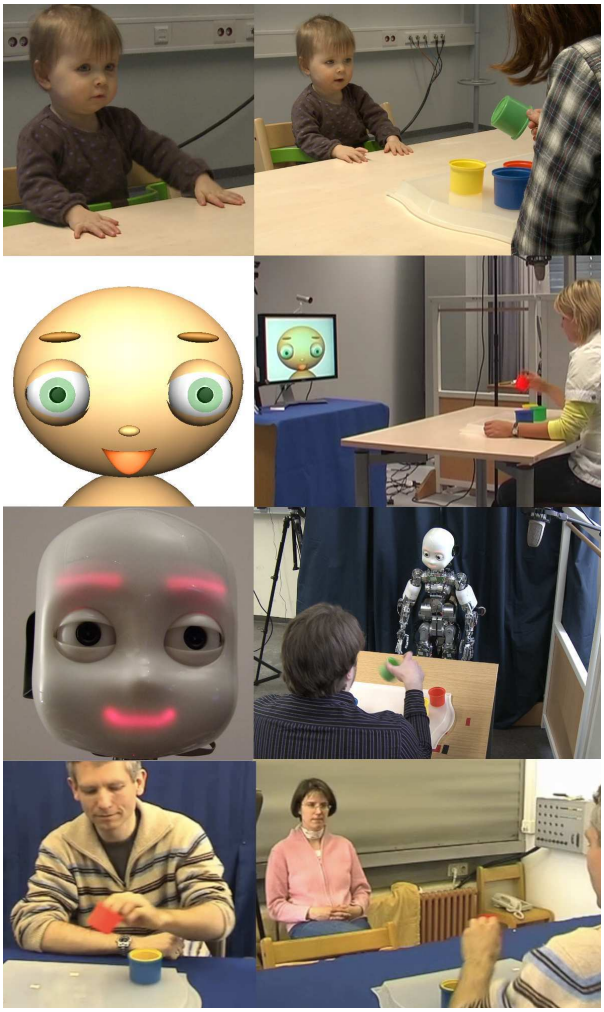

Figure 5. Experiments to gather interaction data. Participants (parents, adults) were asked to demonstrate actions such as stacking cups to a child (top level panels), a virtual robot on a screen (2nd level panels), the iCub robot (3rd level panels) or another adult (bottom panels). See section 4.2

4.1 Contingent human-human and human-robot interaction mediated by motor resonance

Introduction: A fundamental element of the integration of action and language learning is constituted by the way people perceive other individuals and react contingently to their actions. Indeed, beyond explicit and voluntary verbal exchanges, individuals also share beliefs and emotions in a more automatic way that may not always be mediated by conscious awareness. This is the case in communication

based on gaze motion, body posture and movements. The assessment of such implicit communicative cues and the study of the mechanisms at their core helps us understand human-human interaction and investigate how people perceive and relate to non-living agents in human-robot interaction.

Gaze behaviour contributes to language learning in sighted infants: appropriate gazing indicates referential intention when it comes to learning the names of objects and actions. Furthermore, it helps to create a rapport between teacher and learner, a characteristic explored in section 4.3 on user expectations. Gaze behaviour also plays a role in turn-taking in proto-conversations, a precursor to language learning.

The physiological mechanism at the basis of this implicit communication is known as motor resonance [94] and is defined as the activation of the observer's motor control system during action perception. Motor resonance is considered one of the crucial mechanisms of social interaction, as it can provide a description of the unconscious processes that induce humans to perceive another agent (either human or robot) as an interaction partner. The concept of motor resonance can be applied to investigate both human-human (HHI) and human-robot interaction (HRI), and the measure of the resonance evoked by a robotic device can provide quantitative descriptions of the naturalness of the interaction.

In particular, behavioural investigations can describe the tangible consequences of the tight coupling between action and perception described as motor resonance. By recording gaze movement and motion kinematics during or after action observation, we can directly individuate which features of the observed human or robot action are used by observers during action, understanding and execution. The modification of gaze or bodily movements associated with the observation of someone else's behaviour can indeed shed light on motor planning, indicating if and in what terms implicit communication has occurred. In particular, motor resonance can imply facilitation in the execution of an action similar to the one observed – motion priming – or a distortion while performing a different movement, i.e., motion interference. Other phenomena that reflect motor resonance and that could serve as an efficient measure of interaction naturalness are automatic imitation of sensory information into action and goal anticipation with gaze [95, 96]. For a review of the methodologies currently used for the study of motor resonance in HHI and HRI, see [97, 98].

There are alternative techniques available for measuring the naturalness of HRI: for example, neuroimaging and neurophysiological studies allow for the evaluation of the activation of the putative neural correlates of motor resonance (the mirror-neuron system) during action observation [99]. The limitations of these methods are that they are often quite invasive processes and do not permit the testing of natural interactions. Alternatively, standardized questionnaires have been proposed to measure users' perceptions of robots and to estimate factors involved in HRI. However, the questionnaires simply assess the conscious evaluations of the robotic devices and do not take into account some cognitive and physical aspects of HRI, thereby failing in terms of a complete HRI quantification. To circumvent this issue, physiological measurements such as galvanic skin conductance and muscle and ocular activities have been used to describe participants' responses when interacting with a mobile robot (e.g., [100]). We believe that a comprehensive description of the naturalness of the communication between humans and robots can only be provided through a combination of all the above mentioned techniques.

Experimental work: With the aim of studying action-meditated implicit communication and of evaluating how HRI evolves in a natural interactive context, we adopted two new behavioural measures of motor resonance: the monitoring of proactive gazing behaviour [96] and the measure of automatic imitation [101] (see [102] for a short review).

As the predictive nature of someone's gaze pattern is associated with motor resonance [95, 103], the quantification of this anticipatory, unconscious behaviour can represent a good estimate of the activation of the resonating mechanism and, in turn, of the naturalness of an interaction. This option presents some advantages with respect to the previously adopted methods, as it does not require subjects to perform predetermined movements, but to simply look naturally at an action. Moreover, it allows for the study of the effect of observing complex, goal directed actions. This differs from classic behavioural protocols, which generally require simple stereotyped movements. The method we employed was to replicate the experiments previously conducted in HHI studies, i.e., examine anticipatory gaze behaviour when subjects observed someone performing a goal directed action, such as transporting an object into a container [95]. This was done by replacing the human demonstrator with the robotic platform iCub. In this way, we could directly contrast the natural gaze pattern adopted during the observation of human and robot actions. A comparison between the timing of gazing (the number of predictive saccades) in the two conditions provided an indication of the degree of resonance evoked by the different actors. In particular, the appearance of the same anticipation in gaze behaviour during robot and human observation indicated that a humanoid robotic platform moving as a human actor can activate a motor resonance mechanism in the observer [96], thus implying its ability to induce pro-social behaviours [98, 104].

At the same time, studying the automatic imitation phenomena allowed us to quantitatively describe if and how human actions adapt in the presence of robotic agents, that is, if motor resonance mechanisms appear. This was done by studying the automatic imitation effect induced by

movement observation in movement production [105], whether the observed action was performed by a human agent or by the humanoid robot iCub [101]. The modification of the observer's movement velocity as a result of the changes in the human or robot actor's velocity is behavioural evidence of the occurrence of motor resonance phenomena.

Outlook: The behavioural methods proposed here present crucial advantages with respect to other methods of investigating action-mediated communication in HRI contexts. In particular, the evaluation of gazing and automatic imitation behaviours allows for spontaneity and smoothness in HRI, and for an ecological testing of natural interaction. However, they also present some drawbacks, including the impossibility of exactly determining the neural activation associated with interaction, which can be obtained by more invasive techniques like neurophysiological and neuroimaging investigations. Moreover, beyond the basic, unconscious reactions to the human and robot actions measured by these behavioural methods, several other cognitive processes might be involved during action observation and interaction that influenced robot perception, including attention, emotional states, previous experiences and cultural background. From this perspective, the methodologies we propose aim at covering the existing gap between the completely unconscious information obtained by neural correlates examination and the conscious evaluation of robotic agents provided by questionnaires. Our methodologies provide a quantitative description of human motor response during HRI, with a focus on contingent, action-based communication.

4.2 Co-development and interaction in tutoring scenarios

Introduction: In this section, we focus on methods that concern a parent's tutoring behaviour as directed towards a child, or similar human behaviour directed towards a robot (simulated on a screen or physically embodied). The scenarios reflect both the social nature of learning interactions and necessary co-development, where the actions of the learner also affect the actions of the teacher. The first approach was a parent-infant interaction and the second a human-robot interaction [106, 107]. With respect to parent-infant interaction, we conducted a semi-experimental study in which 64 pairs of parents were asked to present a set of 10 manipulative tasks to their infant (aged 8 to 30 months) and to another adult by using both talk and manual actions. During the tasks, parent and child were sitting across a table facing each other while being videotaped with two cameras [108, 109, 110]. Parents demonstrated several tasks to their children. Some of the parents were recruited for a second study, where they were asked to demonstrate similar objects and actions to a virtual robot (see Figure 5).

Experimental work:

A Quantitative approach For the quantitative approach, we focused on investigations of child-directed speech called *motherese* and child-directed motions, called *motionese* [108]. The quantitative results pursued two goals: firstly, to provide a multimodal analysis of action demonstrations and speech in order to understand how speech and action are modified for children. Secondly, we applied our multimodal analysis methods for comparative purposes in order to characterize the interaction with a simulated robot. When we compared the data obtained from a tutor in a parent-child situation to that originating from a human-robot interaction, we found that in the case of a simulated robot, actions were modified more than speech. This virtual robot was designed to provide the tutor with visual feedback in the form of eye-gaze directed at the most salient part of the scene. Results suggest that the tutor reacted to this feedback and adapted his/her behaviour accordingly.

B Qualitative approach For the qualitative approach, we used ethnomethodological conversation analysis (EMCA) [111] as an analytical framework, providing both a theoretical account of social interaction and a methodology for fine-grained micro-analysis of video-taped interaction data. This perspective invited us to consider "tutoring" as a collaborative achievement between tutor and learner. It aims to understand the sequential relationship between different actions and to reveal the methods participants deploy to organize their interaction, and solve the practical tasks at hand.

We undertook systematic annotation of the corpus using both manual and computational methods. Central to a qualitative approach is the relation of the data from different annotation sources to each other, so that a close interaction loop can be demonstrated [110].

C Integrative approach By integrative methods, we mean computational approaches that allow us to analyse phenomena in developmental studies. More specifically, we assume that we can better understand the function of parental behavioural modifications when we consider the interplay of different modalities. We need to better understand how specific features of motherese, such as stress, pauses, specific aspects of intonation on a phonological level or particular construction on a syntactical level are related to specific parts of actions on objects in the real, physical world. Then we can begin to build a model of how multi-modal cues observed in tutoring situations help to bootstrap learning. Furthermore, they may help us to better understand how the emergence of meaning can be modelled in artificial systems. Examples of the integrative approach are given below.

Models of acoustic packaging At the current state of research, we assume that our model of acoustic packaging [112] is the most appropriate method for investigating the interplay of action and speech, as this algorithmic solution enables us to firstly, combine information about language and speech at an early processing level and secondly, analyse how parents package their actions acoustically. Models of acoustic packaging give us insight into the

functions of multimodal child-directed modifications and how multi-modal information enables a system to "understand", that is to bootstrap and then continuously refine an initial concept. This concept reflects the basic structure of actions that are being demonstrated.

Models of cognitive systems Another example of integrative methods are parallel experiments with humans and artificial cognitive systems with the aim of building simple but realistic models. We tested categorization in human-human and machine-machine experiments, in which there were relevant and irrelevant features. For the human-human side of the experiment, participants learned the required categories through interaction with a teacher. The machines deduced categories through feedback on their actions. This methodology was first used in a study by Morlino et al. [113, 114].

Outlook: In our experiments, we focused on social interaction. The objective was to see what type of teaching behaviour would improve an agent's learning. Human and artificial agents were tested in parallel. The focus of the experiments was on the types of instructions that the tutor gave in the experiments. Two types of teaching strategies emerged. One centred on negative and positive feedback, whereas the other strategy attempted to symbolize the action required from the learner,

In this work, the feedback given by the tutor via symbols was quantified, so that the different types of feedback could be modelled to create an artificial tutor. The experiment was then conducted using the artificial cognitive system. The tutor was modelled on a human tutor, whereas the learner was an artificial neural network. The aim was to yield insights into what type of feedback allows for and improves category learning in artificial agents, and to give insight into the consequences for cognitive and social robotics.

Future research needs to explore (i) the question of synchrony and (ii) the question of contingent interaction. With respect to (i), we need to investigate correlations between action and speech, for example, how are attention keeping functions in motionese, such as slow or exaggerated actions, accompanied by motherese. Similarly, how are verbal attention-getters accompanied by actions?

With respect to (ii), the question of contingent interaction, our qualitative analysis shows that for successful tutoring, it is not sufficient to simply look at synchrony between speech and action, its interactional dimension must also be considered. The way in which tutors present an action is not only characterized by synchrony between talking and action, but also by the interpersonal coordination between tutor and learner [115].

4.3 Analysing user expectations in human-robot interaction

Introduction: Interactions do not take place in a void: they are influenced by certain prior assumptions, preconceptions and expectations about the interaction partner and the interaction.

Methodologically, this is useful, because the impact of such assumptions becomes apparent in asymmetric interactions. In human-robot interaction, different preconceptions have been shown to have a considerable influence [116, 117]. As such, first, in order to predict people's behaviour in interactions with a robot and second, to guide them into appropriate behaviours that facilitate the interaction, as well as the bootstrapping of language, experimental studies are necessary to determine what influences users' expectations, as well as their subsequent behaviours. For example, in interactions with children, caregivers employ numerous cues that may facilitate language learning. Whether and to what degree users can be made to employ such features when interacting with robots is therefore an important question [118]. Understanding the similarities and differences between child-directed and robot-directed speech, as well as their determining factors is furthermore crucial to predicting how people will interact with an unfamiliar robot in novel communication situations. Thus, it is desirable to understand what drives the linguistic choices people make when talking to particular artificial communication partners.

Experimental work: In the context of the project, we carried out controlled interaction experiments in which only one aspect of the interaction was varied. Factors that influence users' expectations comprise, for example, the appearance of the robot and its degrees of freedom, as well as further aspects of robot embodiment [119], its communicative capabilities [120] and behaviours [121]. These factors were investigated in experimental settings in which all participants were confronted with the same robot and very similar robot behaviours; however, one aspect of the robot was varied at a time. We then analysed users' behaviour in these interactions, especially their linguistic behaviour, since the ways in which people design their utterances reveal how they think about their communication partner [119].

It is unlikely, however, that people do not update their preconceptions during an interaction and as such, the relationship between users' expectations and processes of alignment and negotiation on the basis of the robot's behaviour needs to be taken into account. For example, we identified preconceptions in greetings and analysed how users' behaviour changed over the course of the interaction or a set of interactions [121]. Furthermore, we had the robot behave differently with respect to one feature, for example, contingent versus non-contingent gaze [115].

Outlook: The experiments that were carried out support the view that user expectations continue to play a considerable role over the course of interactions, since they essentially constrain their own revision; thus, if people understand interaction with the robot as social, they will be willing to update their partner model, based on the robot's behaviour.

If, however, they understand human-robot interaction as tool-use, they will not be willing to take the robot's behaviour into account to the same extent. Future work will need to identify further means for eliciting and possibly changing users' preconceptions, and design interventions that can potentially shape users' expectations and subsequent behaviours [118].

4.4 Linguistic corpora studies to investigate child language acquisition

Introduction: In child language acquisition research, a major empirical methodology is the study of child language and child-directed speech, as documented in linguistic corpora [122]. One of the aims of our study was to assess whether ease of acquisition is better viewed as a function of semantic concreteness/qualitative salience or rather, of input frequency/quantitative salience to the child.

For an overview of corpus-based studies of child language acquisition, see Behrens (2008) [123]. Compared to experimental approaches, the advantages of using corpus data for the study of child language acquisition include their ecological validity (all elements in the dataset are naturally occurring) their principled suitability for longitudinal research (covering longer time spans than is feasible with individual experimental sessions), the fact that they are freely available in large quantities, as well as the fact that they are machine-readable and, given appropriate annotation, conveniently processed in ways that give rise to unique analytical possibilities.

The disadvantages of corpus studies vis-a-vis experimental approaches are that the context of the productions in the corpus is not controlled, that there is no direct cueing of the specific phenomenon or behaviour at issue in a given study and that many potentially relevant context properties of the transcribed interactions (e.g., participants' gaze behaviour or gestures) are often not preserved. Furthermore, depending on the specific corpus that is chosen for a given study, factors such as corpus size, sample density and where applicable, longitudinal span may impose additional limitations on the types of research questions that can be reasonably investigated with a given resource.

Another limitation of data transcribed from audio recordings is that orthographic transcripts do not properly represent the actual sounds that are heard (see section 3.1.

Experimental work: Apart from the immediate theoretical implications of the empirical results of such studies, they can also inform the process of constructing suitable stimuli for later experiments and computational investigations. For example, in a scenario where a robot is faced with the challenge of acquiring several different constructional patterns in parallel (comparable to the situation of a child), statistical properties of the input that are assumed to influence the acquisition process in children can be transferred to the robotic scenario in which they can be systematically manipulated and explored [124]. Experiments can vary such quantitative parameters as the availability or strength of distributional cues to a particular category in the input, the frequency proportions between different variants of a given pattern, the amount of lexical overlap between two or more different patterns in the input and so on [125].

In this project, corpus studies of naturalistic input patterns were conducted for a number of the most elementary grammatical constructions of English. Typical uses and functions of these patterns in child-directed speech were investigated in large-scale corpus studies of caregiver utterances in 25 English language corpora found in the CHILDES database [122].

The aim of the study was to assess whether ease of acquisition is better viewed as a function of semantic concreteness/qualitative salience or of input frequency/quantitative salience to the child. For the investigated corpora, the results pointed in the direction of quantitative salience – another example of statistical learning. However, questions remain whether some of the seemingly late-acquired variants in fact only exhibited late in the data because the corpora were not large and/or dense enough to register possible earlier uses of these variants.

Outlook: Some of the difficulties described above appeared in our corpus studies. However, these are clearly not principled problems and in fact, much current work goes into compiling ever larger, denser and more fully annotated corpora that are often aligned with audio and/or video recordings in order to capture ever more features of the scene [126]. Thus, linguistic corpora studies can be employed to investigate hypotheses concerning language acquisition in children that may be transferable to robots.

5. Conclusion

The research described in this paper was by design based on a multifaceted approach. The initial premise was that co-development of action, linguistic skills, conceptualization and social interaction jointly contribute to the scaffolding of language capabilities, and an overview of research areas addressed is shown in Table 1. However, though these elements ultimately all come together, they can also be profitably studied in smaller combinations, or even separately.

This heterogeneous approach to complementary, mutually scaffolded skills is supported by findings in neuroscientific research. Consider the experiments described in sections 2 and 3, which employ varied statistical learning processes as they focus on different aspects of language learning. In a wider domain, statistical learning is constrained to operate within specific modalities, which contribute to domain-general mechanisms [2].

Our work included simulations of integrative neural processes analogous to those of humans (section 2.1) and the development of the epigenetic robotics architecture (ERA), a structure for integrating a wide range of cognitive phenomena (section 2.2). An example of the result of

training iCub for the meaningful use of words and actions can be seen in a video clip on YouTube: youtu.be/5l4LHD2lYJk. Note that 'l' is the lower case letter 'ell'.

Experiments with iCub were carried out combining visual, audio and proprioceptive perceptions for learning the meanings of words (section 2.3), leading on to work pertaining to proto-grammatic compositional forms. The acquisition of negation was investigated, adding valenced preferences in the cognitive architecture to the robot's "experience" of objects (section 2.4).

Prior to integration, work on components of language learning processes were studied separately. Real-time interactive experiments with iCub demonstrated how an unsegmented audio stream might be processed and how the transition from babbling to word form productions might occur (section 3.1). Another approach to learning the meanings of words through social interaction was investigated using the language games paradigm (section 3.2).

Since the integration of action and language is central to our hypotheses, the implementation of goal-directed actions in robots is a key factor. Section 3.3 describes the theory and practice of the passive motion paradigm (PMP). This approach avoids the classic problems of optimizing movements of robot joints with multiple degrees of freedom, where the indefinitely large number of possible moves to achieve a goal generates ill-posed problems.

One consequence of the PMP approach is a shift from low-level analysis to the structure of internal models (see the end of section 3.3). This needs to be reconciled with enactive, sensorimotor theory underlying the project's approach (see section 1, Introduction), which proposes raw, uninterpreted perceptual experience for scaffolding the acquisition of behaviours [8].

Analogies with child language learning A theme that runs throughout this work is that language learning and conceptualization in our agents are inspired by the development of these capacities in the child. However, though the work is *inspired* by child development, there are many obvious fundamental differences, starting with the contrast between a human body and humanoid hardware. No claim is made for modelling child language acquisition in general, though some specific features have been modelled; for example, bodily bias in child word learning, as described in section 2.2 [37].

Similarities with child language learning are reflected in the experimental design. As such, our work included research into contingent interaction through speech and gestures, (sections 4.1, 4.2, 4.3) and studies on corpora of child language (section 4.4). This work feeds back into experimental design, notably into methods described in sections 2.3, 2.4 and 3.1, where contingent interaction with teachers plays a dominant role.

The novel work on motor resonance, a crucial mechanism in the integration of action and language learning (section 4.1), brings a new, multi-disciplinary approach to investigating communication between human and robotic agents.

However, we need to keep in mind that some learning processes are essential to the acquisition of speech, while others facilitate learning but are not absolutely essential. Although they can learn to use signed or written language, children that are profoundly deaf from birth cannot learn to speak. On the other hand, blind infants can learn to speak, albeit typically at a slower rate than their sighted contemporaries.

Among the differences between language learning arising from the experimental process and language learning in humans, we noted that a real child was generally immersed in a learning environment all day long, whereas the robotic subjects of our experiments had short, task-based sessions, perhaps more akin to therapeutic scenarios. Other aspects of our work that are not in full accord with the real child include neural modelling based on back-propagation learning, which does not have a biological basis. Other divergences include the use of orthographic transcripts in child speech corpora, which only partially represent auditory perceptions. Though this is well-known in speech recognition engineering, it has received little attention in corpus research. We also note that the articulatory abilities of the iCub in babbling-to-word forms experiments did not properly match infant productions.

Overall, however, the project progressed our understanding of language learning and cognitive bootstrapping, and how it might be applied in robotics. One important aspect of the methods applied in the field of language and action learning is that the methods themselves can lead to new and interesting insights for further theoretical proposals. One example of this is where the methods outlined in section 2.2, which discuss the use of the epigenetics robotics architecture (ERA), were applied in order to discover whether effects found in psychological experiments on early language learning with children would also occur in similar experiments with the iCub humanoid robot [38, 37]. The results of these experiments led to a revision of the theoretical ideas supporting such proposals and was further analysed using newer variations on such methods. This process, which we label *research loops*, is an important outcome that fuses together work from the field of robotics and physically embodied studies that includes work on human development. In effect, methods are used to verify theoretical ideas on an experimental robotic platform in a way that would not be possible with human children or adults.

In conclusion, we hope that this article will allow researchers in the field of embodied language learning to assess and enhance the methodologies exhibited and we look forward to seeing further progress in this field.

6. Acknowledgements

The work described in this article was conducted within the framework of the EU Integrated Project ITALK ("Integration and Transfer of Action and Language in Robots"), funded by the European Commission under contract number FP7-214668.

7. References

[1] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C.l. Nehaniv, et al. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2(3):167–195, 2010.

[2] R Frost, B C Armstrong, N Siegelman, and M H Christiansen. Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Science*, 19(3), 2015.

[3] S. Harnad. The symbol grounding problem. *Physica D.*, 42:335–346, 1990.

[4] L Steels and M Hild. *Language Grounding in Robots*. Springer-Verlag, New York, 2012.

[5] Michael Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003.

[6] F Varela, E Thompson, and E Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.

[7] J. K. O'Regan and A. Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24(5):939–973, 2001.

[8] C L Nehaniv, F Förster, J Saunders, et al. Interaction and Experience in Enactive Intelligence and Humanoid Robotics. In *Symposium on Artificial Life, IEEE ALIFE*, 2013.

[9] Caroline Lyon. Beyond vision: Extending the scope of a sensorimotor account of perception. In J. M. Bishop and A. O. Martin, editors, *Contemporary Sensorimotor Theory*. Springer, 2014.

[10] Luc Steels. Self-organizing vocabularies. In C. Langton and T. Shimohara, editors, *Proc. Artificial Life V (Alife V)*. MIT Press, 1996.

[11] Luc Steels and Tony Belpaeme. Coordinating perceptually grounded categories through language. A case study for colour. *Behavioral and Brain Sciences*, 24(8):469–529, 2005.

[12] Frank Broz, Chrystopher L. Nehaniv, Angelo Cangelosi, et al. The ITALK Project: A developmental robotics approach to the study of individual, social and linguistic learning. *Topics in Cognitive Science*, 6(3), 2014.

[13] F. Pulvermueller, M. Haerle, and F. Hummel. Walking or talking? Behavioral and neurophysiological correlates of action verb processing. *Brain and Language*, 78:134–168, 2001.

[14] A. M. Glenberg. Language and action: creating sensible combinations of ideas. In G. Gaskell, editor, *The Oxford Handbook of Psycholinguistics*. Oxford University Press, 2007.

[15] H. H. Clark. Space, time, semantics and the child. In T.E. Moore, editor, *Cognitive Development and the Acquisition of Language*, pages 27–64. Academic Press, New York, USA, 1973.

[16] L. B. Smith, S. S. Jones, and B. Landau. Naming in young children: A dumb attentional mechanism? *Cognition*, 60:154–171, 1996.

[17] K. Nelson. Some evidence for the cognitive primacy of categorization and its funcional basis. *Merrill-Palmer Quarterly*, 69:21–39, 1973.

[18] J. M. Mandler. How to build a baby: II. Conceptual primitives. *Psychological Review*, 99:587–604, 1992.

[19] J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[20] M. H. Christiansen and N. Chater. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205, 1999.

[21] D. Chalmers. Syntactic transformations on distributed representations. *Connection Science*, 2:53–62, 1990.

[22] D. Marocco, K. Fischer, T. Belpaeme, and A. Cangelosi. Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot. *Frontiers in Neurorobotics*, 4(7), 2010.

[23] 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.

[24] Y. Sugita and J. Tani. Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, 13(3):211–225, 2005.

[25] P. Dominey. Emergence of grammatical constructions: Evidence from simulation and grounded agent experiments. *Connection Science*, 17(3-4):289–306, 2005.

[26] 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):201–233, 2008.

[27] M. Peniak, D. Marocco, J. Tani, Y. Yamashita, K. Fischer, and A. Cangelosi. Multiple time scales recurrent neural network for complex action acquisition. In *IEEE ICDL-EPIROB*, 2011.

[28] 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.

[29] R. D. Beer. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509, 1995.

[30] P. J. Werbos. Backpropagation through time: What it does and how to do it. In *Proceedings of the IEEE*, volume 78(10), pages 1550–1560. IEEE, 1990.

[31] E. Tuci, T. Ferrauto, G. Massera, and S. Nolfi. Co-development of linguistic and behavioural skills: compositional semantics and behaviour generalisa-

tion. In *Proc. Conf. on Simulation of Adaptive Behavior (SAB2010)*. Springer, 2010.

[32] Martin Peniak and Angelo Cangelosi. Scaling-up action learning neuro-controllers with GPUs. In *Neural Networks (IJCNN), 2014 International Joint Conference on*, pages 2519–2524. IEEE, 2014.

[33] A. F. Morse, J. DeGreeff, T. Belpeame, and A. Cangelosi. Epigenetic robotics architecture (ERA). *IEEE Transactions on Autonomous Mental Development*, 2(4):325–339, 2010.

[34] Alva Noë. *Action in Perception (Representation and the Mind)*. MIT Press, 2004.

[35] J. de Greeff and T. Belpaeme. Why robots should be social: Enhancing machine learning through social human-robot interaction. *PLoS one*, September 2015.

[36] A. F. Morse. Snapshots of sensorimotor perception. In V Muller, editor, *Philosophy and Theory of Artificial Intelligence*. Springer, 2013.

[37] A. F. Morse, T. Belpeame, A. Cangelosi, and L. B. Smith. Thinking with your body: Modelling spatial biases in categorization using a real humanoid robot. In *Paper presented at Cognitive Science*, 2010.

[38] A. F. Morse, P. Baxter, T. Belpeame, L. B. Smith, and A. Cangelosi. The power of words (and space). In *Proc. (ICDL-EPIROB)*, 2011.

[39] A. F. Morse, T. Belpeame, A. Cangelosi, and C. Flocia. Modeling U-shaped performance curves in ongoing development. In *Paper presented at Cognitive Science*, 2011.

[40] Anthony Morse, V.L. Benitez, T. Belpaeme, A. Cangelosi, and L.B. Smith. Posture affects word learning in robots and infants. *PLOS One*, 2015.

[41] Joe Saunders, Chrystopher L. Nehaniv, and Caroline Lyon. Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors. In *Proc. New Frontiers in Human-Robot Interaction, AISB Convention*, 2010.

[42] P. K. Kuhl. Is speech learning gated by the social brain? *Developmental Science*, 10(1):110–120, 2007.

[43] Paul Bloom. *How Children Learn the Meaning of Words*. MIT Press, 2002.

[44] J Saunders, H Lehmann, F Förster, and C L Nehaniv. Robot acquisition of lexical meaning: Moving towards the two-word stage. In *IEEE ICDL-EPIROB*, 2012.

[45] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[46] R. Pea. The development of negation in early child language. In D.R. Olson, editor, *The Social Foundations of Language and Thought: Essays in Honor of Jerome S. Bruner*. W.W. Norton, 1980.

[47] F. Förster, C. L. Nehaniv, and J. Saunders. Robots that say 'no'. In *Proceedings of the 10th European Conference on Artificial Life, ECAL 2009*, 2009.

[48] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer. Multimodal saliency-based bottom up-attention a framework for the humanoid robot iCub. In *International Conference on Robotics and Automation, ICRA*, pages 962–967. IEEE, 2008.

[49] Walter Daelemans and Antal van den Bosch. *Memory-Based Language Processing*. Cambridge University Press, 2005.

[50] Frank Förster. *Robots that Say 'No': Acquisition of Linguistic Behaviour in Interaction Games with Humans*. PhD thesis, Adaptive Systems Research Group, University of Hertfordshire, 2013.

[51] C Lyon, J Saunders, and C L Nehaniv. Interactive language learning by robots: The transition from babbling to word forms. *PLoS One*, 7(6), 2012.

[52] M. Vihman, R. DePaolis, and T. Keren-Portnoy. A dynamic systems approach to babbling and words. In E. Bavin, editor, *Handbook of Child Language*, pages 163–182. CUP, 2009.

[53] H Yeung and J Werker. Learning words' sounds before learning how words sound: 9-months-old infants use distinct objects as cues to categorize speech information. *Cognition*, 113(2), 2009.

[54] Benedicte de Boisson-Bardies. *How Language Comes to Children*. MIT, 1999.

[55] A. Fernald and V. A. Marchman. Language learning in infancy. In M. J. Traxler and M. A. Gernsbacher, editors, *Handbook of Psycholinguistics*, pages 1027–1071. Elsevier, 2nd edition, 2006.

[56] J. Saffran, R. Aslin, and E. Newport. Statistical learning by 8-month-old infants. *Science*, 274:1926–1928, 1996.

[57] Patricia K. Kuhl. Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11):831–843, 2004.

[58] A. Bigelow and C. Decoste. Sensitivity to social contingency from mothers and strangers in 2-, 4-, and 6-month-old infants. *Infancy*, 4:111–140, 2004.

[59] Britta Wrede, Stefan Kopp, Katharina Rohlfing, Manja Lohse, and Claudia Muhl. Appropriate feedback in asymmetric interactions. *Journal of Pragmatics*, 42:2369–2384, 2010.

[60] Britta Wrede, Katharina Rohlfing, Marc Hanheide, and Gerhard Sagerer. *Towards Learning by Interacting*, pages 139 – 150. Springer, 2009.

[61] P. Jusczyk and R. Aslin. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29:1–23, 1995.

[62] A Rothwell, C Lyon, C L Nehaniv, and J Saunders. From babbling towards first words: the emergence of speech in a robot in real-time interaction. In *IEEE Symposium on Artificial Life (IEEE Alife 2010)*, pages 86–91, 2011.

[63] C Lyon, C L Nehaniv, and J Saunders. Preparing to Talk: Interaction between a Linguistically Enabled Agent and a Human Teacher. In *AAAI Fall Symposium Series*, Dialog with Robots, FS-10-05, 2010.

[64] Joe Saunders, Hagen Lehmann, Yo Sato, and Chrystopher L. Nehaniv. Towards using prosody to scaffold lexical meaning in robots. In *Proc. IEEE ICDL-EPIROB*, 2011.

[65] S Greenberg. Speaking in shorthand: A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29:159–176, 1999.

[66] Gregory Hickok and David Poeppel. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, 92(1-2):67–99, 2004.

[67] D Saur, B W Kreher, S Schnell, et al. Ventral and dorsal pathways for language. *Proc. of the National Academy of Science*, 105(46), 2008.

[68] S Chang, M Wester, and S Greenberg. An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. *Speech Communication*, 2005.

[69] J. de Greeff, F. Delaunay, and T. Belpaeme. Human-robot interaction in concept acquisition: a computational model. In *Proc. of IEEE ICDL*, 2009.

[70] C S Sherrington. The integrative action of the nervous system. *Yale University Mrs Epsa Ely Silliman memorial lectures*, 1906.

[71] N A Bernstein. *The Coordination and Regulation of Movements*. Pergamon Press, 1967.

[72] E Todorov, W Li, and X Pan. From task parameters to motor synergies: A hierarchical framework for approximately optimal control of redundant manipulators. *Journal of Robotic Systems*, 22(11):691–710, 2005.

[73] E Todorov. Optimal Control Theory. *Bayesian Brain Probabilistic Approaches to Neural Coding*, pages 269–298, 2006.

[74] Karl Friston. Perspective What Is Optimal about Motor Control ? *Neuron*, 72(3):488–498, 2011.

[75] Vishwanathan Mohan and Pietro Morasso. Passive motion paradigm: an alternative to optimal control. *Frontiers in Neurorobotics*, 5, 2011.

[76] Emanuel Todorov. Efficient computation of optimal actions. *Proc. National Academy of Sciences*, 106(28): 11478–11483, 2009.

[77] F A Mussa Ivaldi, P Morasso, and R Zaccaria. Kinematic networks. A distributed model for representing and regularizing motor redundancy. *Biological Cybernetics*, 60(1):1–16, 1988.

[78] V Mohan, P Morasso, G Metta, and G Sandini. A biomimetic, force-field based computational model for motion planning and bimanual coordination in humanoid robots. *Autonomous Robots*, 27(3):291–307, 2009.

[79] V Mohan, P Morasso, G Sandini, and S Kasderidis. Inference through embodied simulation in cognitive robots. *Cognitive Computation*, 51, 2013.

[80] A. Feldman. Functional tuning of the nervous system with control of movement or maintenance of a steady posture. *Biophysics*, 11:925–935, 1966.

[81] E Bizzi, F A Mussa-Ivaldi, and S Giszter. Computations underlying the execution of movement: a biological perspective. *Science*, 253(5017):287–291, 1991.

[82] E Bizzi, N Hogan, F A Mussa-Ivaldi, and S Giszter. Does the nervous system use equilibrium-point control to guide single and multiple joint movements? *Behavioral and Brain Sciences*, 15:603–613, 1992.

[83] N. Hogan. Modularity and causality in physical system modeling. *ASME Journal of Dynamic Systems Measurement and Control*, 109:384–391, 1987.

[84] Vishwanathan Mohan, Pietro Morasso, Giorgio Metta, and Stathis Kasderidis. Actions and Imagined Actions in Cognitive Robots. In *Perception-Action Cycle*, pages 539–572. Springer, 2011.

[85] Vishwanathan Mohan and Pietro Morasso. Towards reasoning and coordinating action in the mental space. *International Journal of Neural Systems*, 17(4):329–341, 2007.

[86] Pietro Morasso, Maura Casadio, Vishwanathan Mohan, and Jacopo Zenzeri. A neural mechanism of synergy formation for whole body reaching. *Biological Cybernetics*, 102(1):45–55, 2010.

[87] Vishwanathan Mohan, Pietro Morasso, Jacopo Zenzeri, Giorgio Metta, V Srinivasa Chakravarthy, and Giulio Sandini. Teaching a humanoid robot to draw shapes. *Autonomous Robots*, 31(1):21–53, 2011.

[88] V. Mohan and P. Morasso. A forward / inverse motor controller for cognitive robotics. *Artificial Neural Networks - ICANN 2006*, pages 602–611, 2006.

[89] P Morasso and V Sanguineti. *Self-organization, computational maps, and motor control*, volume 119. North Holland, 1997.

[90] A G Feldman. Functional Tuning of the Nervous System with Control of Movement or Maintenance of a Steady Posture. II. Controllable Parameters of the Muscle. *Biophysics*, 11(3):498–508, 1966.

[91] J Decety. Do imagined and executed actions share the same neural substrate? *Brain Research*, 3(2):87–93, 1996.

[92] M Jeannerod. Neural simulation of action: a unifying mechanism for motor cognition. *NeuroImage*, 14(1 Pt 2):S103–S109, 2001.

[93] A. Karniel. Open questions in computational motor control. *J Integr Neurosci*, 10(3):385–411, 2011.

[94] G. Rizzolatti, L. Fadiga, L. Fogassi, and V. Gallese. Resonance behaviors and mirror neurons. *Archives Italienne de Biologie*, 137(2-3):85–100, 1999.

[95] J. R. Flanagan and R. S. Johansson. Action plans used in action observation. *Nature*, 424:769–771, 2003.

[96] A Sciutti, A Bisio A, F Nori, G Metta, L Fadiga L., and G Sandini. Robots can be perceived as goal-oriented agents. *Interaction Studies*, 2013.

[97] A Sciutti, A Bisio, F Nori, et al. Measuring human robot interaction through motor resonance. *International Journal of Social Robotics*, 4(3), 2012.

[98] T Chaminade and G Cheng. Social cognitive neuroscience and humanoid robotics. *J Physiol. Paris*, 103(3-5), 2009.

[99] G. Rizzolatti and L. Craighero. The mirror-neuron system. *Annual Review of Neuroscience*, 27:169–192, 2004.

[100] F Dehais, E A Sisbot, R Alami, and M Causse. Physiological and subjective evaluation of a human-robot object hand-over task. *Appl. Ergon.*, 42(6), 2011.

[101] A Bisio, A Sciutti, F Nori, G Metta, L Fadiga, G Sandini, and T Pozzo. Motor Contagion during Human-Human and Human-Robot Interaction. *PloS one*, 9(8), 2014.

[102] G Baud-Bovy, P Morasso, F Nori, G Sandini, and A Sciutti. Human machine interaction and communication in cooperative actions. In R Cingolani, editor, *Bioinspired Approaches for Human-Centric Technologies*. Springer, 2014.

[103] T. Falck-Ytter, G. Gredebäck, and C. von Hofsten. Infants predict other people's action goals. *Nature Neuroscience*, 9(7):878–879, 2006.

[104] R van Baaren, R Holland, K Kawakami, and A van Knippenberg. Mimicry and prosocial behavior. *Psychol Sci*, 15 (1), 2004.

[105] A Bisio, N Stucchi, M Jacono, L Fadiga, and T Pozzo. Automatic versus voluntary motor imitation: effect of visual context and stimulus velocity. *PLoS One*, 5(10), 2010.

[106] Kerstin Fischer, Kilian Foth, Katharina Rohlfing, and Britta Wrede. Mindful tutors – linguistic choice and action demonstration in speech to infants and to a simulated robot. *Interaction Studies*, 12(1), 2011.

[107] Katrin Lohan, Katharina Rohlfing, Karola Pitsch, et al. Tutor spotter: Proposing a feature set and evaluating it in a robotic system. *International Journal of Social Robotics*, 4(2), 2012.

[108] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199, 2006.

[109] A. L. Vollmer, K. S. Lohan, K. Fischer, et al. People modify their tutoring behavior in robot-directed interaction for action learning. In *Proc. of DEVLRN â€™09: IEEE ICDL*, 2009.

[110] K Pitsch, A-L Vollmer, K J Rohlfing, J Fritsch, and B Wrede. Tutoring in adult-child interaction. On the loop of the tutor's action modification and the recipientâ€s gaze. *Interaction Studies*, 15:55-98, 2014.

[111] M. Rapley. Ethnomethodology/conversation analysis. In D. Harper and A. R. Thompson, editors, *Qualitative Research Methods in Mental Health and Psychotherapy*. John Wiley, 2011.

[112] L. Schillingmann, B. Wrede, and K.J. Rohlfing. A computational model of acoustic packaging. *IEEE Transactions on Autonomous Mental Development*, 1(4):226 –237, 2009.

[113] G. Morlino, C. Gianelli, A. M. Borghi, and S. Nolfi. Developing the ability to manipulate objects: A comparative study with human and artificial agents. In *Proc. Epigenetic Robotics*, pages 169–170, 2010.

[114] S Griffiths, S Nolfi, G Morlino, et al. Bottom-up learning of feedback in a categorization task. In *IEEE ICDL-EPIROB*, 2012.

[115] Kerstin Fischer, Katrin S. Lohan, Joe Saunders, Chrystopher Nehaniv, Britta Wrede, and Katharina Rohlfing. The impact of the contingency of robot feedback on HRI. In *Proc. Workshop on Collaborative Robots and Human Robot Interaction (CR-HRI 2013)*, 2013.

[116] Steffi Paepcke and Leila Takayama. Judging a bot by its cover: An experiment on expectation setting for personal robots. In *Proceedings of Human Robot Interaction (HRI), Osaka, Japan*, 2010.

[117] Kerstin Fischer. Interpersonal variation in understanding robots as social actors. In *Proc. HRI'11*, 2011.

[118] Kerstin Fischer. Alignment or collaboration? how implicit views of communication influence robot design. In *Proc. Conference on Cooperative Technological Systems*, 2014.

[119] Kerstin Fischer, Katrin Lohan, and Kilian Foth. Levels of embodiment: Linguistic analyses of factors influencing HRI. In *Proc. HRI'12*, 2012.

[120] Kerstin Fischer, Bianca Soto, Caroline Pantofaru, and Leila Takayama. The role of social framing in initiating human-robot interaction. In *Proc. IEEE Symposium on Robot and Human Interactive Communication, Ro-Man '14*, 2014.

[121] Kerstin Fischer and Joe Saunders. Getting acquainted with a developing robot. In *Human Behavior Understanding*. Springer LNCS 7559, 2012.

[122] B. MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Erlbaum, 1995.

[123] H Behrens. *Corpora in language acquisition research: History, methods, perspectives*. John Benjamins, 2008.

[124] A Goldberg. *Constructions at Work*. OUP, 2006.

[125] Arielle Borovsky and Jeff Elman. Language input and semantic catergories: A relation between cognition and early word learning. *Journal of Child Language*, 33, 2006.

[126] D Roy, R Patel, P DeCamp, et al. The human speechome project. In *Proc. Cognitive Science Conference*, 2008.