

Embodied Language and Number Learning in Developmental Robots

Angelo Cangelosi (1), Anthony Morse (1), Alessandro Di Nuovo (1,2), Marek Rucinski (1), Francesca Stramandinoli (1,3), Davide Marocco(1), Vivian De La Cruz (4), Kerstin Fischer (5)

1. Plymouth University, UK
2. University of Enna Kore, Italy
3. Italian Institute of Technology
4. Messina University, Italy
5. University of Southern Denmark

1. Developmental Robotics for Embodied Cognition and Symbol Grounding

Computational models can play an important role in cognitive science as they permit the test and validation of psychology theories by forcing the operationalization of general (and sometimes loose) theoretical concepts into detailed operations which must run on a computer program. Models capable of replicating known behavioral and cognitive phenomena, including the reproduction of errors and impaired performance, can also be used to make further predictions to revise and refine psychological theories (Pezzullo et al. 2013; Cangelosi and Parisi 2002).

Robotics is the obvious candidate for the design of computational models to test embodiment theories, given the focus of embodied cognition on perceptual and motor phenomena. Robots, by their own nature, rely on the implementation of behavioral control architectures integrating multimodal sensing (through vision, audition, tactile sensors) and of motor capabilities (e.g. actuators for locomotion and manipulation, and speech production). The robotic approaches to the modeling of behavioral and cognitive phenomena are normally referred to as “cognitive robotics” and include a variety of methods such as evolutionary robotics, for the modeling of evolutionary phenomena, developmental robotics, for the modeling of learning and development phenomena, and neurorobotics, for a focus on the neural control of behavior. This chapter presents a set of studies based on the developmental robotics approach looking specifically at the modeling of embodied phenomena in the acquisition of linguistic and numerical cognition capabilities.

Developmental Robotics is the “interdisciplinary approach to the autonomous design of behavioral and cognitive capabilities in artificial agents (robots) that takes direct inspiration from the developmental principles and mechanisms observed in natural cognitive systems (children).” (Cangelosi & Schlesinger, 2015: 4). As such this approach puts a strong emphasis on the constraining of the robot’s cognitive architecture and behavioral and learning performance onto known child psychology theories and data. This allows the modeling of the
developmental succession of qualitative and quantitative stages leading to the acquisition of adult-like cognitive skills.

Developmental robotics has been applied to the modeling of a variety of cognitive phenomena as intrinsic motivation, motor and perceptual development, social learning, language acquisition and the learning of abstract knowledge skills (see Cangelosi and Schlesinger 2015 for a recent and comprehensive review of the state of the art in this field; See also Asada 2009 and Lungarella et al 2003 for additional reviews). Most of these studies put a strong emphasis of the interaction between the developing (baby) robot and its physical and social environment. Such an approach is naturally suited to model embodied and situated cognition for the grounding of cognition (Pezzulo et al. 2013). Especially with respect to the embodied basis of language learning, the use of robots which have to learn to name objects they see and name actions they perform, to communicate with other robots and human participants, offers the natural tool to model the grounding of symbols in sensorimotor knowledge and experience. Robotics thus provides a modeling tool to address the issue of the symbol grounding problem in cognitive modeling and artificial cognitive systems (Harnad 1990; Cangelosi 2011).

This chapter provides a brief overview of a series of recent studies on the use of developmental robotics specifically for the modeling the embodied acquisition of linguistic and numerical skills in robots and their grounding in perceptual and motor experience. The first section looks at the modeling of early word learning of object names based on the embodiment biases of posture and space. This model is then extended to the learning of simple but abstract grammatical structure and multi-word sentences, and further to the acquisition of words with more abstract meanings. The final section goes further in the symbol grounding hierarchy to show how the use of counting gestures and finger counting strategies help the robot to acquire number cognition skills.

2. Learning first words: Modeling bodily biases in early object name learning

Recent infant studies investigating how naïve infant learners come to map a name to an object suggest that body posture might be playing a critical role (Samuelson et al., 2011). To learn anything at all from real time experiences, a physical learner must be able to orient its sensors, and thereby its physical body, in order to attend to the referred object. Part of the learning challenge for a physical agent then is to react appropriately, e.g. orienting to the spatial locations of objects. Here we investigate how this may be achieved egocentrically using body posture. We present an embodied approach, mapping posture to expected sensory experience, and explore what implications this approach might have.

Samuelson et al (2011), continuing the work of Baldwin (1993) showed that there are clear posture and spatial biases in infants’ learning of the mapping
between words and objects. In their experimental setup, infants repeatedly experience two new objects (the target and the foil) in consistent but different locations. Subsequently they hear the object name ‘modi’ while attending to the foil object which has now been placed in the location normally associated with the target object. On testing with both objects present in new locations, the infant is asked ‘where is the modi’, and the statistically significant majority of children select the target object. This means that they are selecting the object normally associated with the spatial location they were attending to, rather than the actual object they were looking when they heard the name. By implication, this means that infants cannot be using a simple mapping between the object features being observed and the word detected at that point in time. They rather rely on a memory for own posture and the related object location to associate objects and their names. Samuelson et al (2011) provide a neural field model replicating these results in which spatial locations and physical appearance are combined to solve the binding problem – which object features belong to which object – and learn the associated label. Their model, however, used an abstract, “disembodied” representation of space. Here we extend their investigation with a new embodied model and the iCub humanoid robot, exploring further how this spatial component can be achieved via the robot’s physical interaction with objects and locations. In our model we associate object features directly to the body posture in which they are encountered, and also associate words to the posture in which they are encountered. As a result, body posture acts as a “hub” such that spreading activation via the associating leading to postural representation allows words or visual features to indirectly prime each other, via the intermediary hub of the body posture. The resulting model qualitatively captures the infant data and makes interesting predictions that are currently being explored with new child experiments. The model also shows how initial learning constrained by the body’s momentary dispositions in space can lead to a memory and behavior apparently free from bodily constraints.

The model is an implementation of the Epigenetic Robotics Architecture (Morse et al 2010), a robot cognitive architecture specifically design for studying embodied language learning. The core of such an architecture consists of three self-organizing maps with modified Hebbian learning between their units. The first (visual) map is driven by pre-processed visual information (an HSV spectrogram of the color of each object in view), the second (body) map is driven by postural information (the current motor encoder values of the eyes, head, and torso of the robot), and the final (word) map responds uniquely to each word encountered (pre-processed by the commercial speech to text software Dragon DictateTM). The visual color map and the word map are both fully connected to the body posture map, with connection weights adjusted by a normalized positive and negative Hebbian learning rule (details can be found in Morse et al 2010). Units within each map are also fully connected within each map by constant inhibitory connections, mimicking the structure of the connectionist Interactive Activation and Competition models. The iCub robot’s initial behavior is driven by sensitivity to movement, i.e. a motor saliency maps that detects which objects or body parts move. Similarly, the priming of object features in the self-organizing maps also enhances the saliency of those features in the original
image, thereby causing iCub to orient and reach for those areas. See Morse et al. 2010 for full details.

In one version of the experiment, the target object (a red ball) is placed to the left of the iCub. The robot looks at the target for approximately 10 seconds, before the target object is removed and the foil object is placed to the right of the iCub, which again orients for approximately 10 seconds. This procedure is repeated 4 times. At the fifth presentation cycle, the foil object is placed in the position normally associated with the target object, and the word ‘modi’ is spoken. The original placements of each object are repeated one final time and then both objects are positioned in new locations, to test the robot by asking ‘find the modi’. iCub then orients and reaches for one of the objects. Various versions of the experiment were carried out, each repeated 20 times (with all learned weights reset and self-organizing map’s randomly initialized, and left-right positions counterbalanced). Results comparing the robot and child data at two different learning rates are shown in Figure 1. Having replicated existing child data, we then conducted a new experiment, following the same procedure outlined above but with the addition that we changed the iCub’s posture (from sitting to standing, or from standing to sitting, counterbalanced) for the naming event only at the fifth presentation cycle. As a result of this change in posture, the naming event occurs in a posture that hadn’t been previously associated with either the target or foil object, and so on testing the interference between previously experienced objects and that posture causes the iCub to select the foil object (the object it was observing when it first heard the name). This result has now been verified in new child experiments (Morse et al. submitted). This additional experiment shows that the infants (and the robots) use memory of postures as a way to organize their learning task. If two different postures are used, at this early stage of development they are used by the robot to separate different cognitive tasks.

The implications of this hypothesis – that body posture plays a critical role in learning and recalling object-word mappings – are far reaching. Most notably, the fact that atypical patterns of motor development are co-morbid with many cognitive developmental disorders, and abnormal movement patterns are linked with poor attentional control in children, is well known but not well understood. This work may provide a path for a mechanistic understanding of the developmental dependencies between sensorimotor processes and early cognitive development.

An extended version of this model, while retaining the abilities discussed herein, has already been used to replicate a range of fast-mapping experiments (Twomey et al. 2013), and is now being used to replicate experiments across the developmental stages of early language acquisition, providing an experienced based account of the developmental stages and transitions that occur in early language learning. The architecture is also used in the learning of semantic categories from linguistic cues, as the next section shows.
3. Learning grammar: Acquiring semantic categories from structural word order cues

Grammar learning means the learning of structural cues that encode abstract semantic configurations. One such cue is word order. Especially in languages with little morphological marking such as English, word order plays an important role (Kirkwood 1969). For example, in the sentence “The frog kissed the princess”, the order of the nouns informs the competent speaker of English who did the kissing to whom.

Word order is not a simple mapping between elements: instead, it requires an understanding of the structure of the patterns and of the kinds of semantic categories involved. For instance, a noun phrase can have many different realizations, for example, the frog, the green frog, the big green frog etc., where there is even a preference for the big green frog over the green big frog (e.g. Wulff 2003). Thus, both semantic and structural issues feed into the interpretation of grammatical structure.

In the current study, we use a developmental robotic model to design robots able to infer the semantic categories of unknown words from the word order.
acquired from previous exposure to similar sentences; that is, the robot should be able to infer that “touch the purple ball” means to touch a ball with a particular color, even if it has never seen the term “purple” before, or that “look at the green airplane” means to direct the gaze to a particular object that is green, even if the term “airplane” is new to the robot. Thus, we want our robot to learn to exploit the grammatical cue word order in order to infer the semantic category of unknown words in novel utterances.

The model is a combination of two different systems created with the aim to combine semantic information with a simple form of grammatical analysis. In the model proposed, our robot learns, like children, language in a meaningful way, which means that the utterances it encounters are grounded in its own sensorimotor experience (see Marocco et al. 2010; Morse et al. 2010). In this way, it is situated in the embodied language learning tradition, where linguistic structure is paired with sensorimotor data (e.g. Sugita & Tani 2005, 2008; Steels 2008). However, children, and even young infants (Gómez 2007), have also been shown to carry out distributional analyses of the utterances they hear as well; they extract regularities and co-occurrence relationships for several language-related categorization tasks, such as identifying the elements of the phonological inventory, segmenting words, distinguishing lexical (‘content’) from grammatical (‘function’) words, and bootstrapping syntactic categories (Saffran & Thiessen 2003; Küntay & Slobin 2001). This kind of learning from distributional cues has been shown to be very successful in the learning of structural linguistic information (e.g. Borovsky & Elman 2006, Onnis et al. 2008). Nevertheless, in the child language learner, these two processes, embodied and distributional learning, interact and influence each other. For our current learning problem, we therefore employ an architecture that combines the two approaches and that allows our humanoid robot iCub to learn new words from the interaction between distributional and sensorimotor grounded information.

The grammar learning model combines a module based on the Epigenetic Robotics Architecture (ERA, described in the previous section), which provides the grounding of words, and a Recurrent Neural Networks (RNNs) for the ability to extract temporal features from serial order analysis of linguistic structure. The ERA module can learn cross-situationally from ongoing experience abstract representations that combine and interact dynamically to produce and account for multiple cognitive and behavioural phenomena. The recurrent neural network module is trained with a standard Error Back Propagation algorithm, which learns the dynamical sequences of input-output patterns as they develop in time (see Elman 1990). In particular, we presented the robot with the following utterances with corresponding situations: touch ball, touch cube, touch red, touch green, touch green ball, touch green cube, touch red ball, touch red cube. Note that the input was so constructed that a simple mapping between the position of a word in the sentence and its semantic category is not possible, but that instead it is the order in which adjectives and nouns occur that is informative.
During learning, each word is associated with a specific activation pattern in the ERA model’s color, shape, and body posture maps. As the ERA module associates words to properties, such as the shape and color of an object, in training the neural network, every word of a sequence in the input is associated to the corresponding semantic categories provided by the ERA module as desired output. For example, in the case of “touch the red cup”, given that “touch”, “red” and “cup” are already known to the ERA module, the neural network will learn the correct association: “touch”/action, “red”/color, and “cup”/shape. In this way, the neural network implicitly learns the association between the word order in a sentence with the corresponding semantic category of each word that forms the sentence itself. After the training, therefore, the neural network is able to predict, on the basis of the position of the word in a sentence, the expected semantic category; for instance, it will predict that ‘purple’ will be the color of the object in the utterance “touch the purple cup” even if it has not seen the word ‘purple’ before.

We have tested the system proposed in an experiment using the iCub simulator (see also Marocco et al. 2010). Regarding sensorimotor experience, we use a joint on the shoulder that allows the robot to reach and move an object placed on a desk in front of it, as well as a binary tactile sensor on the hand in order to provide tactile sensory feedback. The robot’s vision system provides information about the shape of the object and its color. A parameter of the shape is calculated from the image of the object acquired by the robot and its value is added as input to the neural network controller together with the color of the object in RGB value. The robot automatically generates a movement when it receives a target joint angle as input. The movement corresponds to the target angle and is generated by means a pre-programmed proportional-integral-derivative (PID) controller. The sensorimotor state of the robot is updated every 500ms.

In the experiment, the robot was presented with two objects (cube and ball) and two colours (red and green). True and false sentences were provided to the robot, such as touch red ball (true) if a red ball is present or touch green ball (false) if only a green cube is present. Sentences with a third colour, blue, were used for testing, e.g. touch blue ball. The action required of the robot is to touch or not to touch the object. The total number of input sequences available was 32, but only 24 of those sequences were used during the training. The remaining 8 were used for performing generalisation tests. After the training, an analysis of the internal representations before and after the linguistic input was performed. A cluster analysis of the robot’s internal representations (see Figure 2) shows that the internal representations are reshaped in a way that all colour terms, also the previously unseen colour term blue, are correctly categorized as colours, based on their position in the word order. This indicates that it is possible to correctly identify semantic categories from distributional cues provided by the word order, which, in turn, allows the artificial system to apply the correct meaning to a new word on the basis of its position in the sentence and the robot’s sensorimotor data stored in the self-organising maps of the ERA model.
The architecture suggested thus allows the robotic learner to learn new words by pairing information from previous distributional analyses, provided by, for instance, word order regularities in the target language, with current sensorimotor data. Thus, if a word is unknown, the correct meaning will be assigned to the word by the ERA module thanks to the ability of the RNN module to infer the semantic category on the basis of the distributional information. In particular, while Elman-style classifiers can group novel words together based on their distributional properties, only the connection to the sensorimotor data provide the robot with an understanding of the new word as a particular colour. The architecture proposed thus allows the robot to learn grammatical cues to semantic categories by combining sensorimotor grounded and distributional information. While this experiment applies to a rather restricted domain, and natural language is really much messier than the data used as input here suggest, the results still indicate that relatively simple mechanisms can account for abstract form meaning pairings, as required for grammar learning in general. Especially construction grammar (Goldberg 1995, 2006; Tomasello 2003) understand grammar to consist exclusively of such pairings, and the combination of experiential grounding and distributional learning combine two of the most crucial skills for language acquisition.

Figure 2 – Hierarchical cluster diagram of hidden unit activation vectors that constitute the internal representations of the robot as a result of the testing – the cluster analysis shows that the novel adjective blue is correctly represented as a colour term together with red and green. For the calculation, only true sentences were used. Capital letters stand for sentences included in
4 Learning abstract words: Grounding transfer from sensorimotor experience to abstract concepts

When infants start acquiring word meanings, they develop the ability to match the stream of perceptual-cognitive information (i.e. entities perceived through senses) to the stream of spoken language (i.e. sound associated to words). Studies conducted on children's early vocabulary acquisition have shown that the learning of concrete word meanings precedes the acquisition of abstract concepts (Gentner, 1982). While concrete terms (e.g. object's names) refer to tangible entities characterized from a direct mapping to perceptual-cognitive information, abstract words pertain to intangible entities that have weaker perceptual-cognitive constraints with the real world; that is, abstract words are linguistically more variable, given that they can refer to many events, situations and bodily states. Hence, during the process of word meanings acquisition, the mapping of perceptual-cognitive information related to concrete concepts into the linguistic domain occurs earlier than the mapping of perceptual-cognitive information related to abstract concepts.

Many scholars have suggested that the distinction between abstract and concrete words is a continuum according to which all entities can be varied in their level of abstractness (Wiemer-Hastings et al., 2001). As such to model the grounding and embodied bases of abstract word learning in robot, abstract action verbs, such us “to use”, “to make”, that represent a class of terms that describes actions with a general meaning (Wiemer-Hastings et al., 2001) have been used. Exploiting the hierarchical recursive structures observed both in language and the motor system (Cangelosi et al., 2010), an iCub model, that integrates simple motor primitives and concrete words in order to create the semantic referents of abstract action words that do not have a direct mapping to the perceptual world, has been developed. Indeed, in the proposed model the semantic referents of abstract action words are formed by recalling and reusing the sensorimotor knowledge directly grounded during the interaction of an agent in the real world (Stramandinoli 2012, Stramandinoli 2011). This is based on the mechanism of “symbol grounding transfer” (Cangelosi and Riga 2006).

The robotic task, following the “verb-argument structure” of the action-object frame as the basic component of human language (Arbib, 2002), consisted in training the iCub robot to learn a set of behaviors by acting with specific tools and acquiring the associated two-words sentences consisting of a verb and a noun describing the specific action performed on the selected object. Building on this premise, first the robot is trained to recognize a set of tools characterized by different colors, sizes and shapes (e.g. knife, hammer, brush) and to perform object related actions (e.g. respectively cut, hit, paint). Subsequently, the robot is taught to name these objects and actions (e.g. “cut with knife”). Finally, the robot is taught the abstract motor words of “use” and “make” by combining these new action words with their tool (e.g. “use knife”).
For the modeling of the mechanisms underlying motor and linguistic sequences processing in robots, partial recurrent neural networks (P-RNNs) were used as a neural controller for the iCub. A multi-modal 3-layer Jordan P-RNN (Jordan, 1986) was used to receive linguistic, visual and proprioceptive input modalities and to output words, motor responses and object representations. The visual and sensorimotor inputs have been recorded from the iCub sensors, while the linguistic inputs consists of binary vectors for which the “one-hot” encoding has been adopted, according to which each unit representing an individual word.

Robot experiments, carried out with the simulated model of the iCub, shown the ability of the robot to correctly understand and respond to the linguistic instruction using the abstract action words “use” and “make”. The experiments also investigate the effects of using different combination of the three input modalities (vision, language and proprioception). For example, the robot’s categorization of the perceptual, proprioceptive and linguistic inputs decreased in case the linguistic or visual inputs were not provided. Furthermore, incompatible condition tests have been performed; in case of inconsistency between the perceptual and linguistic input, simulation results have shown that the robot ignored the linguistic command by executing the actions elicited by the seen objects. These results are consistent with evidence in neuroscience and behavioral sciences that shows that visually perceived objects activate motor information (Jeannerod, 1994, Arbib, 1997). Hence, the knowledge associated to objects relies not only on objects perceptual features, but also on the actions that can be performed on them (i.e. affordances). Further simulation experiments suggest that the acquisition of concepts related to abstract action words requires the activation of similar internal representations activated during the acquisition of the concrete concepts contained in the linguistic sequences used for the grounding of abstract action words. This finding suggests that the semantic representation of abstract action words requires the recall and reuse of sensorimotor representational capabilities (i.e. embodied understanding of abstract language). Indeed, neurophysiological evidence of the modulation of the motor system during the comprehension of both concrete and abstract language exists to support this. For example, Glenberg et al. (2008) show that the processing of words, both concrete and abstract, involves the modulation of the motor system.

Future extension of this robotic model of abstract words will focus on the learning of action word meanings through the discovery of new affordances related to objects and the environment. Despite it is clear that language is grounded in sensorimotor experience, it is also evident the importance to go beyond simple direct sensorimotor grounding. This will be investigated with “hybrid models”, where some concepts can be directly grounded in a robot’s sensorimotor experience, while other concepts are acquired through statistical inference that will permit to go beyond the available data.
5. Learning to count: The role of pointing gestures and finger counting

Number cognition is another key example of the contribution of embodied cognition in the acquisition of abstract, symbol-like manipulation capabilities. Various embodied strategies, such as pointing and counting gestures, object touching, and finger counting, have been shown to facilitate the development of number cognition skills (e.g. Alibali & DiRusso, 1999; Moeller et al. 2011). The embodied basis of numbers is also shown in adults, as with the size, distance and SNARC effects (Spatial-Number Association of Response Codes; Dehaene, Bossini & Giraux, 1993). In this section we specifically look at two developmental robotics models of number embodiment, namely the role of counting gestures and of finger counting.

The contribution of the pointing gestures to the children’s learning of the capability to count is an interesting developmental phenomenon from the point of view of the embodiment of linguistic and symbolic knowledge, as it seems that it is through learning to count that children build a link between pre-verbal, approximate quantification skills and precise symbol manipulation capabilities (Le Corre & Carey, 2007). When learning to count, children spontaneously point to, touch, or move the objects, and there exist a large body of studies which show the beneficial effect of sensorimotor strategies on counting performance (see Graham, 1999, and Alibali & DiRusso, 1999, for reviews). There are three main groups of hypotheses about the role and the mechanism behind this phenomenon. First, gestures may help children overcome the limitations in available cognitive resources, for instance by helping to keep track of counted items. Second, they may perform a coordinative function by combining a temporal correspondence with speech and a spatial correspondence with the counted items in one bodily activity. Third, gestures may also facilitate social learning by providing the tutor with feedback about the child’s learning progress. By investigating the contribution of the counting gestures to learning to count with the use of developmental robotics we aimed at providing additional evidence for (or against) these hypotheses.

Our developmental robotics model of the contribution of the counting gestures to learning to count (Rucinski et al. 2012; Rucinski 2013) aimed at answering two questions: (1) Can counting gestures improve the counting accuracy if they are represented in the form of the values of arm joint angles that change over time; (2) Is the spatial correspondence between the items being enumerated and the indicating act performed during counting an important characteristic of the counting gestures? The first question is connected with providing evidence toward the usefulness of the counting gestures in learning to count beyond known behavioral studies. Answer to the second question would provide novel insight into the nature of the contribution.

Our robot experiment design was modeled after Alibali & DiRusso’s (1999) behavioral study of the role of counting gestures in children. The robot model employed in our experiments was based on an Elman simple recurrent network. The counting task was simulated as requiring the network to output a count list
(one-hot coding) corresponding to counting the objects shown in the visual input layer, optionally in the presence of counting gestures, in response to the trigger stimuli. Proprioceptive information was simulated based on counting gestures performed by the iCub robot (see figure 3). The robot’s neural network was trained and tested in several experimental conditions, such as counting with vision only, with natural counting gestures and with artificial rhythmic gestures. Research questions were addressed by comparing final counting performance, assessed in the same way as is done for children, across the experimental conditions.

The robot simulation experiments showed that supplying the network with proprioceptive information on the pointing gestures allowed it to significantly improve the counting accuracy, as compared with the condition of counting using only visual information. Furthermore, the improvement was not explained simply by the additional input signal, as the model also counted significantly worse if supplied only with the proprioceptive input. This provided first evidence outside of behavioral studies that counting gestures are a useful embodied cue in learning to count. In addition, contrasting the effects of natural spatio-temporal counting gestures with those of artificial rhythmic ones revealed that it is important that counting gestures are characterized by a spatial correspondence to the counted items – in the latter case the gestures did not facilitate the extraction of information by the neural network from the visual input.

Although our model reproduced the main effect of the counting gestures on counting accuracy, it fell short of exhibiting all behavioral effects reported by Alibali & DiRusso (1999). More specifically, our data did not fit to the experimental data quantitatively, we did not observe statistically significant effect of the set size, and the patterns of counting errors of our model were different from those of children. The most probable reasons were relative simplicity of the proposed model and discrete-time nature of the employed neural network framework. Furthermore, our model was so far tested only in the set-up with proprioceptive information as an input, whereas simulations in which model has to produce the correct sequence of gestures by itself would also be valid from the theoretical standpoint.
The direct link between finger counting and number learning is also an evidence of the role of embodied cognition in numbers and abstract symbol processing. Consistent neurocognitive and psychological data with children and adults show that finger counting strategies and finger-based representations play an important role in the development of numerical and arithmetical skills and in the learning of number words. Moreover, finger counting in particular and gesture- and action-based embodied strategies in general, have been shown to support more effective acquisition of number words (e.g., Alibali and Di Russo 1999) and to affect the teaching of mathematical concepts (e.g., Moeller et al. 2011).

The developmental robotics paradigm was used specifically to explore whether finger counting and the association of number words (or tags) to each fingers could serve to bootstrap the representation of number in a cognitive robot. This study uses a recurrent artificial neural network to model the learning of associations between (motor) finger counting, (visual) object counting and (auditory) number word and sequence learning (De La Cruz et al. 2014; Di Nuovo et al. 2014b). In particular this studies manipulates the coupling between different modalities, as with the comparison of the Auditory-Only condition, i.e. when the robot solely learns to hear and repeat the sequence of number words (“one”, “two”, ... up to “ten”), and of the Finger+Auditory condition, i.e. with the robot’s simultaneously learning of the sequence of acoustic number words and the sequence of moving fingers (the American sign language finger counting configuration was used to match the iCub robot’s finger actuator system).

The results obtained in various modeling experiments with both the simulated and the physical iCub robot show that learning the number word sequences together with finger sequencing helps the fast building of the initial representation of number in the robot. Robots who only learn the auditor sequences achieve worst performances. Moreover, the neural network’s internal representations for these two counting conditions result in qualitatively
different patterns of the similarity between numbers. Only after the
Finger+Auditory sequence learning the network represents the relative distance
between numbers, which corresponds to the quantitative difference between
numbers. In Finger+Auditory trained robots, the cluster analysis diagram of the
hidden layer’s activation shows that the representation for the number word
“one” is adjacent to that of “two” and is increasing more different (distant) from
the higher numbers. Instead, in the auditory-only condition, there is no
correspondence between the cluster diagram similarity distance and the
numerical distance.

Furthermore, the neural network’s internal representations of the finger
configurations, developed by the robot as a result of the experiments, sustain the
execution of basic arithmetic operations. In the Finger+Auditory condition,
number words heard repeatedly, when coupled to the experience of moving the
fingers, serve as tools, used in the subsequent manipulation of the quantities
they come to represent.

In fact, the internal representations of the finger configurations themselves,
found as a result of the experiments, can be considered to be a basis for the
building of an embodied number representation in the robot, something in line
with embodied and grounded cognition approaches to the study of mathematical
cognitive processes. Just as has been found with young children, through the use
of finger counting and verbal counting strategies, such a robotic model develops
finger and word representations that subsequently sustain the robot’s learning
the basic arithmetic operation of addition. To support this statement, in Figure 4
it is presented the dendrogram after the optimal leaf order that shows how the
internal finger representation is more similar to the number sequence, indeed,
numbers that are close in the actual sequence are linked together. Meanwhile,
the grouping of number words (learned in or out of sequence) is more random,
and affects the learning as shown in the classification experiment.

The use of such embodied developmental robotic models can also have
implications for research on the understanding of the role of motor strategies in
mathematical education. The utility of children’s learning of finger counting
strategies early in their mathematical education continues to be debated in
mathematics education research, despite the evidence coming from
neurocognitive and psychological studies indicating that it does (for review of
debate see Moeller et al., 2011). The robot experiments show that learning to
The robotics models and experiments presented in this chapter show the potential of the cognitive and developmental robotics approach to model a variety of phenomena linking embodiment and symbol manipulation skills. They range from the modeling of embodiment cues, as posture and space biases, in learning the names of objects, to the strong relationship between sensorimotor experience and representation and the learning of labels for action. Furthermore, the embodiment strategies that a robotic agent can use exploiting its own intrinsic sensorimotor nature, as using pointing gestures when counting objects or linking finger sequencing and counting with number sequences.

Most of the models presented above are closely based on empirical data on cognitive development. For example, the posture and word learning model is directly based on the set of child psychology experiments (modi experiments) conducted by Smith and collaborators to investigate the role of the body and space biases in early word acquisition. In such cases where specific robot experiments directly replicate child experiments it has also been possible to use the computational model to make detailed predictions, which have been subsequently validated in new child psychology studies (cf. Morse et al., submitted). Other models, as those on number learning, use a more loose approach to the modeling of developmental psychology data. For example, the work by Rucinski et al. and by Di Nuovo et al. take inspiration from general developmental evidence on the contribution of pointing gestures and finger counting, rather than modeling specific experiments in the literature. Even in these cases, the robot models can contribute to a better understanding of the relationship between embodiment and language and number cognition in children.

Such a direct link between developmental robotics models and child psychology experiments and data shows the fruitful scientific and technological benefits of a highly interdisciplinary approach to the understanding and modeling of cognition in natural cognitive agents and to the design and implementation of sensorimotor mechanisms in the development of language and symbol manipulation skills. This is in line with the growing trend to use “grounded” (i.e. grounded)...
The analysis of these developmental robotics models also offers a methodological consideration on the need and benefit of using a physical vs. simulated robot agents to model cognitive development phenomena. The above models include both experiments carried out with the physical iCub robot platform, as in the Morse word learning experiments, and experiments carried out solely on the simulated iCub, as in Rucinski’s pointing gesture experiments. In some cases, as in the De La Cruz and Di Nuovo’s model of finger counting, studies originally based on the simulator have later been extended to experiments with the physical robot. The choice of a robot platform versus a simulated robot agent might depend on a variety of constraints (Tikhanoff et al. 2011). One concerns the need, or not, to model detailed sensorimotor strategies in cognitive development. If the embodiment strategies investigate regard higher-level motor strategies, as with point gestures, a robot simulator might be enough to model how the production of gestures, and their proprioceptive feedback to the robot, supports the acquisition of number knowledge. If, on the other hand, embodiment mechanisms are hypothesized to depend on fine differences in motor strategies, then experiments with the physical robot platforms are needed, as they better permit the investigation of fine sensorimotor tactics and of the noisy and dynamic interaction between the robot and its physical environment. For example, Di Nuovo’s et al. (2014a) comparison of the simulated vs physical experiments on finger counting have shown that even if from the cognitive point of view we can derive the same conclusion stated above the comparison highlights that physical problems of the real platform weakens the final result, in terms of the likelihood of the number classification. But, on the other hand, the results with the real iCub platform are more in line with other studies present in the literature that do not show a strong influence of finger counting over number words on numerical development. Another constraint to consider when choosing between the physical versus the simulated robot is the intended primary scope of the robot models. Research aimed at the development of sensorimotor and cognitive capabilities in actual robotic agents require the use of the psychical platform for the controlled to be able to handle the noisy and variable aspects of the robot’s sensors and actuators (though initial pilot experiment in simulation can help explore some of the initial experimental parameter sets).

A further methodological and scientific consideration in the use of cognitive robotics models of embodied cognition is the need to consider open-ended, cumulative learning in cognitive development. The variety of experiments and models presented in this chapter, though it shows the potential of such an approach to model various embodied developmental mechanisms, at the same time has the limitation of considering such skills as separated phenomena. For example, both pointing gestures and finer counting have been shown to contribute to the development and bootstrap of number cognition. But in the two number models presented above, each embodiment strategy is studied in separated experiments and with a cognitive architecture based on different
types of neural network (though both share the use of recurrent network, given the need to handle time series and action sequences). At the same time, the objects and action naming models are treated as separated experiments, though they both share the ERA cognitive architecture. Future work in this field will then require the use of the same (expandable) cognitive architecture to be used to model the acquisition and control of a variety of cognitive skills and study the open-ended, cumulative aspects of development as cognitive bootstrapping phenomena based on the critical accumulation and integration of various modalities and skills.

Acknowledgement

The work presented in this chapter was funded by the EU FP7 ICT Integrating Projects "ITALK", the FP7 ICT projects “POETICON++” and "Robot-Era", the Marie Curie ITN “RobotDoc”, and the UK EPSRC project “BABEL”.

References


Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. Champaign, Ill.: University of Illinois at Urbana-Champaign, Center for the Study of Reading.


