A computational framework for the study of parallel development of manipulatory and linguistic skills

Ioana D. Marian **Aude Billard**
Autonomous Systems Laboratory
Swiss Institute of Technology Lausanne (EPFL)
1015 Lausanne, Switzerland
{ioana.marian;aude.billard}@epfl.ch

Abstract

This paper investigates the neural correlates underlying children cognitive ability to process sequences. This research stresses the developmental parallel between the appearance of complex manipulatory skills and the usage of two words compounds, and investigates the hypothesis that similar neural structures might be recruited for the processing of sequences in these two tasks.

We develop a model composed of a hierarchy of connectionist architectures, that accounts for low-level processing of sensory and motor information, the building of complex sequences of sensory-motor loops, to the construction of abstract symbolic (linguistic) knowledge. We validate the model against behavioral data, through an implementation of the *seriate nesting cup* scenario (Greenfield et al., 1972) in a dynamic simulation of a child-caretaker pair of humanoid robots. The seriate nesting cup experiment investigates the correlates between the development of the child's ability to serially handle objects, and the child's ability to form and understand sequences of words. We investigate the effect of varying a number of parameters of the neural model, as a way of accounting for the different capabilities of the child to manipulate objects serially. We present preliminary results. Ongoing work models grounding of basic linguistic capability, and investigate its role as a bootstrapping mechanisms for the learning of manipulation tasks.

1. Introduction

A developmental approach represents a fruitful paradigm for the construction of artificial systems capable of developing complex behaviors from simpler components (Zlatev and Balkenius, 2001). A considerable body of cognitive theories support the idea that higher level functions of the brain, such as cognition and language, exploit lower level functions, used more generally for processing perceptual information and driving motor systems (Piaget, 1970; Greenfield 1991). Moreover, such a *motor theory* of language evolution and development is also supported by neurophysiological data, that show the central role of cortical motor systems in the evolution and acquisition of different linguistic capabilities (Pulvermuller, 2002). We ground our research on these theories and investigate the hypothesis that neural structures, responsible for the processing of sequences, that develop for purely motor tasks, might be recruited for the development of language.

The ability to *seriate*, i.e. the capacity to relate an intermediate element both backwards to the previous element and forwards to the next element in the sequence, is an important topic of psychodevelopmental studies. Like imitation and language acquisition, the ability to seriate objects follows developmental stages that are well documented (Piaget, 1970). Work by Greenfield and colleagues developed the *seriate nesting cup* experiment (Greenfield et al., 1972) to investigate the correlates between the development of the child's ability to serially handle objects, and the child's ability to form and understand sequences of words. They report that children between 11 and 36 months of age exhibit different strategies, correlated to their developmental age, for combining cups of different sizes. Three distinct strategies were identified: (1) the pairing method, when a single cup is placed in/on a second cup; (2) the pot method, when two or more cups are placed in/on another cup; (3) subassembly method, when a previously constructed structure consisting of two or more cups is moved as a unit in/on another cup or cup structure. The child's choice of the acting/acted upon cups seems to be based on either one of these three criteria: size, proximity and contiguity. The youngest children seem to use the proximity criteria (i.e., same side of the table with the moving hand) for pairing cups. Children of 16-24-month-old seems to follow only the contiguity criteria (i.e., never reach behind a nearer cup to use a more distant cup), while the 28- to 36-moth-old seems to follow the size criteria.

Greenfield and colleagues stressed the homology between these three action strategies and specific grammatical constructions. When a cup "acts upon" another cup to form a new structure, there is a relation of *actor-action-acted upon*; such a relation is realized in sentence structures like *subject-verb-object*. The first action strategy would, thus, correspond to the use of simple two-words sentence. The second and third strategies, on the other hand, allow the formation of multiple actor-action-acted upon sequences, and, as such, would correspond to the usage of more complex sentences. The difference is that in the second stage the child performs a *conjunction* of the sequences/words, while in the last stage, the embedding of the cups is accomplished, reflected into the capacity of using *relative clauses* in the language.

The seriate cups experiment is of relevance to the theory of goal-directed imitation (Bekkering et al., 2000). While watching the demonstration, the children form an representation of the *goal* of the task. The seriate cups experiment highlights three goal-directed strategies based on different metrics (the criteria listed above).

In this research, we investigate the parallel between the development of the child's ability to process sequences of manipulatory action and the child's ability to process sequences of words. In this paper, we present preliminary steps towards the modeling of the seriate nesting cup scenario. This works follows a general approach to understand the principles behind imitation learning, and, in particular to study the strategies required to discover the goal, i.e. *what to imitate*, and the metric from a demonstrated task (Billard et al., 2003). It builds on previous work of ours that investigated the role of imitation in learning simple sequences of movements (Billard and Hayes, 1999) and the role of imitation in the emergence of synthetic proto-language in an autonomous robot (Billard, 2002).

2. Related modeling work

Learning to seriate nesting cups requires the ability to represent time. The common need of the architectures aimed at processing temporal sequences is the presence of a short-term memory and of a prediction mechanism. Recurrent networks can in principle implement short-term memory using the feedback connections. Temporal sequences are learned as set of associations between consecutive components, and recall of the next elements in the sequence is possible based on the previous components. Different temporal variants of connectionist networks can be constructed by locally modifying the architecture to keep a trace of history (i.e., partial recurrent networks), reconfiguring the network parameters to accommodate enough temporal information (i.e., time delay networks) or by designing special architectures (see Chappelier et al., 2001 for a review).

To deal with temporal temporal dependencies beyond consecutive components different solutions have been developed. Early attempts to deal with long time lags, were based on using time constants to influence changes of unit activations. Recent models propose different ways to enhance the memory capacity at the neural level, by carefully designing the basic computational unit (Hochreiter and Schmidhuber, 1997). Our previous work on robot learning (Billard and Hayes, 1999) used a time-delay associative network, consisting of a Willshaw network, to which self-recurrent connections have been added and a capacity to encode (in the weights) the time and frequency of nodes activation.

Directly relevant to the developmental correlation hypothesis, are the computer simulations of Reilly (1997), who demonstrated a computational advantage in building a language production capability upon a motor-planning foundation. He proposed the concepts of *cortical software re-use* and *asymmetric collaboration* to explain how language processing and cognition can be built upon sensory-motor programs. Related work have been carried by Dienes et al. (1999) on the transfer of structural knowledge between different domains.

A number of attempts have also been made towards world-to-word mapping. Regier' model (1995) on spatial knowledge learning was built based on a tripartite trajectory representation of type *source-path-destination*, aimed at grasping the event logic in both motion and language. Siskind (1995) proposed the use of visual primitives which encode notions of support, contact and attachment to ground the semantics of events for verb learning. Event logic has been recently applied by Dominey (2003) for learning grammatical constructions in a miniature language from narrated video events, and by Billard et al. (2003) to the learning and reproduction of a manipulation task by a humanoid robot.

Our work complements these different efforts, by bringing forward the importance of imitation in the building of manipulatory and linguistic abilities.

3. The computational framework

Ongoing work presented here is aimed at investigating the effect of varying a number of parameters of the neural model, as a way of accounting for the different capabilities in the child. In particular, we consider the effects of joint attention, mnezic capacities, and development of object concept on the child/robot's reduced ability of composing manipulation and linguistic steps.

There is a growing body of evidence that extended periods of adult-child attentional focus on nonlinguistic entities, scaffold the child's early language development (Tomasello, 1988). When the infant hears an utterance and perceives a visual scene, he/she has to discover to what aspect of the scene the sentence is referring. While joint attention solely cannot solve this problem (i.e., cross-situational learning is necessary) it makes it possible and computationally tractable. One of

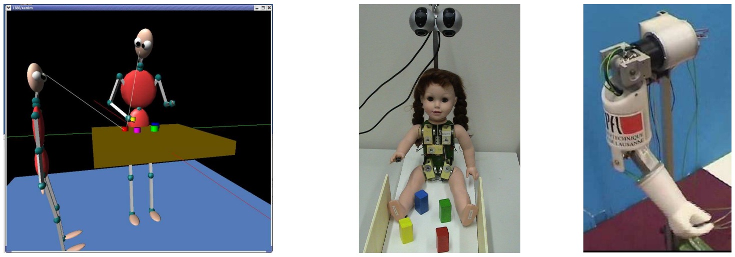

Figure 1: (a) The Xanim dynamic simulation of a Child-Caretaker pair of humanoid robots. The Caretaker demonstrates the seriate cup task. The Child and Caretaker shared focus of attention is highlighted by the crossing of their gaze directions. (b) The experimental setting for the seriate cups task with Robota (under construction). (c) Robota's 5 DOF arm.

the assumptions of our system is that a powerful joint attentional mechanism acts as an information-processing bottleneck for the incoming sensory information.

By the end of the first year of life, infants show delayed imitation capacities and long-term memory for serial order (Bauer, 2002). However, it was suggested that the capability of infants of this age to remember serially organized sequences depends on being primed with the preceding component. Correspondingly, their capacity would be due to a paired-associate learning, rather than to learning of distant temporal relations. We consider that there is no evidence to attribute the deficit in seriation solely to a mnezic deficit, and rather, we investigate possible limitations of the learning employed. Learning it is also a function of the way external information is represented. Carey and Xu (2001) brought evidence that children up to 10 months of age cannot draw on featural properties to establish objects identity. Rather, they rely on spatiotemporal information (i.e., location) to build distinct representations of the objects.

As concerns the neural and architectural biases of the model, these are arising from the sub-symbolic paradigm: temporal overlapping of neural states during interaction with the environment, graded activation of simultaneously active representations, and incomplete reduction of information during generalization (Dorffner, 1992). We draw on both computational and neurobiological data indicating that symbolic knowledge in the brain is implemented through the distributed activation of sub-symbolic features (Pulvermuller, 2002). The big promise of a distributed approach is the integration of learning and representation. We implement an abstract notion of a *cell assembly* as a basic computational unit for both the encoding of information (i.e., a collection of features) and learning of temporal sequences.

4. The simulation environment

The current implementation of the model was conducted in the Xanim dynamic simulator (Schaal, 2000), to model a pair 30 degrees of freedom (Head: 3, Arms 7 *2, Trunk 3, Legs 3*2, Eyes 4 D.O.F.) humanoid robots, (see Figure 1 left). The external force applied to each joint is gravity. Balance is handled by supporting the hips; ground contact is not modeled. There is no collision avoidance module. The dynamics model is derived from the Newton-Euler formulation of Rigid Body Dynamics. The simulation package SL has a modular structure that includes a motor servo, used to read the current state of the robot/simulation and to send commands to the robot/simulation; a task servo that allows switching between different tasks; a vision servo to collect data from camera systems; and inverse dynamics and inverse kinematics servos to allow the control of the robot from Cartesian states. In the next months, the seriate nesting experiment will be implemented in the mini-humanoid Robota (Billard, 1999), see Figure 1 right.

5. The model

The model consists of a hierarchy of connectionist architectures, whose basic computational unit is depicted in Figure 2. The first two layers of the network implement a mechanism for object recognition and an attentional mechanism, respectively. At the last layers a formal implementation of a cell assembly is used for information representation and temporal learning. The layered architecture was developed to account for temporal integration at different time scale levels. The reduction of the hierarchy to three levels, results in a tight coupling between the conceptual/structural layer (i.e., cell assemblies) and the perception layer. The output sent to the motor system is represented by the coordinates of the target object, which are transformed by Xanim servos into commands to the simulated robot links.

Let us describe how the system functions. At the first parsing of the visual scene, the agent creates a list with the objects present in the image at their initial locations. The object recognition (OR) subnetworks from all locations are activated and remain active as long as the

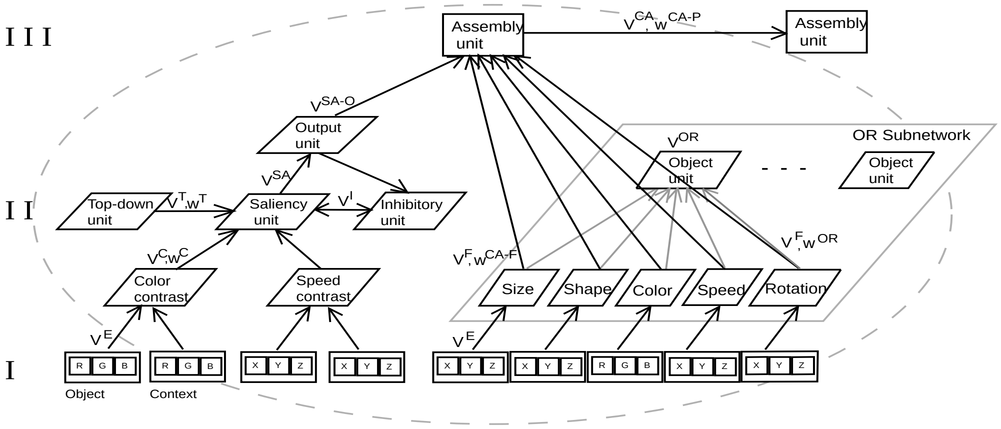

Figure 2: Basic computational unit of the hierarchical network consisting of three interconnected components: a saliency signal, an object recognition subnetwork and a cell assembly unit. The saliency signal is computed by integrating the feature contrast with top-down cues. The activity of a winning salient unit is modulated by inhibition of return. An object recognition unit is feeded from 5 feature units. A subnetwork is formed from all co-located objects. Both attentional and recognition components are taking input from an external layer. The cell assembly unit receives a saliency signal and it is grounded in the visual feature layer.

objects are in the visual field (i.e., objects can become invisible if they are hidden/embedded by other objects). Saliency (SA) is computed in a distributive manner at all locations of the visual scene and the unit with the highest activation wins the focus of attention. A winning salient unit enhances the object representation and allows the creation of a 'copy' of the information from that location in space. This copy is referred to as a *cell assembly* (CA) and it is created whenever a significant variation (i.e., event) in one sensor has been detected. At the creation, a cell assembly unit inherits all the features currently active at that location.

Learning of event sequences is implemented by mapping each new event-processing state of the system into a distinct CA. This represents a method to unfold temporal relations into spatial structures, with the advantage of preserving the representation of information. Knowledge is represented in the weights of the CA features and learning at this level can aim for variance reduction and generalization. Learning of precedence relations takes place in the weights of all simultaneously activated CAs.

5.1 Attentional module

The attentional module consists of a mixed bottom-up and top-down neural network model. Studies in psychophysics suggested that the contrast of the features with respect to the contextual surround, rather than the absolute values of the features, drive the bottom-up attention (Nothdurft, 2000). We compute bottom-up saliency based on the processing of the contrast of two low-level features: color and motion. The focus of attention is deployed to the most salient location in the scene, which is detected using a winner-take-all strategy. Once the most salient location is focused, the system uses a mechanism of *inhibition of return* to inhibit the attended location and to allow the network to shift to the next most salient object (Itti and Koch, 2001).

The activity of the saliency map is further modulated by top-down cues, which focus the attention in accordance with the learned significance of each cue (i.e., to follow the gaze of the demonstrator or to look in the direction of the hand pointing). The current version of the attentional module includes as top-down cues: gaze following and skin color preference.

5.1.1 External units

Each *feature contrast unit* V^F receives input from two pairs of three *external units* V^E, that encode either the color (R,G,B) or 3D location (X,Y,Z) of an object. Color contrast is computed using one value of the sensor at the object location and $N = 6$ values corresponding to 6 contact points with the surrounding context. The motion contrast is computed using one reading of the speed for the object and N context readings corresponding to the speeds of all objects in the visual image.

The activation of the external units corresponding to the component j read at location l are given by:

$$V^{\mathrm{E}}_{j,l}(t) = \begin{cases} n_{j,l}, & \text{if location l is visible at t,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$$V^{\mathrm{E}}_{j,l_{context}}(t) = \frac{1}{N}\sum_{k=1}^{N} n^{k}_{j,l}, \tag{2}$$

where t is the time and $n_{j,l}$ is a normalized value between $[0, 1]$ of the value returned by the sensors.

5.1.2 Contrast units

The activity of a feature contrast unit is given by the euclidean distance between the components of the feature corresponding to the object and to the context surrounding the object:

$$V_i^{\mathrm{C}}(t) = \mathcal{F}\Big(\sqrt{\sum_{j=1}^{3}(V_j^{\mathrm{E}}(t) - V_{j_{context}}^{\mathrm{E}}(t))^2}\Big) \tag{3}$$

where $\mathcal{F}$ is the sigmoid function $\mathcal{F}(x) = 1/(1+e^{-x})$. For simplification of notation, we consider that the activity of each unit corresponds to a location l, which is not shown in the equation.

5.1.3 Saliency units

The set of bottom-up features is $C = \{\text{color-contrast}, \text{motion-contrast}\}$. The saliency of a certain location is also modulated by top-down processes. We consider only one top-down cue $T = \{\text{gaze-following}\}$. Skin preference is integrated at the level of color contrast processing.

The activity of a saliency unit i is given by the weighted summation of the features $j \in C$ and the contribution of the top-down cue $k \in T$:

$$V_i^{\mathrm{SA}}(t) = \mathcal{F}\Big(\sum_{k=1}^{|T|} w_k^{\mathrm{T}} \cdot V_k^{\mathrm{T}}(t) + \sum_{j=1}^{|C|} w_j^{\mathrm{C}} \cdot V_j^{\mathrm{C}}(t)\Big) \tag{4}$$

where the activity of the top-down unit is given by:

$$V_k^{\mathrm{T}}(t) = \begin{cases} 1, \text{ if object i is referred by the top-down cue k,} \\ 0, \text{ otherwise.} \end{cases} \tag{5}$$

Weighting of the bottom-up features w^{C} and top-down cues w^{T} results from the satisfaction of a number of constraints, as described in Section 5.2.

All units in the saliency map compete according to a winners-take-all strategy and the winning unit i sets the activity of its output unit to 1:

$$V_i^{\text{SA-O}}(t) = \begin{cases} 1, \text{ if } V_i^{\mathrm{SA}}(t) > V_j^{\mathrm{SA}}(t), \forall j \neq i \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

The activation of an output unit is sent both forward to the cell assembly layer and backward to the saliency map where it triggers the activation of the inhibitory unit.

The activation of the inhibitory unit is a function of the input from the salient unit and its previous memory:

$$V_i^{\mathrm{I}}(t) = \mathcal{F}\left(V_i^{\mathrm{SA}} \cdot \mathcal{H}(t - t_0) + \tau_{ii} \cdot V_i^{\mathrm{I}}(t-1)\right) \tag{7}$$

where τ_{ii} is the time decay rate of unit i, and $\mathcal{H}$ is a function given by the equation:

$$\mathcal{H}(t - t_0) = \begin{cases} 1, \text{ if } t = t_0 \\ 0, \text{otherwise.} \end{cases} \tag{8}$$

V_i^{SA} is the activity of the winner unit and t_0 represents the time when the output of the inhibition unit is triggered, and it is a function of the saliency unit activation $t_0 = f(V_i^{\mathrm{SA}})$. When the activity of the inhibitory unit reaches its maximum it shunts down the salient unit. After the inhibition unit is activated, the activity of a salient unit will be given by:

$$V_i^{\mathrm{SA}}(t) = \mathcal{F}\Big(\sum_{k=1}^{|T|} w_k^{\mathrm{T}} \cdot V_k^{\mathrm{T}} + \sum_{j=1}^{|C|} w_j^{\mathrm{C}} \cdot V_j^{\mathrm{C}}(t) - V_i^{\mathrm{I}}(t)\Big) \tag{9}$$

The larger the value of the unit saliency, the longer it will stay active, but also the higher will be its inhibition. After shutting down the salient unit, the inhibitory unit preserves a memory of its activation, which decays in time and allows the unit to win again further in future.

5.2 Setting up the attentional constraints

A number of 6 constraints were defined to describe the internal model of the imitator on the significance of the top-down cues relative to the bottom-up features. The constraints are:

1. **Skin color preference**. For any static scene, the bottom-up saliency (equation 4) of an end-effector (i.e., hand) should be higher than that of any object.
2. **Preference for moving stimuli**. For any moving object its bottom-up saliency should be higher than that of any static object, including the end-effectors.
3. **Motion versus skin color preference**. Saliency of a moving object should be higher than that of an end-effector moving at a slower speed, but smaller the the end-effector saliency moving at a comparable speed.
4. **Gaze following versus moving objects**. The global saliency of any static object located in the focus of attention should be higher than the bottom-up saliency of any moving object located outside from the focus.
5. **Gaze following versus skin color**. The global saliency of any static object located in the focus of attention should be higher than the bottom-up saliency of any static end-effector located outside from the focus.
6. **Gaze following versus moving end-effectors**. The bottom-up saliency of a moving end-effector should be higher than the global saliency of any static object placed in the focus of attention, but smaller than the saliency of an object moving in the focus of attention.

The weights w^{C} and w^{T} were set after solving the system of equations given by the 6 constraints. Figure 3 illustrates the functioning of the attentional mechanism. In Figure 3a are shown the evolution in time of the saliency

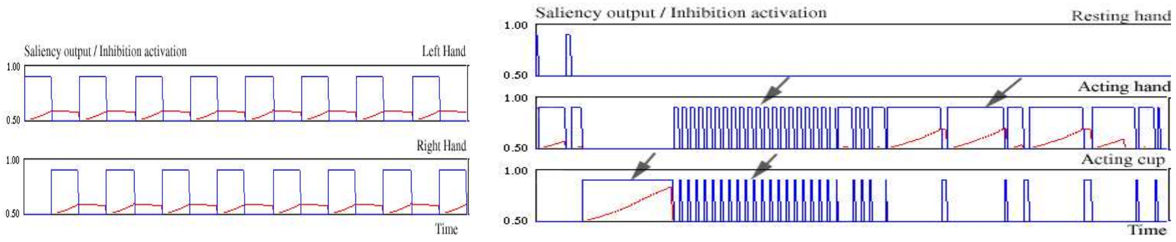

Figure 3: The time evolution of saliency output ($SA\!-\!O$) vs. inhibition activation. (a) The shift of focus between two competing end-effectors is shown. Note that inhibition shunts down the saliency unit (SA) (not shown in figure). (b) Competition between the resting hand, acting hand and acting object is shown, in the presence of the gaze signal (indicated by the arrows).

output vs. inhibition activation of two end-effectors that compete for the focus of attention. Figure 3b illustrates the shift of the focus in the presence of the demonstrator gaze signal. When gazed, the acting cup wins and maintains the focus of attention for a long period. Different locations can be attended to, due to the inhibition mechanism. As a result, a powerful selective mechanism implements an information-processing bottleneck of what is learned at the higher levels of the system.

5.3 Cell assembly module

The cell assembly module consists of the OR network and the CA layer (see Figure 2 right). The object recognition module accounts for visual awareness of an object, based on the automatical processing of five features in $F = \{\mathrm{color}, \mathrm{shape}, \mathrm{size}, \mathrm{rotation}, \mathrm{motion}\}$. Information processing at the CA level is grounded in the accurate representation of the world state constructed here. The activation of the external units is given by Equation 1.

5.3.1 Feature units

The activation of a feature unit is given by:

$$V_i^{\mathrm{F}}(t) = \mathcal{F}\Big(\sum_{j=1}^{3}(V_j^{\mathrm{E}}(t))^2\Big) \tag{10}$$

For the detection of an event, the signal variation ΔV_i^{F} is integrated over time and compared with a positive, arbitrary set threshold Θ_i^{F}. A feature is referred to as being *active* if at least one event has been detected.

5.3.2 Object recognition units

The activity of an object recognition unit i is given by the weighted summation of the features j:

$$V_i^{\mathrm{OR}}(t) = \mathcal{F}\Big(\sum_{j=1}^{|F|} w_{ji}^{\mathrm{OR}} \cdot V_j^{\mathrm{F}}(t)\Big) \tag{11}$$

The strengths of the weights w_{ji}^{OR} is subject of adaptation according to an Anti-Hebbian learning rule:

$$\Delta w_{ji}^{\mathrm{OR}}(t) = \eta \cdot \big(V_j^{\mathrm{F}}(t)) - w_{ji}^{\mathrm{OR}}(t-1)\big) \tag{12}$$

where η is the learning rate. The OR units are used during associative learning between word form and semantic meaning, to ground referrents for individual objects.

5.3.3 Cell assembly units

A cell assembly unit i is constructed whenever a new event has been detected by the OR subnetwork, and it is defined by a set F^{CA} of features simultaneously active at some location in space: $F_i^{\mathrm{CA}} = \{j \in F \,|\, \int_0^t \Delta V_j^{\mathrm{F}} \geq \Theta_j^{\mathrm{F}}\}$. In what follows we will refer to the cell assembly' features, as its feature constraints.

Once it is created, a cell assembly can be in one of the following states: satisfied, maximally satisfied, activated and silent. Satisfaction is a graded measure computed in rapport with the current state of the external world (CSW), reflected in the weights of the OR network. A cell assembly is satisfied, if its feature constraints are satisfied by the CSW above a certain threshold arbitrary set for all the CAs. A maximally satisfied CA is the CA whose feature constraints satisfy best the CSW. A CA is active if its activation is above an arbitrary set threshold. Otherwise it is considered silent.

An activated CA learns a general representation of the objects present at that location. At the creation of the CA_i, the weights $w_{ji}^{\text{CA-F}}$ are initialized to the values of the weights of the object recognition network. During the demonstration of the task, the weights are subject of Anti-Hebbian learning, meant at preserving only the invariant features:

$$\begin{aligned}\Delta w_{ji}^{\text{CA-F}}(t) = {} & \eta \cdot V_i^{\mathrm{CA}}(t)\Big((1-\Delta V_j^{\mathrm{F}}(t))(V_j^{\mathrm{F}}(t) - w_{ji}^{\mathrm{CA}}(t)) \\ & + V_j^{\mathrm{F}}(t)(-w_{ji}^{\mathrm{CA}}(t))\Big)\end{aligned} \tag{13}$$

where η is the learning rate.

In the current version of the model, constraint satisfaction is implemented by learning for each CA a feature threshold using the formula: $\Theta_i^{\text{CA-F}}(t) = \frac{1}{n+1}\Big(\Theta_i^{\text{CA-F}}(t-1) \cdot n + \sum_{j \in F_i^{\text{CA}}} w_{ji}^{\text{CA}} \cdot V_j^{\text{F}}(t)\Big)$, where n represents the current time step. A disadvantage of this implementation is that different combinations of features can satisfy the threshold. Ongoing work is aimed at developing a more efficient framework for general resolving of multiple constraints satisfaction.

The activity of a cell assembly is a function of the global saliency V_i^{SA-O} of the location for which it was build, the level of feature constraint satisfaction, and the memory of its previous activation:

$$V_i^{\text{CA}}(t) = \mathcal{F}\Big(S_i^{\text{CA-F}}(t) + \tau_{ii} \cdot V_i^{CA}(t-1)\Big) \qquad (14)$$

Satisfaction of the feature constraints is given by:

$$S_i^{\text{CA-F}}(t) = \mathcal{H}(\frac{A}{B}, \theta_s) \cdot V_i^{SA-O}(t), \qquad (15)$$

where

$$\frac{A}{B} = \frac{\min\Big(\Theta_i^{\text{CA-F}}(t), \sum_{j \in F_i^{\text{OR}}} w_{ji}^{\text{OR}} \cdot V_j^{\text{F}}(t)\Big)}{\max\Big(\Theta_i^{\text{CA-F}}(t), \sum_{j \in F_i^{\text{OR}}} w_{ji}^{\text{OR}} \cdot V_j^{\text{F}}(t)\Big)}$$

and θ_s is an arbitrary set threshold. $\mathcal{H}(x,y)$ is a Heaviside function that outputs x, if x is greater than y, and 0 otherwise. A cell assembly i is satisfied if $S_i^{\text{CA-F}} > 0$ and becomes unsatisfied, either if $A/B < \theta_s$ or if $V_i^{SA-O} = 0$.

The satisfaction degree is a positive, symmetric measure of the distance between the CA learned bias (i.e., what it knows) and the current state of the world. F_i^{OR} represents the set of currently active features of an OR network (defined as in equation 15), and it is subject of adaptation with every change in the CSW. By contrast, the set of active features F_i^{CA} of a created CA remains constant, and with the creation of any new CA, early CAs tend to satisfy in a smaller manner the CSW. This is the case only, until the system returns to the initial conditions, what in our system happens due to the recursive nature of the nesting cups task. The activation of a CA that left the focus of attention decreases as a function of the decay rate τ_{ii}.

5.4 Learning of precedence relationships

A satisfied CA can learn a set of precedence links $w_{ji}^{\text{CA-P}}$ from other activated CAs. Precedence in the system is encoded in the relative order between the construction of the CAs. To deal with time dependencies larger than the time decay rate, we have introduced the graded measure of CSW satisfaction. The CA activation is a function of the time lag between the moment when the CA best satisfied the state of the external world and the CSW.

Learning of the precedence links is defined as a function of the rapport between the activities of two CAs. For a postsynaptic CA_i, a subunitary rapport V_j^{CA}/V_i^{CA} encodes the temporal precedence of CA_j and the distance between the moments when the two CAs maximally satisfied the state of the world. Adaptation of the weight from CA_j to CA_i takes place by:

$$\Delta w_{ji}^{\text{CA-P}}(t) = \begin{cases} \eta \cdot V_j^{\text{CA}} \cdot \big(r_{ji} - w_{ji}^{\text{CA-P}}(t-1)\big)\ , & \text{if } r_{ji} < 1, \\ \eta \cdot V_j^{\text{CA}} \cdot \big(-w_{ji}^{\text{CA-P}}(t-1)\big)\ , & \text{otherwise,} \end{cases}$$

where r_{ji} equals the rapport of CAs activities $V_j^{\text{CA}}/V_i^{\text{CA}}$. The learning rule favors strengthening of weights due to systematic causal relationships between assemblies. A CA who systematically precedes another cell assembly, will become a reliable predecessor and the strength of the connection will reflect the temporal lag between the CAs satisfaction. On the other hand, a large fluctuation of the order in which CAs are satisfied, will decrease the weight, reflecting the fact that there is no systematic causality between two CAs. The gradual adjustment of the weight is ensured by the small value of the learning rate (i.e., $\eta = 0.001$), which also avoids the formation of strong connections for accidentally perceived events.

5.4.1 Retrieval of the time structure

The activation of a CA during retrieval depends on the satisfaction of feature and precedence constraints. If the CA_i feature constraints satisfy the CSW, than its activity is given by Equation 15 and it is marked as being satisfied. If this is not the case (i.e., the feature constraints do not satisfy CSW, but the CA_i is active), CA_i becomes a sub-goal of the system and it can be in two states: achievable or postponed. A CA is achievable either if there is a minimal distance between the CSW and the satisfaction of its feature constraints, or if its precedence constraints are met. The minimal dissimilarity between the CSW and a CA state is met if the CA can be satisfied by acting upon only one dimension (i.e., motion or rotation). A CA is postponed if its precedence constraints are not met.

The activity of a CA during retrieval is given by:

$$V_i^{\text{CA}}(t) = \mathcal{F}\Big(S_i^{\text{CA-F}}(t) + S_i^{\text{CA-P}}(t) + \tau_{ii} \cdot V_i^{CA}(t-1)\Big) \quad (16)$$

where $S_i^{\text{CA-P}}(t)$ represents the level of satisfaction of precedence constraints. P_i^{CA} represents the subset of the precedence links whose weight is higher than an arbitrary threshold: $P_i^{CA} = \{w_{ji}^{\text{CA-P}} | w_{ji}^{\text{CA-P}} > \theta_w\}$ Satisfaction of the precedence constraints is defined in a similar manner with the satisfaction of feature constraints:

$$S_i^{\text{CA-P}}(t) = \mathcal{H}(\frac{A}{B}, \theta_p) \cdot V_i^{SA-O}(t), \qquad (17)$$

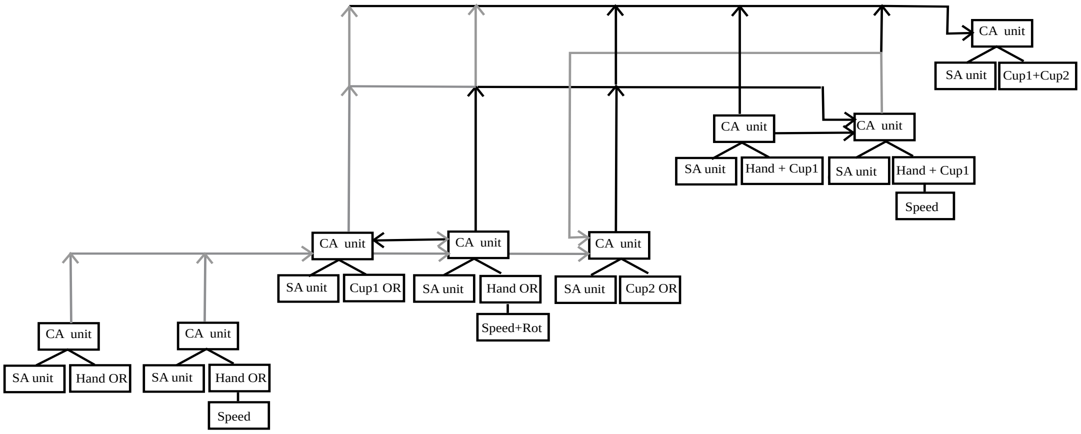

Figure 4: The CAs structure resulted after the embedding of two cups. CAs are shown from left to right in the order of construction. The strength of precedence weights is indicated by the darkness of line.

where

$$\frac{A}{B} = \frac{\min\left(\Theta_i^{\text{CA-P}}(t)\cdot I, \sum_{j\in P_i^{\text{CA}}} w_{ji}^{\text{L}}\cdot V_j^{\text{CA}}(t)\cdot S_j^{\text{CA-F}}\right)}{\max\left(\Theta_i^{\text{CA-P}}(t)\cdot I, \sum_{j\in P_i^{\text{CA}}} w_{ji}^{\text{L}}\cdot V_j^{\text{CA}}(t)\cdot S_j^{\text{CA-F}}\right)}$$

and θ_p and $\mathcal{H}(x,y)$ as above. The threshold $\Theta_i^{\text{CA-P}}$ was computed as the arithmetic mean value of all products $w_{ji}^{\text{CA-P}}\cdot V_j^{\text{CA}}$ corresponding to a satisfied CA_i. $I\in[0,1]$ is defined as the smallest percentage of the precedence constraints necessary to pass the threshold condition. Precedence is met by ensuring that the threshold $\Theta_i^{\text{CA-P}}(t)\cdot I$ is reached only by summing up the inputs in the order they have been learned (i.e., high activations associated with strong weights). The condition that only a satisfied cell assembly CA_j can activate a successor $S_j^{\text{CA-F}}>0$ is imposed to force the system to act externally (as opposed to the internal simulation) towards minimizing the distance between its goals and CSW.

An achievable CA_i can become satisfied, by triggering two types of actions: (1) send a motor command to modify the CSW (i.e., if it is immediately achievable); (2) *call-back* its predecessors, to increase their probability of being satisfied. The call-back is executed by setting the top-down saliency of the called CA to 1 and it is necessary to ensure that all predecessors are satisfied before the goal satisfaction.

6. Results and Discussion

6.1 Analysis of the system behavior

In the experiments presented here, imitation is bottom-up driven by the saliency of the objects/links in the visual image (see the forward model in Section 7). The agent starts by gazing the end-effectors and shifting between these and the objects in the image. As generalization is still limited (see discussion in Section 6.2), the CAs are activated only by the learned objects. In Figure 4 is depicted the CA structure resulted after two cups are nested. Activation enters the structure through the satisfaction of the CAs corresponding to static objects and end-effectors (bottom side of figure). We refer to these CAs as visible, in contrast to hidden CAs, which are not directly reachable (i.e., satisfiable) from the initial state. Each of the visible CAs is attached to a set of successors, to which activation is distributed in parallel. A hidden CA cumulates the activity received from its predecessors and checks for the satisfaction of precedence constraints. If this happens, the CA becomes achievable (i.e., a goal) and it triggers the motor commands necessary to bring the CSW into a state where its satisfaction is possible.

Figure 5 shows the actions taken by the system for the completion of a sequence. Precedence constraints are first met for the hidden CA corresponding to the shaping (i.e., rotation) of the hand to grasp the cup. Next CA achievable is that for carrying the cup 1 towards cup 2, who calls all its predecessors (see Figure 4), to ensure, for instance, that grasping the cup will be performed before carrying it. In parallel with the activation these goals, there are short activations of the final state CA, corresponding to the image of cup 1 into cup 2. The system exhibits both sequential (i.e., carrying is activated only after the hand-shaping is satisfied) and parallel activation of the goals (i.e., carrying with embedding) as a function of the level of the satisfaction threshold for the precedence constraints.

For any action to be taken, the system must send to the motor servo the coordinates of the target and recipient objects, which at any moment must be unique. A coherent behavior can be acquired only if several external constraints are applied: (1) *first come-first served*, that is, the goal first activated sets its target cup; (2) *continuity* of action, that is, once a cup become object of action,

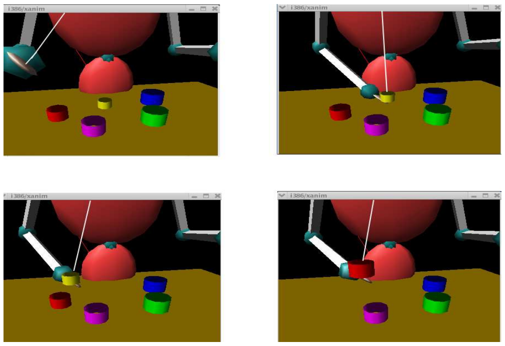

Figure 5: Imitation of a two steps embedding task. The actions taken by the agent are, in order: shape the hand, grasp cup 1, carry cup 1 to cup 2; shape the hand, grasp cup 2, carry 2 to cup 3.

motion is continued until the current goals satisfaction. Parallel vs. sequential activation of the goals represents, in our view, an appealing feature for the modeling of cognitive and linguistic processes.

6.2 Generalization capacity

Generalization is currently limited by the fact that a distinct CA is created corresponding to each location and event. Learning at the cell assembly level faces a generalization vs. seriation problem. Strict sequencing requires the creation of distinct CAs for each new system state. In order to generalize, however, a CA must be able to perceive different objects and events. A possible solution is to allow the development of higher CAs levels, where the object position can be discarded and generalization over the properties of several objects can occur.

Another solution can emerge exactly from what we thought to be a limitation, that is, the fact that each CA perceives only co-located objects. At the first parsing of the visual scene, the system makes a distinction assumption, that is, to build a distinct representation for each object. Whenever two objects come into contact, a new representation (CA) of the event is constructed, grounded in the recognition network of both objects. Presume that the hand grasps cup 1 and carry it to the position A, then grasps cup 2 and carry it to position B. Despite the different trajectories, the CAs corresponding to the manipulation of the objects, share the OR network of the hand and they are activated, at certain levels, whenever the hand is focused. Each of them will be able to perceive and learn, through the mediation of hand representation, the features of the other object, and develop a general representation of the type 'hand carrying any cup'. Because the level of activity of a CA is a function of the feature constraints satisfaction, it will be activated, and generalize only over sets of features which are not too dissimilar to what it knows already (i.e., any cup, bot not a cube). This behavior may be described as 'better perception through action' (Metta and Fitzpatrick, 2002), because the agent can learn more about the environment by acting upon it.

For the simulation of the seriate cups task, attention must be paid to the acquisition of the size concept. By implementing Anti-Hebbian learning the system discards variant features, and the agent will infer that size is not a significant parameter for the task. However, by assigning an increased saliency to the target states (i.e., nested cups), the agent can compare the objects, and learn that the difference in their affords their embedding.

7. Accounting for the developmental differences in seriating strategies

The differences in the infants' nesting behavior can be attributed to internal deficits at: encoding, retrieval or both. At encoding, we considered only one limitation, that is, an embedded object disappears from the view and its memory is only preserved in the activity of previously created CAs. The generalization capacity of the system plays an important role during encoding. Our model suggests, so far, that the neural structures grounded in the hand representation and involved in manipulation, provide the substrate for an over-generalization and lead to the formation of an enhanced representation of the type 'hand moves any cup'. This general representation would be responsible for the infants generalized behavior 'put cups into each other'.

For the retrieval of the sequence, we propose three hypothetic models:

1. The **forward model** considers that retrieval takes place through the forward, bottom-up activation of the hierarchical structure. Activity is driven in by the saliency mechanism and actions are chosen as a function of the saliency, as well as of other constraints (i.e., contiguity, coherence of action, principle of least effort). Goals are satisfied in the order they become activated. Initial priming it is also possible. This model is put to test to account for the first developmental stage of infants seriating strategies.

2. The **call-back model** considers that retrieval takes place by activating one or two goals and driving the system towards their completion. The difference to the previous model, is that the system starts from a hidden state, rather than from a visible one, and attempts the satisfaction of the activated goal, by calling its predecessors. The behavioral progress is due to the capacity during learning to focus on the target states of the task (i.e., the embeddings) and to store them in the long-term memory. After one goal is achieved, another one can be satisfied. The system will exhibit a seriating capability corresponding to the second developmental stage.

3. The **multiple constraints satisfaction model** considers retrieval as a task of reproducing the entire se-

quence demonstrated. The system activates in turn all the sub-goals of the final target state, and does not trigger action unless the complete sequence was not simulated first. This process can be defined as multiple satisfaction of all the sub-goals. External action is driven towards the minimization of the distance between the final goal (i..e, the complete seriate structure demonstrated) and the current state of the environment. During learning, the increased saliency of the target states allows the construction of a higher CA level where only target goals are stored. At this level, saliency during demonstration is mapped into intentionality during imitation.

8. Ongoing work

Where does the linguistic input fits into the seriate cups task? In our view, it represents another source of information to be integrated by a multiple constraint satisfaction process. We embrace Seidenberg and McDonald (1999) unified view on the bootstrapping mechanisms, defined as components of a general constraint satisfaction process that exploits correlations between multiple types of information. Preliminary experiments using an associative architecture and cross-situational learning have been run to learn the meaning of a small set of words, from sentences paired with the actions of the simulated agent (i.e., 'Look the small red cup'). Our method was that of autonomous bootstrapping (Brent, 1996), consisting in extracting tiny bits of linguistic knowledge and using them for further analysis of the inputs. Further development of the system is towards the processing of multiple cues, such as word order and collocations, as well as integrating contextual constraints. The goal is to study not only how the linguistic function can make use of the sequence detectors developed for the seriate ability, but also how goal-directed action can use top-down linguistic cues to discover 'what' and 'how' to imitate.

Acknowledgments We are very grateful to Stefan Schaal to have provided access to the Xanim simulation environment. This work was supported by the Swiss National Science Foundation.

References

Bauer, P. J. (2002). Long-term recall memory: Behavioral and neuro-developemental changes in the first 2 years of life. *Current Directions in Psychological Science*, 11.

Bekkering, H., Wohlschlager, A., Gattis, M. (2000). Imitation is goal-directed. *Quarterly Journal of Exp. Psychology*, 153-64.

Billard, A., Hayes, G. (1999). DRAMA, a connectionist architecture for control and learning in autonomous robots.*Adaptive Behavior Journal*, 7(1).

Billard, A. (2002). Imitation: a means to enhance learning of a synthetic proto-language in an autonomous robot. In K. Dautenhahn and C. L. Nehaniv (eds.), *Imitation in Animals and Artifacts*, MIT Press.

Billard, A., Epars, Y., Schaal, S., Cheng, G. (2003). Discovering Imitation Strategies through Categorization of Multi-Dimensional Data. *Procs. Int. Conference on Intelligent Robots and Systems* IROS' 03.

Brent, M.R. (1996). Advance in the Computational Study of Language Acquisition. *Cognition*, 61.

Carey, S., Xu, F., (2001). Infant's knowledge of objects: beyond object files and object tracking? *Cognition*, 80.

Chappelier, J.C., Gori, M., Grumbach, A., (2001). Time in Connectionist Models. *Lecture Notes in Computer Science*, 1828.

Dienes, Z., Altmann, G., Gao, S-J. (1999). Mapping across domains without feedback: A neural network model of implicit learning. *Cognitive Science*, 23.

Dominey, P.F. (2003). Learning Grammatical Constructions from Narrated Video Events for Human-Robot Interaction. *Procs. of the IEEE Humanoid Robotics Conference*, Karlsruhe, Germany.

Dorffner, G. (1992). A step toward sub-symbolic language models without linguistic representations. In R.G. Reilly and N.E. Sharkey (eds.) *Connectionist approaches to Natural Language Processing*. LEA, UK.

Greenfield, P., Nelson, K., and Saltzman, E. (1972). The development of rulebound strategies for manipulating seriated cups: a parallel between action and grammar. *Cognitive psychology*, 3.

Greenfield, P. (1991). Language, tool and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior, *Behav. Brain Sci.*, 14.

Itti, L., Koch, C. (2001). Computational modeling of visual attention. *Nat. Rev. Neurosci.*, 2(3).

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8).

Metta, G., and Fitzpatrick, P. (2002). Better vision through manipulation. *Procs. of 2nd Int. Workshop on Epigenetic Robotics*. August 2002, Edinburgh, Scotland.

Nothdurft, H.C. (2000). Salience from feature contrast: additivity across dimensions. *Vision Research* 40.

Piaget, J. (1970). *Genetic Epistemology*, New York: Columbia University Press.

Pulvermuller, F. (2003). *The Neuroscience of Language. On Brain Circuits of Words and Serial Order*. Cambridge University Press.

Regier, T. (1995). A Model of the Human Capacity for Categorizing Spatial Relations. *Cognitive Linguistics*, 6(1).

Reilly, R.G. (1997). Cortical software re-use: A neural basis for creative cognition. In T. Veale (ed.) *Computational models of creative computation*.

Seidenberg, M.S., MacDonald, M.C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23(4).

Schaal, S. (2000). The SL simulation and real-time control software package, USC.

Siskind, J. M. (1995). Grounding language in perception. *AI Review*, 8.

Tomasello, M. (1988). The role of joint attentional processes in early language development. *Language Sciences*, 1.

Zlatev, J., Balkenius, C. (2001). Introduction: Why epigenetic robotics ? *Proc. First Workshop on Epigenetic Robotics,* Lund, Cognitive Studies 85, 2001.