13

Development of goal-directed imitation, object manipulation, and language in humans and robots

Ioana D. Goga and Aude Billard

13.1 Introduction

The aim of the present volume is to enrich human language dimensions by seeking to understand how the use of language may be situated with respect to other systems for action and perception. There is strong evidence that higher human cognitive functions, such as imitation and language, emerged from or co-evolved with the ability for compositionality of actions, already present in our ancestors (Rizzolatti and Arbib, 1998; Lieberman, 2000; Arbib, 2003; Arbib, Chapter 1, this volume). Corroborating evidence from psychology (Greenfield *et al*., 1972; Iverson and Thelen, 1999; Glenberg and Kaschak, 2002), neurobiology (Pulvermüller, 2003) and cognitive sciences (Siskind, 2001; Reilly, 2002) strongly support a close relationship between language, perception, and action. Social abilities, such as imitation, turn-taking, joint attention and intended body communication, are fundamental for the development of language and human cognition. Together with the capacity for symbolization, they form the basis of *language readiness* (Rizzolatti and Arbib, 1998; Arbib, 2003).

The work presented in this chapter takes inspiration from this body of experimental evidence in building a composite model of the human's cognitive correlates to action, imitation and language. The model will contribute to develop a better understanding of the common mechanisms underlying the development of these skills in human infants, and will set the stage for reproducing these in robots and simulated agents.

A recent trend of robotics research follows such views, by equipping artifacts with social capabilities. The rationale is that such abilities would enable the artifact to communicate with humans using "natural" means of communication (Schaal, 1999; Breazeal and Scassellati, 2002; Billard and Mataric, 2001; Kozima and Yano, 2001; Demiris and Hayes, 2002). We follow this trend and investigate the role that imitation and joint attention play in early language acquisition. This work builds upon other work of ours that investigate the basics components of imitation learning, such as the ability to extract the important features of a task (*what to imitate*) (Billard *et al*., 2003; Calinon and Billard,

Action to Language via the Mirror Neuron System, ed. Michael A. Arbib. Published by Cambridge University Press.

in press), and the ability to map motion of others into one's own repertoire (*how to imitate*) (Sauser and Billard, 2005). See the chapters by Greenfield and by Zukow-Goldring (this volume) for complementary work on the role of the caregiver in the child's acquisition of these abilities.

We now briefly outline the methodological approach followed in this work.

13.1.1 Outline of the methodological approach

A considerable body of cognitive and robotic theories point to at least three conditions to be met by a system, in order for the system to develop human-like cognition: (a) *sociocultural situatedness*, understood as the ability to engage in acts of communication and participate in social practices and language games within a community; (b) *naturalistic embodiment*, that is, the possession of bodily structures to experience the world directly; (c) *epigenetic development*: the development of physical, social, and linguistic skills in an incremental, step-wise manner (Harnad, 1990; Clark, 1997; Brooks *et al*., 1998; Zlatev and Balkenius, 2001; Steels, 2003).

Embodiment

Embodiment allows artificial systems to ground symbolic representations in behavioral interactions with the environment in such a way that the agent's behaviors, as well as its internal representations, are intrinsic and meaningful to itself (Harnad, 1990; Ziemke, 1999). Symbols are grounded in the capacity to discriminate and identify the objects, events, and states of affairs that they stand for, from their sensory projections (Regier, 1995). In addition, by enabling the agent to act upon its environment, the transduction mechanism develops a functional value for the agent, and can be considered meaningful to itself (Bailey, 1997; Brooks *et al*., 1998).

Embodiment is at the core of our methodology. In previous work, we took the stance that the robot's body was fundamental to convey and ground meaning in words, transmitted and taught by another teacher agent (Billard, 2002). Meaning was, thus, grounded in the learner robot's perceptions (i.e., sensor measurements and motor states).

Development

A second core stance of our methodology stresses the role that development plays in the acquisition of compositional skills, such as language and imitation. Development represents a framework through which humans acquire increasingly more complex structures and competencies (Piaget, 1970). A developmental process starting with a simple system that gradually becomes more complex allows efficient learning throughout the whole process, and makes learning easier and more robust. A developmental framework also supports the system capacity to organize words and concepts in a hierarchical, recursive, and compositional fashion. The epigenetic developmental approach is increasingly exploited in the artificial intelligence (AI) field to account for the building of complex

cognitive structures from low-level sensorimotor schemas (Metta *et al*., 1999; Weng *et al*., 2001; Zlatev and Balkenius, 2001; Reilly and Marian, 2002).

The starting point of our approach to action, imitation, and language is the definition of a developmental benchmark, against which modeling can be compared. This benchmark has to meet several criteria: (a) it must be grounded in the observation of infants' behavior in a complex social scenario, which involves social interaction, imitation, object manipulation, and language understanding and production; (b) it must permit the characterization of developmental stages in the acquisition of the skills under observation; (c) it should be sufficiently realistic, so that it could be replicated experimentally; (d) the infants' behavior under observation must be such that they can be modeled and implemented in artificial systems. The *seriated nesting cups* task (Greenfield *et al*., 1972; see also Greenfield, 1991, and this volume) is our benchmark.

The role of imitation

We argue that the abilities to imitate and to manipulate objects lay at the foundation of language development in humans. The capacity to imitate goal-directed actions plays an important role in coordinating different behaviors. Billard and Dautenhahn (2000) showed that imitation allows sharing of a similar perceptual context, a prerequisite for symbolic communication to develop. In the work reported here, imitation will be studied as exploiting the capacity to recognize and extract others' actions (where mirror neurons have been shown to play a central function), to memorize, learn, and reproduce the demonstrated behavior. Our working hypothesis is that imitation requires the ability to extract meaning, by inferring, and furthermore, by understanding the demonstrator's goal and intention (Byrne and Russon, 1998). Moreover, the ability to infer others' intention, to imitate novel behavior and the capacity for language is investigated as a process that follows a common developmental path, and which may have a common underlying neural process.

Thus, essential to our approach is the Mirror System Hypothesis (Arbib, Chapter 1 this volume), that is, the idea that the communication system develops atop an action system capable of object manipulation, with its capacity to generate and recognize a set of actions. This assumption represents a starting point for the present approach, whose goal is to investigate the computational means by which interaction of the action and language systems develops on a common neural substrate. The model we develop here follows from and complements other works by providing a more detailed description of the role of imitation and joint attention in the acquisition of sequential manipulation of objects.

Modeling

The core of our methodological approach, outlined in Fig. 13.1, resides in the computational modeling of the developmental path that human infants follow in developing the capacity to create assemblages of objects and to generate well-formed sentences. Our model is constrained by evidence from neuroscience and developmental psychology. The

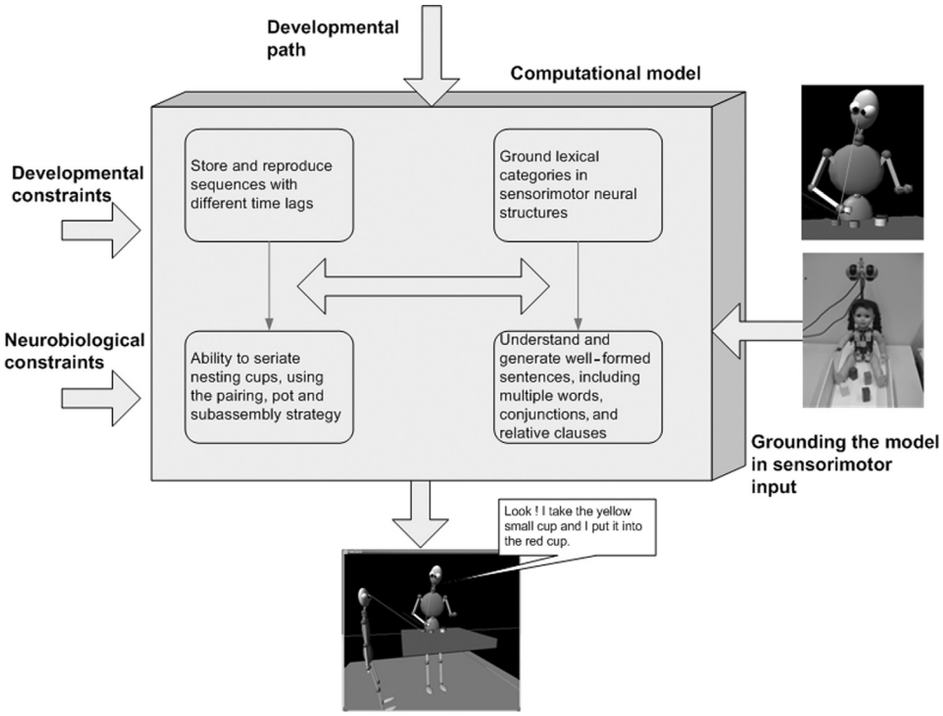

Figure 13.1 A neurobiologically and developmentally constrained approach to action–language modeling task. The developmental path envisaged has four hypothesized stages: (a) development of the ability to learn and reproduce sequences with different time lags and complexity; (b) acquisition of the capacity to seriate nesting cups, through an epigenetic process that replicates the strategic rule-bound behavior observed in human infants; (c) learning of a number of lexical categories grounded on the sensorimotor neural structures developed previously; (d) development of the capacity to generate well-formed, meaningful sentences that describe the actions performed, by following the same linguistic developmental path that human infants follow.

acquired knowledge is grounded in the sensorimotor interaction of the simulated agent with the environment.

We start with a simple system capable first of imitating short sequences of actions, that gradually develops afterwards to accommodate more complex behaviors. As learning and development proceed, prior structures get integrated and provide competencies that can be reused (the vertical flow in Fig. 13.1). Previously acquired behavioral manifestations put constraints on the later structures and proficiencies (the horizontal flow in Fig. 13.1). To facilitate learning, the gradual increase in internal complexity is accompanied by a gradual increase in the complexity of the external world to which the "infant" is exposed (see more on the language modeling framework in Section 13.6).

The rest of this chapter is structured as follows. Section 13.2 describes the original experiment of seriated nesting cups (Greenfield *et al*., 1972) and discusses the relevance

of this task to goal-directed imitation and the action–language analogy. Section 13.3 introduces a general computational framework for the modeling of the action–language re-use hypothesis. This section focuses on the presentation of the neurobiological constraints and learning principles that can support incremental construction of both action and language systems. In Section 13.4 we describe a dynamic simulation of the seriated nesting cups task with a child–caregiver pair of humanoid robots. A number of developmental constraints specific to each imitation stage are presented and integrated in the model. Section 13.5 presents the imitative behavior of the agent corresponding to the first developmental stages of nesting cups behavior. Section 13.6 capitalizes on the seriated nesting cups model and presents a general overview of the action–language model envisaged.

13.2 The seriated nesting cups experiment

The seriated nesting cups experiment by Greenfield *et al*. (1972) was conducted as part of a research program on the development of hierarchical organization in children, based on the observation of manual object combination tasks. Systematic development towards increasingly complex hierarchical organization has been repeatedly observed for nesting cups (Greenfield *et al*., 1972), nuts and bolts (Goodson and Greenfield, 1975), and construction straws (Greenfield and Schneider, 1977).

13.2.1 Development of rule-bound strategies for manipulating seriated cups

Greenfield *et al*. (1972) report that children between 11 and 36 months of age exhibit different strategies, correlated to their developmental age, for combining cups of different sizes. The seriated nesting cups experiment consists first of a demonstration phase, during which the experimenter manipulates the cups to form a nest (i.e., insert all cups into one another using the most advanced (subassembly) strategy (see Fig. 13.2c)), followed by a spontaneous imitation phase, during which the child is left free to play with the cups.

Three manipulative strategies were identified and analyzed: (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) *the 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 (see Fig. 13.2).

The child's choice of the acting/acted upon cups seems to be based on one of three criteria: size, proximity, and contiguity. During the first developmental stage, children typically use the proximity criterion (i.e., same side of the table with the moving hand) for pairing cups. Children also operate as though size is a binary concept, with one cup treated as the "biggest" while all the others belong to the category "little." A common phenomenon is the transfer of a single moving cup from one stationary cup to another without letting go of the original cup. Intermediate constructions between stage 1 and

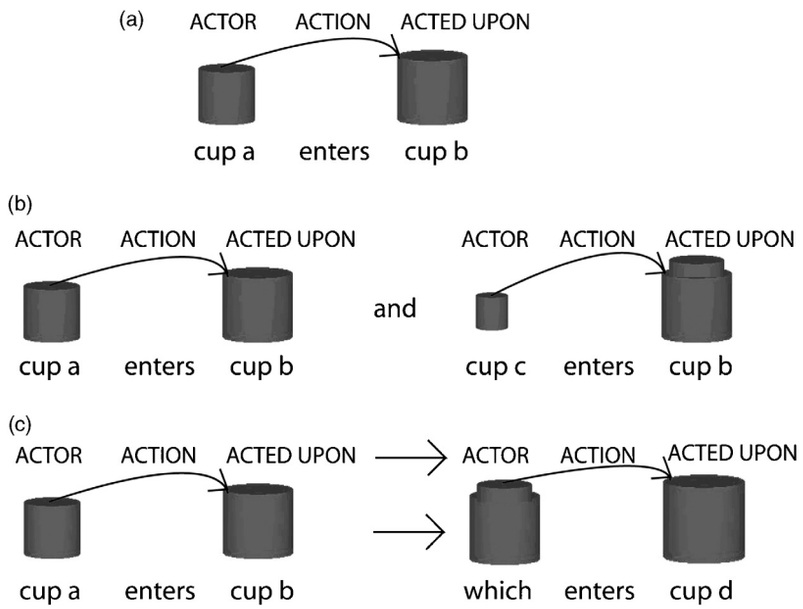

Figure 13.2 Formal correspondence between developmental manipulation strategies and sentence types. In each frame, action relations are shown on the top, the imitative behavior is shown in the center, while a descriptive sentence on the bottom illustrates the corresponding grammatical relation. (a) Strategy 1: the pairing method corresponds to a grammatical relation of type *subject–verb–object*. (b) Strategy 2: the pot method corresponds to a relation of type *subject–verb–object* and *subject–verb–object*. (c) Strategy 3: the subassembly method, corresponds to the composition with ellipsis *subject–verb–object* → *which[subject]–verb–object*. (Adapted after Greenfield *et al*., 1972.)

stage 2 consist in forming two pairs or one pair and a structure of three embedded cups. Children at 16–24 months old seem to follow only the contiguity criterion (i.e., never reaching behind a nearer cup to use a more distant cup), while the 28- to 36-month-olds seems to follow the size criterion.

The consistency of each individual's strategic behavior is defined in terms of the dominant or most frequent strategy. A child's dominant strategy accounted on average for 80% of his structures. On the other hand, each infant uses at least once an intermediate manipulative method and consistency of the strategic behavior slightly decreases with the age.

13.2.2 Development of goal-directed imitation

The seriated cups experiment is of relevance to the theory of goal-directed imitation (Bekkering *et al*., 2000). In order to imitate complex behaviors, one must recognize goals, understand how individual actions are embedded in a hierarchy of subgoals, and

recognize recursive structures. Normally developing children do not copy exactly the act of adults. They adapt their imitation as an effect of the inferred intended goal of the action. For instance, when presented with a simple goal-directed action (e.g., reaching with the left arm to the right ear), children have a tendency to reproduce first the goal (touching the ear), rather than (at first) paying attention to the limb used to achieve the goal (using the left arm) (Bekkering *et al*., 2000).

The imitative behavior of infants during the seriated nesting cups task reflects both the capacities of the imitator and the characteristics of the internal model developed. Greenfield and colleagues (1972) suggested that during the first developmental stages, the model was mainly preserved as a generalized goal state: "to put the cups inside each other." Moreover, it seems that the youngest children used the model mainly to get the nesting behavior going, whereas the older children used it as a basis for terminating activity as well. For them, the final stage of the demonstration appeared to function as a precise goal and signal for termination. In Section 13.5 we propose a computational account for the limitations described in the infants' abilities of goal-directed imitation, based on the particularities of the internal model developed during demonstration.

13.2.3 The action–grammar analogy hypothesis

A related objective of Greenfield *et al*. (1972) was to investigate the question of a formal homology between strategies for cup construction and certain grammatical constructions. Figure 13.2 illustrates the analogy they proposed between the 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 pairing strategy is dominant at 11- and 12-month-old infants, and corresponds to the use at this age of simple sentences, formed from two or three words. 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, paralleling the capacity to use *relative clauses* in language. Sources of evidence on the relative ordering of these types of grammatical constructions are provided by experimental studies showing that conjunction of sentences was frequent and preceded relative clauses in the speech of children aged 18 months to 3 years (Smith, 1970).

Greenfield (1991) put forward the hypothesis of a neural structural homology, rather than just an analogy, between action strategies and grammatical constructions. She adduced evidence from neurology, neuropsychology, and animal studies to support the view that object combination and speech production are built upon an initially common neurological foundation, which then divides into separate specialized areas as development progresses. Her hypothesis is that early in a child's development Broca's region may serve the dual function of coordinating object assembly and organizing the

production of structured utterances. Computational support to this hypothesis was brought by Reilly (2002) (see also Section 13.6.2).

13.3 General overview of the computational model

The task of the present section is to introduce the main concepts of the computational framework. In our view, learning to seriate nesting cups and to generate grammatical constructions have some common needs: (a) the capacity to represent categorical information in a subsymbolic manner; (b) the operation of a mechanism for grounding internal representations on sensorimotor processes; (c) the ability to learn from and to represent time-ordered sequences; (d) the capacity to process and satisfy multiple constraints in a parallel manner; (e) the operation of a computational mechanism that supports cross-domain bootstrapping. In this section we address the requirements (a)–(d), while the challenging issue of inter-domain transfer of information is tackled in Section 13.6 in the description of the language modeling framework.

13.3.1 A decompositional approach to knowledge representation

The subsymbolic paradigm states that the brain represents cognitive symbols through their decomposition into smaller constituents of meaning, usually called *semantic features* or *microfeatures* (Sutcliffe, 1992). Furthermore, concepts in the brain are grounded on semantic neural networks, involved in the perception or execution of the corresponding symbols (see the grounding paradigm: Harnad, 1990). That is because, when the meaning of a concrete content word is being acquired, the learner is exposed to stimuli of various modalities related to the word's meaning, or the learner may perform actions the word refers to.

Recent neuroscientific evidence corroborates the decompositional and grounding computational approaches on language representation. Neuroimaging studies of language (Pulvermüller, 1999, 2002; Hauk *et al*., 2004) support the idea that words are represented in the brain by distributed cell assemblies whose cortical topographies reflect aspects of word meaning (including action-based representations). There is evidence that: (1) assemblies representing phonological word forms are strongly lateralized and distributed over perisylvian cortices; (2) assemblies representing concrete content words include additional neurons in both hemispheres; (3) assemblies representing words referring to visual stimuli include neurons in visual cortices; (4) assemblies representing words referring to actions include neurons in motor cortices (see Pulvermüller (1999) for a review on the neurobiological data).

The great promise of the distributed, neurobiologically inspired approach to knowledge representation is the integration of learning and representation within the same structures. The pitfall of this approach is the increasing complexity of the biologically plausible computational models (Marian *et al*., 2002). For a detailed introduction to biologically

realistic neural architectures we refer the reader to Maas and Bishop (1999). In this work, we get inspiration from neurobiological studies and define a computational primitive, which is both simple and scalable, suited for the operation on cognitive and linguistic structures.

13.3.2 The cell assembly concept

The computational building-block of our system is a neural primitive referred to as a *cell assembly*. The concept is envisaged along the lines proposed by Hebb (1949), Pulvermüller (2002), and Frezza-Buet and Alexandre (2002). Hebb (1949) suggested that a cell assembly is a reverberatory chain of neurons, with many re-entrant loops through which activity waves can travel repeatedly. More recently, Pulvermüller (2002) uses the concept of a *neural set*, to refer to functional webs characterized by a great variety of activity states. Because of its strong internal connections, the neuronal set is assumed to act as a functional unit, which can be primed, ignite, and reverberate.

Our approach is inspired by Pulvermüller's definition of neuronal set. A cell assembly is a neuronal set (i.e., a selection of neurons that are strongly connected to each other, act as a functional unit, can be primed and ignite) with additional special properties that are relevant to category information processing. Each cell assembly receives input from external sensorial units and can have an activating effect on other cell assemblies directly connected to it (Fig. 13.3). A cell assembly can: (a) learn a subsymbolic feature representation of an external concept or event (feedforward flow in Fig. 13.3), and (b) become a node in a sequence detector (precedence flow in Fig. 13.3). A third learning mechanism (relation flow in Fig. 13.3) extracts and stores information concerning systematic relations among sensorial or semantic features that constitute specific categories. Anti-Hebbian learning (Levy and Desmond, 1985) is used to extract the invariants between the items that are in the focus of attention.

13.3.3 Learning framework

The general approach to learning is similar to the biologically inspired model of the cortex described by Frezza-Buet and Alexandre (2002), including elementary functions in the perceptive, motor, or associative domain. The basic computational unit in their work is a maxicolumn, consisting of a set of cortical-like columns. The maxicolumn is grounded on sensorial and motor maps, has associative functions on higher-order maps, and allows three types of learning: (a) learning of the external event that the column represents, through feedforward connections; (b) associative learning within a map, through lateral connections; and (c) sequence learning. Causal sequence learning is based on a rule that finds the conditional stimulus that is most frequently associated with the unconditional stimulus, and which predicts it. The mechanism is employed to find the neural component in a map whose activity predicts the satisfaction of another neural component in the map, referred to as a goal (Frezza-Buet and Alexandre, 2002).

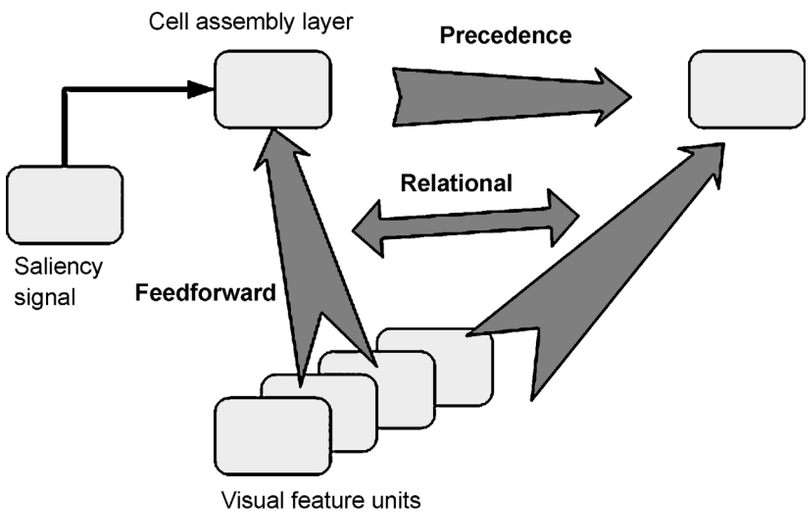

Figure 13.3 Representation of feedforward and lateral information flows of a cell assembly. Each kind of link represents a type of learning that a cell assembly has to manage. The saliency signal is received from the attention module (not shown), and it is inextricably bound up with the creation and satisfaction of a cell assembly.

Category learning

The connectivity patterns represented in Fig. 13.3 define the kinds of information flows, and thereby, the kinds of learning that a cell assembly can be involved in. The first type of learning (i.e., feedforward connections in Fig. 13.3) tunes the cell assembly to respond to a specific distribution of its input information. The learning algorithm is inspired by Adaptive Resonance Theory (ART) (Grossberg, 1976; Carpenter and Grossberg, 1987). ART networks were designed to overcome the stability–plasticity dilemma, that is, how can a system be stable against noisy or irrelevant data and yet remain plastic enough to learn novel inputs without affecting already learned knowledge. A central feature of all ART systems is a pattern-matching process that compares an external input with the internal memory of an active code. ART matching leads either to a *resonant* state, which persists long enough to permit learning, or to a parallel memory search. If the search ends at an established code, the memory representation may either remain the same or incorporate new information from matched portions of the current input. If the search ends at a new code, the memory representation learns the current input. The criterion of an acceptable match is defined by a dimensionless parameter called *vigilance*. Vigilance weights how close an input must be to the prototype for resonance to occur. Low vigilance leads to broad generalization and more abstract prototypes than high vigilance.

The cell assembly unit represents a category from ART with additional properties that are relevant for temporal pattern processing (i.e., graded activation and memory decay)

and for extraction of correlation information. Different kinds of information are extracted during the learning process from the external input. The activated cell assemblies learn through an unsupervised adaptation process the distribution of feature information in the external sensorial map. Classical Hebbian, non-supervised learning (Hebb, 1949) based on the strengthening of synaptic weights between co-activated units, can be successfully used for this task.

Learning of temporal sequences

The second type of learning concerns processing of temporal sequences that are inextricably bound up with behaviors, such as language and object manipulation. Learning structure from temporal sequences, as well as being able to output symbols that are ordered in time, needs a system ability to detect and generate sequences. The common need of the architectures aimed at processing temporal sequences is the presence of a short-term memory and a prediction mechanism for the serial order (Wang, 2002). Recurrent networks can in principle implement short-term memory using feedback connections (Billard and Hayes, 1999). Temporal sequences are learned as a set of associations between consecutive components (see Chappelier *et al*. (2001) for a recent review of the fundamental types of temporal connectionist models).

The initial implementation of the cell assembly concept was inspired by previous work of ours on sequential learning with time-delay networks (Billard, 2002). A cell assembly was represented there by a node to which self-recurrent connections have been added and a time decay rate was used to simulate a short-term memory of the assembly. The cell assembly can in this case transit several states (Fig. 13.4a): (a) an *ignition* state corresponding to the maximal activation of the cell assembly, (b) an *activation* state corresponding to the decaying memory of its ignition, and (c) an *inactive* state, when its activation is below an arbitrary set threshold.

Once information from the sensorial external map reaches the cell assembly, it is then further memorized for a period of time, during which it can be associated with any incoming event in any other sensory system. This leads to a system capable of associating events delayed in time with a maximal time delay equal to the length of the working memory. Time-delay networks (Billard and Hayes, 1999) or simple recurrent networks (Elman, 1993) can be employed with success for learning the sequential order of events.

Storing temporal dependencies using graded activation states

To deal with temporal dependencies beyond consecutive components different solutions were explored. Recent models propose different ways to enhance the memory capacity, by carefully designing the basic computational unit (Hochreiter and Schmidhuber, 1997) or by exploiting learning in hierarchical structures (Tan and Soon, 1996).

In this model, learning of temporal dependencies is supported by the graded activation of the cell assemblies, and is facilitated by the layered architecture of the model. Each cell assembly can transit an increased number of activation states. This is achieved by

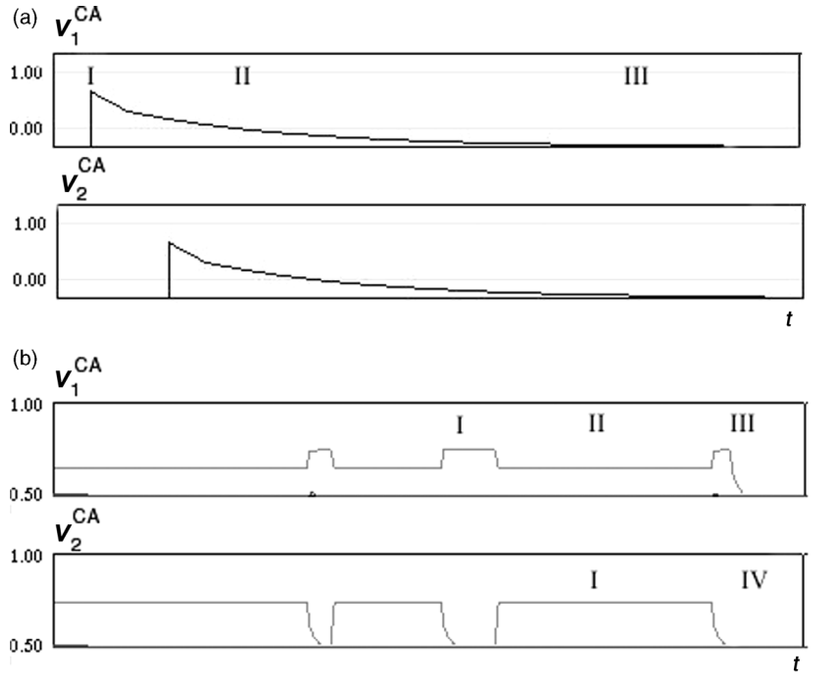

Figure 13.4 (a) Cell assembly activation curves characterized by three operation regimes: *ignition* (I), *memory decay* (II), and *inactivation* (III). (b) Activation curves with four functional regimes: *maximum activation* (I), corresponding to maximal satisfaction of the cell assembly feature constraints; *sustained activation* (II), when the cell assembly feature constraints partially satisfy the state of the external world; *memory decay* (III), when the cell assembly is no longer satisfied and its activation decays as a function of the decay rate; and *inactivity* (IV). The second regime (sustained activation) was introduced to deal with time dependencies larger than the time constant of the decay rate.

implementing at the cell assembly level a computational mechanism that keeps the CA active for larger periods of time. The mechanism relies on the concept of a cell assembly's *degree of satisfaction* and it is explained in more detail in Section 13.4.5. Briefly, the activation of a cell assembly becomes a function of how well its subsymbolic feature representation of the learned event satisfies the current state of the external world. Consequently, a cell assembly can be in one of the following states: (a) *maximally satisfied*, leading to a temporary state of maximal activity (regime I in Fig. 13.4b); (b) *partially satisfied*, allowing a prolonged activation regime (regime II in Fig. 13.4b); (c) *unsatisfied*, when its activation decays as a function of the memory rate (regime III in Fig. 13.4b); or (d) *inactive*.

A layered neural architecture favors temporal integration at different timescale levels. In our view, the creation of a cell assembly is inextricably bound up with the receipt of a saliency signal (Fig. 13.3; see also Section 13.4.4). The sensorial external maps operate on a short timescale, in the sense that they respond spontaneously to external inputs. At

this level of visual awareness, the system's capacity to store associations is limited to the duration of the short-term memory. A saliency signal received by an object which is in the focus of attention can enhance the object's neural representation and enable the creation of a cell assembly unit, as an enhanced "copy" of the original representation (see the discussion of salience and "top–down" attention by Itti and Arbib, this volume). The cell assembly operates on a larger timescale, allowing the system to extract and store temporal sequences with various time lags.

Graded activation of the cell assembly as a function of the satisfaction level leads to the formation of precedence relations that reflect not only serial order, but also causal relations. This is because satisfaction is a graded measure of how well a given category matches the current state of the external world. Consequently, related categories have similar levels of satisfaction, favoring learning of precedence relationships between causally related categories. Cell assemblies representing in turn "right hand," "right hand grasping an object," and "right hand carrying an object" have similar levels of satisfaction, and hence, can learn a set of associative or precedence connections.

The precedence learning mechanism operates as follows: if a cell assembly i is repeatedly satisfied after the activation of a cell assembly j, then satisfaction of j is supposed to be one of the events possibly required for the occurrence of i. The associative connection between j and i is adapted as a function of the rapport of the cell assemblies' activations. A systematic causal relation increases the strength of the weight, and turns the cell assembly j into a predictor of i. On the other hand, a large fluctuation in the order in which i and j are satisfied will decrease the weight, to an insignificant value. This type of mechanism can account for unary precedence relations of the type $A \rightarrow C$ or $B \rightarrow C$, where A and B are learned predictors of C.

When would a cell assembly coding for concept C respond to a binary sequence $A \rightarrow B$ and not to $B \rightarrow A$? We envisage the operation of a sequence detector mechanism similar to that described in Pulvermüller (2002). The mechanism is inspired from the functioning of the movement detectors in the visual cortex (Hubel, 1995) and is based on the following idea: sensitivity to the sequence $A \rightarrow B$ may involve low-pass filtering of the signal A, thereby delaying and stretching it over time. Integration at the site of a third neuron k of the delayed and stretched signal from i and the actual signal from j yields values that are large when activation of i precedes that of j, but small values when the activation of i and j occurs simultaneously or in reverse order. This yields strong weights between j and k but not between i and k (after Pulvermüller, 2002). A similar mechanism is implemented for the retrieval of precedence relationships during the simulation of the seriated nesting cups task (see more details in Section 13.4.6). For our goals, sequence detectors operating on no more than two or three nodes are sufficient.

13.3.4 A multiple constraints satisfaction framework

Sequential learning is sufficient neither for the acquisition of action grammars nor of language grammars. A basic property of language is the translation of a hierarchical

Table 13.1 *Steps of the retrieval process*

1. Focus attention on the most salient objects and words from the environment.
2. Cell assemblies corresponding to the distribution of features that are in the focus of attention become satisfied.
3. Hidden cell assemblies are primed through the lateral connections and precedence constraints are computed.
4. Primed, unsatisfied cell assemblies compete for setting the goals of the system.
5. Action is initiated to satisfy the most important goal and to minimize the internal global dissonance.
6. The objects of action are chosen as a function of the internal and external constraints existent in the system.
7. Termination of action occurs when all (or the most important) goals of the system are satisfied.

conceptual structure into a temporally ordered structure of actions, through a constraint satisfaction process, that integrates several types of constraints in a parallel manner (i.e., phonological, syntactic, semantic, contextual) (Seidenberg and MacDonald, 1999). In a neural network, constraints are encoded by the same machinery as knowledge of language itself, and this is clearly an advantage over approaches in which symbolic knowledge is represented separately from the system that makes use of it.

The computational framework developed in this work is based on the assumption that the reproduction of a sequence of motor or linguistic actions represents the product of the interplay between sequence detectors operating on short temporal dependencies, and a general constraints satisfaction framework. This framework controls the way different types of constraints (including sequence detection) are integrated. In particular, we applied this framework to the reproduction of the sequence of actions demonstrated in the seriated nesting cups task and we are currently exploring its application to the generation of well-formed, descriptive sentences. As stressed in the introduction, imitation is addressed as a means for disassembling and reassembling the structure of the observed behavior. We state that the reassembling task can best be described as *a multiple constraints satisfaction process*.

During the retrieval of a series of actions (either motor or linguistic) the system acts in a constrained manner, in accordance with the knowledge stored in all sets of weights. Most generally, the retrieval process consists of seven steps, which are summarized in Table 13.1.

Initiation of any type of action must be preceded by the activation of an internal goal. A cell assembly represents a *goal* of the system, if the cell assembly is not satisfied and if its precedence constraints are met over a certain threshold. Further on, the behavior of the agent is driven by a process of *dissonance minimization*. That is, the system computes the total dissonance (i.e., difference) between its goals and the current state of the environment, and acts towards the minimization of this dissonance state. The working

definition of dissonance is inspired by the work of Schultz and Lepper (1992) on modeling the cognitive dissonance phenomena described by Festinger (1957).

13.4 A model of the seriated nesting cups task

13.4.1 The simulation environment

The current implementation of the model was created within the Xanim dynamic simulator (Schaal, 2001) designed to model a pair of 30 degrees of freedom (head 3, arms 7×2, trunk 3, legs 3×2, eyes 4) humanoid robots (Fig. 13.5). 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 simulated robot is controlled from Cartesian coordinates through inverse dynamics and inverse kinematics servos. A motor servo is used to read the current state of the robot/simulation (i.e., position, color, orientation and rotation angles, and motion speed) and to send commands to the robot/simulation. The environment is controlled, in other words, only a predefined set of objects and end-effectors are visually recognized and manipulated.

13.4.2 Developmental constraints

In modeling the seriate nesting cups behavior, we investigated the effect of varying a number of parameters of the computational model, as a way of accounting for systematic differences in child behavior. In particular, we considered the effects of shared attention, memory capacity, categorization, and development of the object concept on the child's/robot's ability to compose manipulation and linguistic steps. Each developmental stage is modeled through the combination of the effects of several developmental constraints, as shown in Fig. 13.6. The integration of each developmental constraint at a given simulated age is based on the review of psychological developmental literature (see also Section 13.7 for a discussion of the experimental sources of evidence). The hypothesized model is validated through the replication of the behavioral manifestations of human infants with a pair of humanoid robots, as described in Section 13.5. However, it represents only a possible explanation for the development of nesting cups ability, and it cannot be generalized to other developmental issues.

13.4.3 Architecture and functioning principles

The model consists of a hierarchy of connectionist architectures, whose basic computational unit is depicted in Fig. 13.7. The lower two layers of the network implement a mechanism for object recognition and an attention mechanism. The last layer implements the cell assembly network. The purpose of the layered architecture is to support temporal integration at different timescales. The reduction of the hierarchy to three levels results

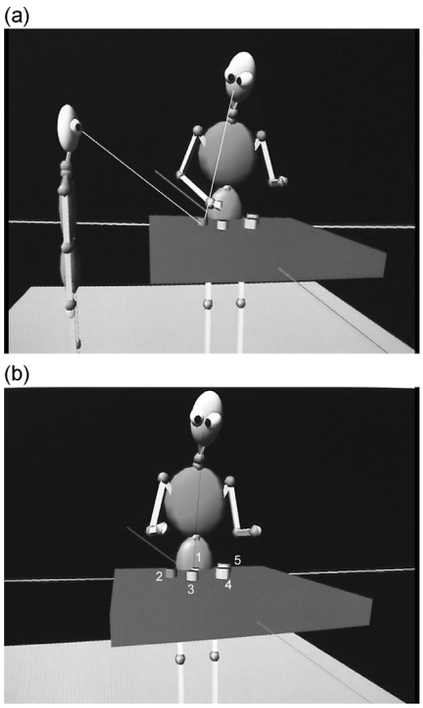

Figure 13.5 The Xanim dynamic simulation of a child–caregiver pair of humanoid robots. (a) The Caregiver demonstrates the seriated cup task. The shared focus of attention of Child and Caregiver is highlighted by the crossing of their gaze directions. (b) The Robot Child imitates the seriated cups task.

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 (see Section 13.4.1) into commands to the simulated robot links. A one-to-one mapping is imposed in this case, between what the agent sees and what it executes.

Let us describe how the system processes visual input. At the first parsing of the visual scene, the object recognition subnetworks (OR) from all locations are activated and remain active as long as the objects remain visible. No segmentation of the visual scene is required, because the simulated agent can recognize a predefined number of objects: the end-effectors, the objects located on the table, and the table. For each object, the color, size, shape, rotation angle, and the speed of motion are read out from the robot's sensors. These values are fed in the network through the activation of the external layer's units.

Saliency (SA) is computed in a distributed manner at all locations of the visual scene and the unit with the highest activation wins the focus of attention. The winning 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

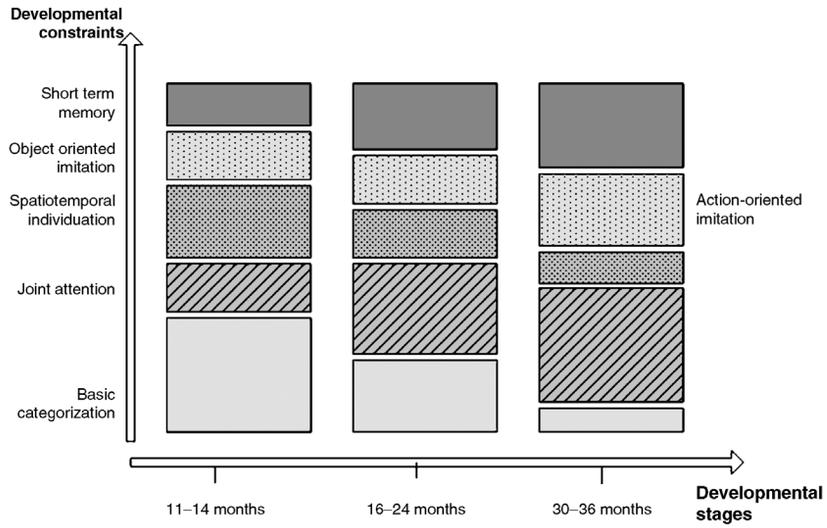

Figure 13.6 Hypothetical model of the developmental stages of the seriate cups ability. On the *Y*-axis is shown the contribution of each developmental constraint to the learning process for a given age. The height of the boxes indicates the weight of the corresponding process, in relation to the learning of the serriated cups task. The earliest developmental stage is accounted for through a basic categorization process, accompanied by spatiotemporal individuation and a limited short-term memory. Mervis (1987) showed that 2-year-olds form basic-level categories and suggested that infants under 2 years of age attend to the function and shape of objects to categorize. There is also evidence that infants up to 1 year rely almost exclusively on spatiotemporal information (i.e., location) to build distinct representations of the objects (Carey and Xu, 2001). Object-directed imitation and an increase of the attention and memory resources for object-related information characterize the second developmental stage. By the end of the first year of life, infants show long-term memory for serial order, and can retain in visual short-term memory about three to four items (Rose *et al*., 2001). A small number of attended objects may be indexed in time, the indexed individuals tracked through time and space, and the spatial relations among indexed individuals represented (Carey and Xu, 2001). When objects are involved as goals, infants learn first about them, and movements are imitated if there is no object goal, or if this is ambiguous (Bekkering *et al*., 2000). The third developmental stage is described in terms of learning the details of the hand movements, due to increased vigilance and memory resources.

created whenever a significant variation (i.e., event) in one sensor has been detected and there is no other cell assembly which can accommodate the novel information perceived by the system.

At each developmental stage, the assumptions described in Fig. 13.6 constrain the means by which the infant robot extracts and remembers information. The following assumptions are integrated: (1) static objects situated at distinct locations are mapped into distinct categories; (2) objects or events that have a common neural structure (i.e., are co-located) can become the subject of generalization. The logic of the second assumption is

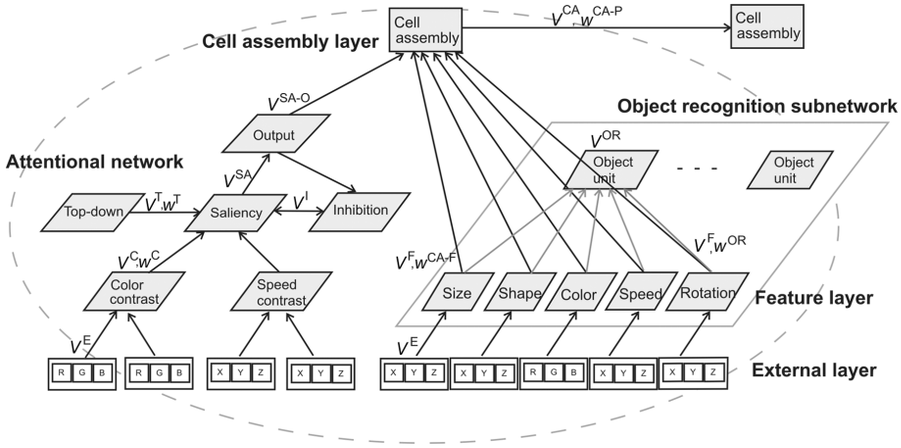

Figure 13.7 Basic computational unit of the hierarchical network consisting of three interconnected components: an attention network SA, an object recognition subnetwork OR, and a cell assembly layer CA. The saliency signal of the attention network is computed by integrating feature contrast with top–down cues. Each feature contrast unit receives input from a pair of three external units, corresponding to the object and to the context surrounding the object. An external unit encodes the color components (R, G, B) or the three-dimensional projections on X, Y, Z of the object feature. The activity of a salient winning unit is modulated by inhibition of return. An object recognition unit is fed from five feature units. An object recognition subnetwork is formed from OR units corresponding to all co-located objects. The cell assembly unit receives a saliency output signal and has a number of feature constraints grounded in the visual layer. Lateral links in the cell assembly layer learn precedence relations between cell assemblies' activations.

simple: if two objects are co-located (i.e., two embedded cups) their visual neural representations are partially shared, and generalization over their subsymbolic feature representation is favored. Similarly, when different objects are held by the hand, the shared neural representation of the hand allows generalization over the feature representations of the held objects.

According to the individuation-by-location hypothesis, a small number of attended objects may be indexed in time, the indexed individuals tracked through time and space, and the spatial relations among indexed individuals, represented (Carey and Xu, 2001). These indexes depend upon spatiotemporal information in order to remain assigned to individuals. In our implementation, each object initially located on a distinct location is mapped into a distinct cell assembly. The spatiotemporal individuation hypothesis is responsible mainly for the modeling of youngest infants' behavior (11- to 18-month-olds) (Fig. 13.6).

As explained above, a central function of the cell assembly module is to map new events into existent memory states or to create a new cell assembly representation, if the distance between the external input and the memory codes is higher than the vigilance parameter. In our model, the vigilance parameter is set in such a way that

changes in the orientation and motion speed of an object are accepted as variations within the cell assembly corresponding to that object. Thus, when the object is moving, the corresponding cell assembly updates the location coordinates of the object, until this becomes occluded. Only when a cup is grasped by the hand, or when two cups are brought together, a new cell assembly is created (see the detailed description of the learning process in Section 13.4.6).

A cup becomes occluded if it is embedded in a larger cup. In this case, a decaying memory of the cup's existence at that location is preserved in the system. There is experimental evidence that the objects' indexes can be placed in the short-term memory, and that infants can create and store more than one memory model of sets of objects, and can compare them numerically in memory (Carey and Xu, 2001). With the increase in the developmental age of the robot infant, several cup representations can be stored in memory at the same location in space, and can be compared with respect to their feature properties (i.e., size, shape).

13.4.4 The attention module

Development of goal-directed imitation and object manipulation skill is supported by selective and joint attention mechanisms. The function of the visual attention module is to direct gaze towards objects of interest in the environment. A two-component framework for attention deployment has been implemented, inspired by recent research in modeling of visual attention (Itti and Koch, 2001; see also Itti and Arbib, this volume). *Bottom–up attention* is computed in a preattentive manner across the entire visual image. It has been suggested that the contrast of the features with respect to the contextual surround, rather than the absolute values of the features, drives bottom–up attention (Nothdurft, 2000). In this model, saliency is computed based on the linear integration of contrast of two features: color and motion. *Top–down attention* is deliberate and more powerful in directing attention. The robot has a pre-wired capacity to follow the gaze of another agent and to recognize the skin color of the hand. Skin color preference is used as an indicator of where are located the hands of the demonstrator. The weights of bottom–up and top–down constraints are set to satisfy a set of attention constraints, described in Table 13.2.

A two-dimensional saliency map is used to control the deployment of attention on the basis of bottom–up saliency and top–down cues. 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).

Fig. 13.8 illustrates the functioning of the attention mechanism. Figure 13.8a illustrates the case when the imitator focuses its attention on the demonstrator's hands. Fig. 13.8b shows the time evolution of the saliency map output vs. inhibition, corresponding to the locations of the hands. After shutting down the salient unit, the inhibitory unit preserves

Table 13.2 *Visual attention constraints*

Skin color preference For any static scene, the bottom–up saliency of the hand should be higher than that of any object.
Preference for moving stimuli For any moving object, its bottom–up saliency should be higher than that of any static object, including the hands.
Motion versus skin color preference Saliency of a moving object should be higher than that of a hand moving at a slower speed, but smaller than the saliency of a hand moving at comparable speed.
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 the focus.
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 hand located outside the focus.
Gaze following versus moving hand 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.

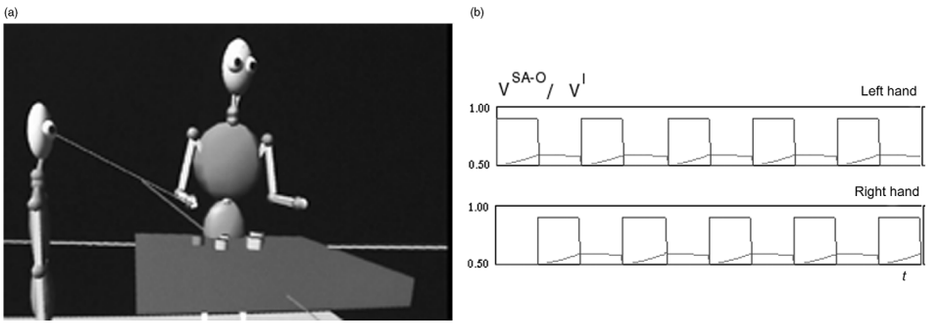

Figure 13.8 Operation of the attention module is illustrated for the case when the learner's focus of attention is driven by bottom–up attention constraints only (i.e., in the absence of the gaze signal). (a) The deployment of the focus of attention on one of the demonstrator's hands. (b) The time course of saliency output vs. the level of inhibition activation, corresponding to the shift of focus between the two hands. Inhibition regularly shunts down the saliency unit, allowing the shift of attention.

a memory of its activation, which decays in time and allows the unit to win again further in future. Different locations can be attended to due to the inhibition mechanism.

13.4.5 The cell assembly module

The cell assembly module consists of the object recognition (OR) network and the cell assembly CA layer (Fig. 13.7). The OR module implements a visual memory function

on a short timescale for the representation of external objects. Each object is represented by its decomposition in five features: color, shape, size, rotation angle, and motion speed. Each feature unit receives input from three external units corresponding to the values read from the robot sensors and projected on the components X, Y, Z. For the detection of an event, the signal variation of a feature unit's activation is integrated over time and compared with a positive, arbitrary set threshold.

The learning algorithm consists of four main steps: (a) deployment of the focus of attention at the most salient location; (b) detection of new events occurring in the focused area; (c) search for the internal representation that matches the new event; and (d) adaptation of the weights as a function of the learning rules. If the best matching cell assembly does not satisfy the vigilance threshold, that is, if novelty cannot be accommodated to the knowledge retrieved from the memory, a new cell assembly is created for the corresponding category.

During learning, the activity $V_i^{\mathrm{CA}}(t)$ of the cell assembly is a function f of the degree of satisfaction of the feature constraints $S_i^{\mathrm{CA-F}}(t)$ between the cell assembly (CA) and the feature (F) layer and of the memory of its previous activation:

$$V_i^{\mathrm{CA}}(t) = f\Big(S_i^{\mathrm{CA-F}}(t), \tau_i \cdot V_i^{\mathrm{CA}}(t-1)\Big) \tag{13.1}$$

where τ_i is the time decay rate of unit i. For simplicity of presentation we will give for the following equations only the parameters that affect the state of the variable. The function which computes the result will be generically noted with f or s. Satisfaction of the feature constraints, $w_{ji}^{\mathrm{CA-F}}$, with $j \in F_i^{\mathrm{CA}}$, depends on the current state of the external environment E and on whether the cell assembly i is in the focus of attention as measured by the saliency output $V_i^{SA-\mathrm{O}}(t)$:

$$S_i^{\mathrm{CA-F}}(t) = s\Big(\sum_{j \in F_i^{CA}} w_{ji}^{\mathrm{CA-F}} \cdot V_i^{\mathrm{F}}(t), E, \theta_s, V_i^{SA-\mathrm{O}}(t)\Big) \tag{13.2}$$

where E is computed as a function of the output of the object recognition network $V_j^{OR}(t)$ and θ_s is an arbitrarily set threshold. A cell assembly becomes unsatisfied if either the difference between its feature constraints $\sum_{j \in F_i^{\mathrm{CA}}} w_{ji}^{\mathrm{CA-F}} \cdot V_j^{\mathrm{F}}(t)$ and E is higher than θ_s or if $V_i^{SA-\mathrm{O}}(t) = 0$. Thus, the satisfaction degree is a positive, symmetric measure of the distance between the cell assembly's feature constraints and the current state of the environment E.

During retrieval, the activation of a cell assembly depends on the satisfaction of feature and precedence constraints. The activity of CA_i during retrieval is given by

$$V_i^{\mathrm{CA}}(t) = f\Big(S_i^{\mathrm{CA-F}}(t), S_i^{\mathrm{CA-P}}(t), \tau_i \cdot V_i^{\mathrm{CA}}(t-1)\Big) \tag{13.3}$$

where $S_i^{\mathrm{CA-P}}(t)$ represents the level of satisfaction of precedence (P) constraints. Satisfaction of the precedence constraints is defined as a function of the saliency output and of the summed activation of all predecessor units in P_i^{CA}

$$S_i^{\mathrm{CA-P}}(t) = s\Big(\sum_{j \in P_i^{\mathrm{CA}}} w_{ji}^{\mathrm{CA-P}} \cdot V_j^{\mathrm{CA}}(t), \Theta_i^{\mathrm{CA-P}}, \mathrm{V}_i^{\mathrm{SA}-O}(t)\Big) \tag{13.4}$$

where $w_{ji}^{\mathrm{CA-P}}$ are the weights of precedence links, and $\Theta_i^{\mathrm{CA-P}}$ is the precedence threshold for the CA_i. Precedence is met by ensuring that the threshold $\Theta_i^{\mathrm{CA-P}}$ is reached only by summing up the inputs in the order they have been learned (i.e., a strong connection requires a high input activation value in order to trigger an output). There is a supplementary condition which requires that only satisfied cell assemblies can activate their successors. This is necessary, in order to force the system to act externally (as opposed to an internal simulation of action) towards minimizing the distance between its goals and the current state of the external world.

13.4.6 Learning the seriated nesting cups task

Demonstration of the nesting cups task consists in the seriation of four cups by the demonstrator agent, using the subassembly strategy. Cups are placed as follows: cup1 (the smallest) in front of the demonstrator and cup2 to cup5 are arranged anticlockwise starting from the first cup. The demonstration order is cup1→cup2, cup1+cup2→cup3, cup1+cup2+cup3→cup4. For each cup the proximal hand is used to grasp and carry it.

During the first developmental stages, learning is driven by a basic-categorization process. The width of the categories is given by the value of the vigilance threshold ρ. A low value causes the formation of general categories, by forcing new states to be mapped into previously built cell assemblies. Each object initially located on a distinct location is mapped into a distinct cell assembly (see Sections 13.4.2. and 13.4.3). The vigilance threshold is set up based on the correlation matrix computed between the feature representations of all the objects in the system in all the possible states. The state of an object or end-effector changes as a function of the values of the following properties: orientation (i.e., rotation angle), motion (i.e., speed value), relation to the hand (i.e., whether it is held in the demonstrator's hand or not), and relation to another object (i.e., whether the objects are situated at the same location). The vigilance parameter is set in such a way that cell assemblies are resistant to variation on speed and rotation features. When the hand and an object or two objects are brought together, a new cell assembly is created. Co-located objects and end-effectors are subject to generalization and formation of more general categories.

The cell assemblies formed with a vigilance threshold $\rho = 0.25$ are shown in Fig. 13.9. Whether during each demonstration action, the system extracts specific goals of the type: "hand grasps cup1" and "hand places cup1 into cup2", until the end of the demonstration, these are gradually refined to most general categories: "hand manipulates cup" and "place cup into cup." At the end of the demonstration, "manipulates" stands for "grasp" and "carry," "hand" can be any of the hands, and "cup" stands for any of the *five* action cups.

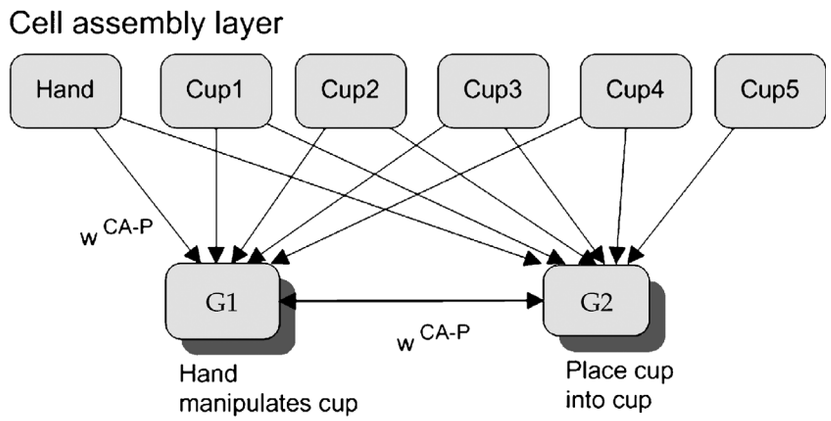

Figure 13.9 Structure of cell assemblies developed by the system during the simulation of the earliest developmental stages of the nesting cups task. Six basic categories are formed, corresponding to the five cups and hand. Distinct cell assemblies are created for all objects and end-effectors that are initially perceived at distinct locations. Each of these cell assemblies is resistant to variations of the speed and rotation features. Two more cell assemblies are formed, corresponding to "hand manipulates cup" and "place cup into cup." We refer to these cell assemblies as *hidden*, as opposed to *visible* cell assemblies, because at the initiation of imitation they do not correspond to any state in the external world. Each hidden cell assembly learns a set of predecessors that can trigger its activation. If the precedence constraints are met over a certain threshold, the hidden cell assembly becomes a goal.

Feature weights learning

At the creation of a cell assembly CA_i, all the weights w_{ji}^{CA-F} received from the feature units $j \in F_i^{CA}$ are initialized to the values of the weights of the object recognition network. The object recognition network is locked to the object's position while this is tracked in time and space. If two objects are brought together, their OR subnetworks are updated correspondingly. A satisfied cell assembly (i.e., a cell assembly whose features satisfy the current state of the external world above a certain threshold) learns the distribution of the sensorial features that constitute the objects existent at the cell assembly location. The satisfaction condition ensures that the cell assemblies remain distinct and do not end up by storing the same representation. During the demonstration of the task, the weights of the satisfied cell assemblies are subject to a Hebbian adaptation rule, and converge to the last values $V_j^F(t)$ perceived on feature units j. A better solution would be to adapt weights such as to converge to a mean value of the values received from the OR network during demonstration.

Correlation weights learning

The imitative behavior of children indicates that infants as young as 12 months of age possess a size concept that is manifested in the capability to form simple nested structures.

At this age, children operate as though size is a binary concept, with one cup treated as the "biggest" while all the others belong to the category "little". Only infants of 28 to 36 months old seem to follow the size criteria for all structures formed.

We propose here a solution for how one can simulate the gradual development of the size concept, grounded on the interaction of the artificial system with the external world. The solution is based on the system's ability to learn invariant relationships between the features that constitute specific objects. Correlation weights w_{ji}^{CA-R} between the features of two objects j and i are adapted using an anti-Hebbian learning rule, which weakens the strength of the connection when only pre- or postsynaptic neurons are activated. Invariant features or unsystematic variations of the feature values cause weights to decay. Only systematic variations, such as the relation between the sizes of the acting and recipient cup, lead to an increase of the weights.

When the hand and an object or two objects are brought together for the first time, a new cell assembly is created for each of these external states. The resulted internal representation should persist long enough to permit learning of the relation existent between the features of the co-located objects. In our model, the external states corresponding to "place cup into cup" and "hand manipulates cup" receive a top–down attention signal, which preserves focus of attention on the corresponding objects and allows the internal comparison of their representations.

Figure 13.10 shows the time evolution of the correlation weights corresponding to the goals extracted during the demonstration of the nested cups task.

Precedence weights learning

With the creation of new cell assemblies, early cell assemblies tend to satisfy in a smaller manner the current state of the environment, and have a lower level of activity. A satisfied cell assembly CA_i learns a set of precedence links w_{ji}^{CA-R} from other activated cell assemblies CA_j. Precedence in the system is encoded in the relative order between the construction and satisfaction of the cell assemblies. Temporal learning between co-activated cell assemblies is favored by their graded activation as a function of the satisfaction level. The precedence links learned during the first developmental stage are shown in Fig. 13.9.

13.5 Reproduction of the first developmental stages of nesting cups behavior

Modeling of the developmental path of the ability to seriate cups necessitates the reproduction of two behavioral manifestations. First, the model has to account for the systematic differences existing between the infants' strategies. We consider as responsible for this, the learning process whose parameters are developmentally constrained and which leads to the construction of specific internal models. Second, the model must replicate the variety of imitative behaviors characteristic to human infants. This is achieved through a process of probabilistic satisfaction of multiple types of constraints.

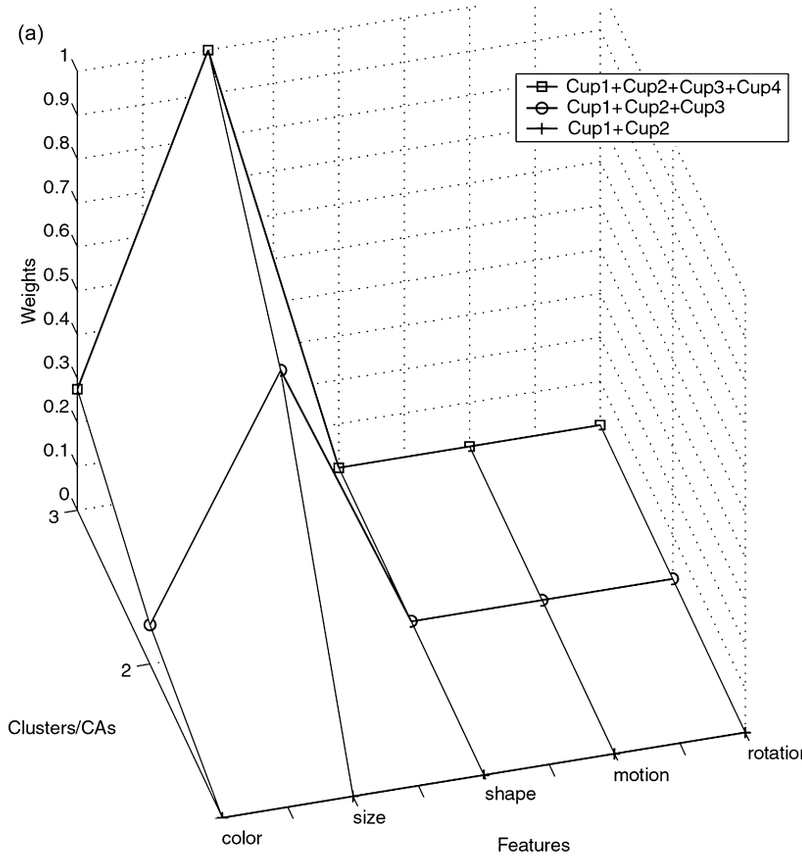

Figure 13.10 Evolution of the correlation weights corresponding to different clusters of sensorial features, during the demonstration of the nested cups task. (a) Shown on the *z*-axis are correlation weights for successive states when two or more cups are embedded (i.e., from state 1 to 3 on the *y*-axis). The size weight increases from 0 in the state "cup1 into cup2" to a maximal value in state "cup3 into cup4." This increase of the weight reflects the systematic relationship existent between the sizes of the acting and the recipient cup. (b) Correlation weights on the *z*-axis corresponding to the states when the hand grasps a cup (i.e., from 1 to 3 on *y*-axis). At the end of the demonstration, the weight values corresponding to size, shape and rotation features are maximal, indicating the existence of an invariant relationship between the hand and the acting cup for each of these features. These constraints will be integrated at retrieval during the third developmental stage.

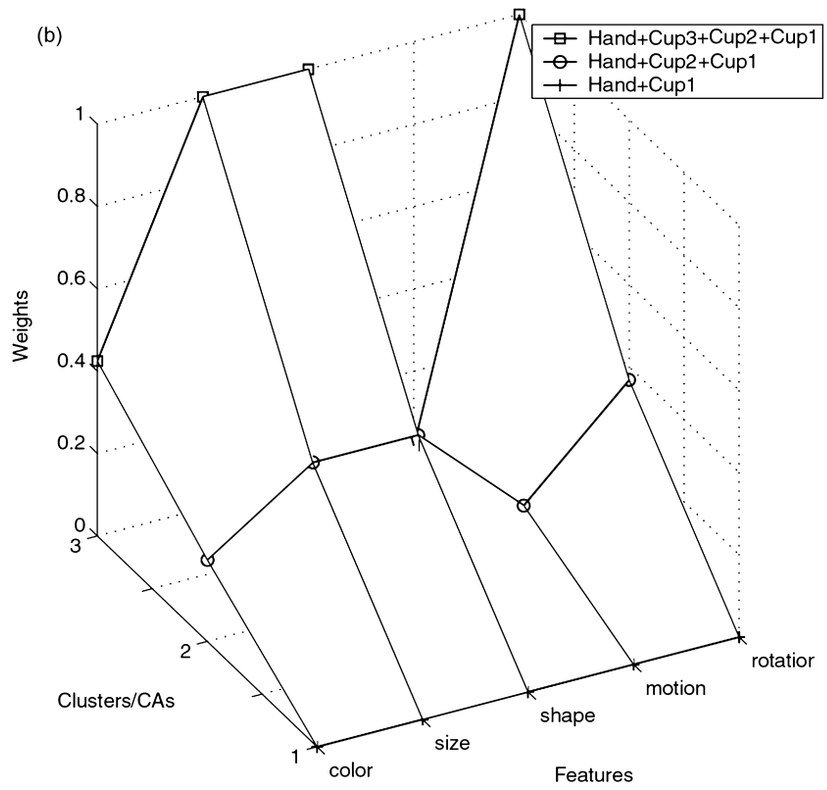

Figure 13.10 (cont.)

During retrieval, three types of constraints are operating. The *conservation* C constraint reflects the tendency of the system to minimize the amount of effort during the execution of the task. Its operation is reflected in the choice of proximal cups and in the system's preference to rest in a state with minimal energy (i.e., where all goals are satisfied). The operation of the *saliency* (SA) constraint is reflected in the choice of objects with a high color or motion contrast. Finally, *size* (S) constraints are satisfied when the imitator embeds a smaller cup within a larger cup. The size constraints are learned in the set of weights $W_{ji}^{\mathrm{CA-R}}$, while saliency constraints are applied as a function of the attention module's output $V^{\mathrm{SA-O}}(t)$. The conservation constraints C are applied: (a) by computing the physical distance between the hand and a given object; and (b) by ending the imitative behavior as soon as all the system goals are satisfied.

During imitation, the actions of the agent are driven by its internal model, more specifically, by those goals which are not yet satisfied. The internal model has the role of activating and deactivating the goal states: "hand manipulates cup" (G_1) and "place cup into cup" (G_2). Activation of a goal occurs when the precedence constraints are met and deactivation of the goal takes places when its feature constraints are maximally

satisfied. An activated goal drives the action A of the agent, by setting the type of action, the final state, and the category of objects which is subject of action. Goals are encoded by categorical cell assemblies, hence, they stand for classes of objects, rather than specific objects.

As shown in Fig. 13.9, the hidden cell assemblies can be activated by the active representations of the five cups and the end-effectors. Visible cell assemblies become active as a matter of the bottom–up saliency signal received from the attention module $V_i^{\mathrm{SA-O}}$(see 13.2 and 13.3). During imitation, the attention of the agent is driven by the same attention constraints as during the demonstration (see Section 13.4.4), excepting the existence of demonstrator's gaze signal. The inhibition mechanism permits the switch of the attention's focus between end-effectors and all objects in the environment. As soon as an object enters the focus of attention, its cell assembly becomes satisfied and can compete for being chosen as target of action. The choice of the acting, respectively the recipient cup, results from a process of multiple constraints satisfaction (i.e., size, saliency, and conservation constraints).

The behavior of the agent results from solving the set of equations corresponding to the satisfaction of all types of constraints existent in the system. If we consider that the satisfaction of a goal yields a certain adaptive value $\Gamma(G)$ for the agent, then, at any moment in time, the goal G_i of the system is given by the cell assembly i with $\Gamma(G_i) > \Gamma(G_j)$, $\forall j \in \mathrm{CA}$, which meets the conditions: the cell assembly is not satisfied $S_i^{\mathrm{CA-F}}(t) = 0$, precedence constraints are met over a given threshold $S_i^{\mathrm{CA-P}}(t) > \Theta_i^{\mathrm{CA-P}}$, and all its predecessors are satisfied $S_j^{CA-\mathrm{F}}(t) > 0, \forall j \in P_i^{CA}$.

The activated goal sets the type of the action (i.e., grasp or move), which is executed by the system through a process of successive operations aimed at minimizing the distance between the current state of the world and the desired state corresponding to the goal. Elsewhere (Goga and Billard, 2004) we discussed how the system could learn the sequence of actions required to grasp the object (i.e., rotate the end-effector, move it towards the object, lift the object). Here, we follow a similar approach, based on the computation of the difference between the current and the desired value for each feature, and sending an appropriate command to the robot's actuators for the minimization of this distance.

The acting cup i and recipient cup j are set as a function of the maximal consonance computed over all constraints in the system, $\max_{i,j} \chi_{ij}(\sum_{k=1}^{3} p_k \cdot C_k)$ where p_k are the probabilities of size, saliency and conservation constraints C_k and χ_{ij} is the consonance computed between objects i and j (see Section 13.5.2). The agent has a pre-wired drive towards the maximization of the internal consonance computed. The system also possesses a regulating mechanism in order to decide when to terminate behavior. The infant robot stops imitating when the goal G_i with the highest value $\Gamma(G_i)$ is maximally satisfied.

In the following, we present several behavioral scenarios, corresponding to the probabilistic combination of the criteria for choosing the acting and recipient cups: size, saliency, and conservation of energy. Behavior is consistent for any one setting of the constraints.

13.5.1 Saliency constrained behavior

In the first behavioral scenario, constraints are applied as follows. The acting cup is set to the first cup that wins the focus of attention, which is the most salient object in the environment (i.e., has the highest color contrast). For all satisfied cell assemblies, size (S) and conservation (C) constraints are computed, and the recipient cup is set to the first cup that satisfies an arbitrary combination of these. Figure 13.11 shows different imitative behaviors, corresponding to different settings of the probabilities of size and conservation constraints. In Fig. 13.11a the most salient cup becomes the acting cup and minimization of the path is applied, leading to the formation of one pair. In Fig. 13.11b, after the acting cup is chosen, size constraints are computed for the cups being in the focus of attention (see Section 13.5.2 on how size consonance is computed), and the cup is nested in a larger cup. In both scenarios, after the acting cup is chosen, the hidden cell assembly corresponding to "hand manipulates cup" becomes the goal of the system, and triggers the set of

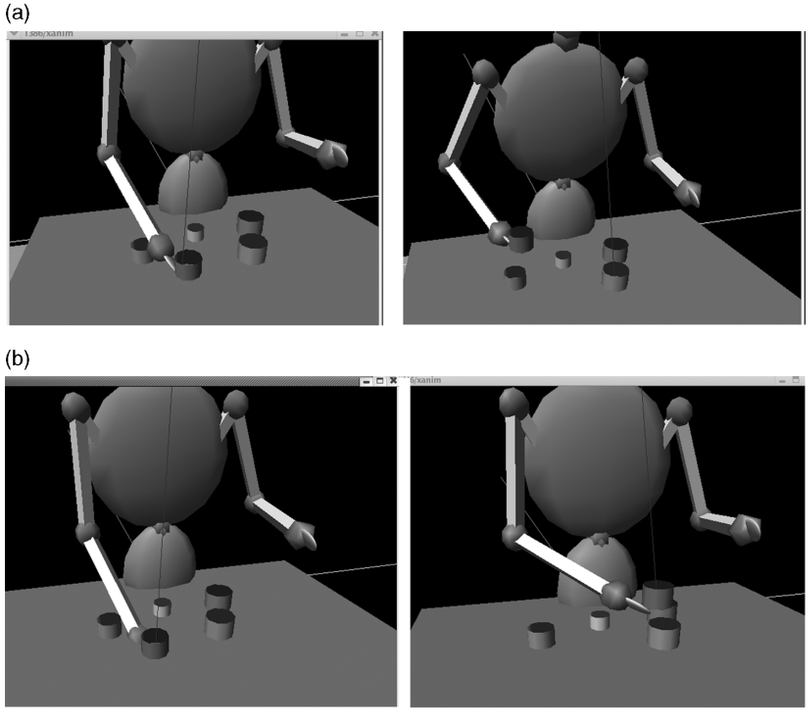

Figure 13.11 The behavior of the simulated agent, when the system is acting under the primacy of saliency constraints. (a) The most salient object (cup3) wins the focus of attention and becomes the acting cup. The recipient cup (cup2) is chosen in order to minimize the path. (b) The most salient object (cup3) becomes the acting cup and the recipient cup is chosen after size consonance is computed for the satisfied cell assemblies.

actions required for its achievement. The robot's actuators receive the coordinates of the acting and target cup, and the end-effector is moved to the desired location through the inverse kinematics and dynamic motor servos.

13.5.2 Size constrained behavior

When the probability of the size constraint is higher than those of conservation and saliency constraints, the system chooses the acting and the recipient cups as a function of the global consonance χ computed based on the internal correlation constraints. The term *consonance* is used to reflect the fact that the behavior of the agent is constrained by its internal model on how different objects can be combined together. The consonance between two objects i and j is obtained through the summation over all features $k \in \mathrm{F}$ of the products between the correlation weight $w_{ji}^{\mathrm{CA-R}}$ and the difference between the activation values on the corresponding feature:

$$\chi_{ji} = \sum_{k \in \mathrm{F}} w_{ji}^{\mathrm{CA-R}} \cdot (V_j^k - V_i^k) \tag{13.5}$$

Consonance is computed as a function of all relation weights developed. The first two developmental stages are characterized by top–down attention biases towards the extraction of relational information concerning the size feature of objects (Fig. 13.10a). In this respect, we describe the corresponding behavior as being size constrained.

The system acts towards the maximization of the global consonance. For each object i its consonance is computed with all objects j corresponding to cell assemblies which are currently satisfied. For the pairing and pot strategy model, the maximal consonance for any object is given by the product between the size relation weight (i.e., highest weight) and the largest size difference between the compared objects. The global consonance χ is given by the maximum of all consonances computed.

Figure 13.12 illustrates two behavioral scenarios corresponding to the seriation of different cups. Note that only satisfied cell assemblies can be internally compared in the process of global consonance computation. In Fig. 13.12a, the smallest cup is placed into the next largest cup. The behavior corresponds to the computation of global size consonance for all demonstrated cups (the largest cup was not used during the demonstration) and the acting and target cups were chosen to maximize this value. Fig. 13.12b illustrates the case when internal consonance is computed for only two cups and the choice of the acting and recipient cups reflects the local maximum found. The two cups correspond to the first objects that attract the agent's focus of attention. Initiation of action depends on the activation of the internal goals. In this case, the satisfaction of the cups' cell assemblies is sufficient for the activation of the "hand manipulates cup" goal and for the triggering of first actions.

Various nesting behaviors (i.e., behaviors that satisfy size constraints, but do not maximize global consonance) can be simulated as a function of three parameters: the *ignition value*, the precedence threshold and the number of internal comparisons

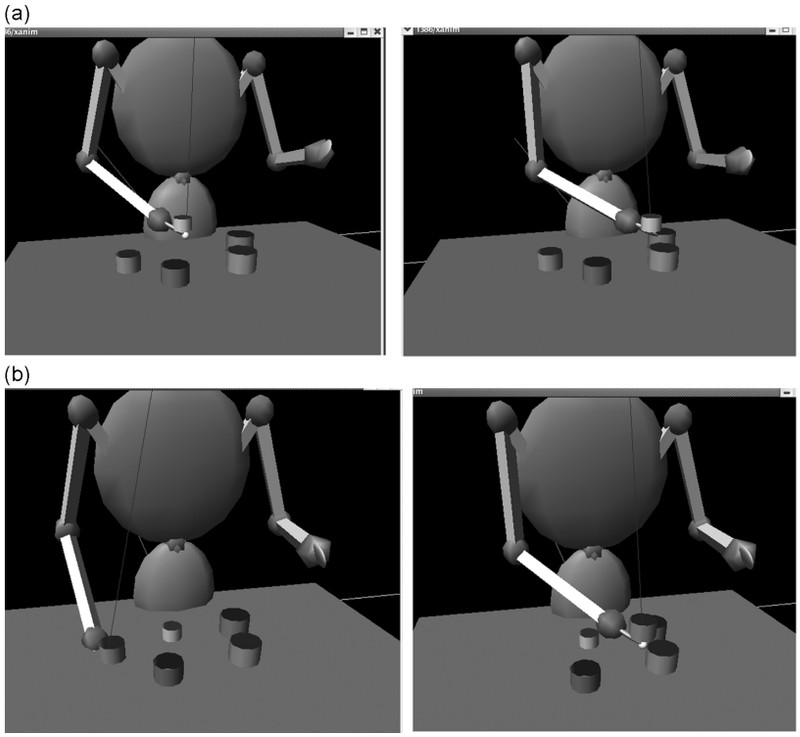

Figure 13.12 Two behavioral scenarios where action is primarily driven by the satisfaction of size constraints. (a) Internal consonance is computed for all satisfied cups (see text for explanations) and the acting and recipient cups are chosen to maximize its value. This will correspond to the movement of cup1 (the smallest cup) into cup4 (the next largest cup). (b) Internal consonance is computed for the first two cups which enter the focus of attention. The choice of the acting and recipient cups reflects the local maximum found (see text for explanations on the role of the internal parameters on the reaction time of the robot). The agent behavior is to move cup2 into cup4.

performed. The precedence threshold Θ^{CA-P} (13.4) is an internal, emergent parameter of the model, while the ignition value and the number of internal comparisons are developmentally constrained parameters. The ignition value represents the percent of precedence constraints that should be met for a goal to become activated. These parameters affect the time elapsed until the activation of the first goal of the system, and the variation of this *reaction time* affects the computation and maximization of global consonance.

13.5.3 One pair, two pairs, transfer of the cup

The model described above can replicate the behavioral manifestations characteristic to human infants using the pairing cups strategy. The simulated agent can form a nested structure or a tower, as a matter of the probability of size constraints satisfaction. The

model can also decide whether to stop the seriation behavior after the completion of a pair or to continue it. The robot infant can form one, two, or several pairs, as a function of the termination condition imposed.

Let us analyze the state of the system when the hand carrying a cup reaches the target location. At this moment, both goals, G_1 corresponding to "hand manipulates cup" and G_2 corresponding to "place cup into cup," can be satisfied. If we consider that satisfaction of a goal yields an adaptive value $\Gamma(G)$ for the agent, two situations can occur:

- If $\Gamma(G_1) > \Gamma(G_2)$, that is, if grasping the cup is more important than forming a nested cups structure, the system will act to maximize the satisfaction of G_1. In this case, the cup is held in the hand and G_2 does not become satisfied. It corresponds to an imitative behavior where one cup is transferred from one target position to another. Continuation of behavior is assured by the non-satisfaction of the G_2 goal. This behavior is illustrated in Fig. 13.13.
- If $\Gamma(G_2) > \Gamma(G_1)$, the system will act towards the maximization of the satisfaction of G_2. Accordingly, the cup will be dropped at the target location. The system stops if the pair formed satisfies the goal G_2, or it continues if the pair does not satisfy the features of G_2 (i.e., if the formed pair has never been perceived before) (Fig. 13.13).

13.5.4 The pot strategy

Hypothetically, the transition from the pairing to the pot strategy can be modeled in two ways: (1) by increasing the number of goals that are learned during the demonstration of the task (i.e., the formation of a sequence of subgoals corresponding to the embedding of several cups into the pot cup); or (2) by preserving the main structure of the internal model, and increasing the information stored within each goal representation. The second alternative comes more naturally within our developmental model, which acts towards the reduction of the amount of information stored in the memory.

We consider that during the second stage of development, cognitive resources (i.e., attention and memory) are employed to extract more information concerning the goal "place cup into cup." This leads to an increased capacity to store and retrieve the information concerning the number of cups that are embedded during demonstration. In the cell assembly structure, this acquisition is reflected in the specialization of the state goal "place cup into cup," by storing a symbolic representation of all the cups that can be embedded. The behavior of the agent evolved in this second developmental stage towards the maximization of the size consonance, computed for all the objects stored in the representation of the state goal "place cup into cup." With each new pair formed, the global consonance increases, and makes less probable such behaviors as, the transfer of one cup through several positions.

Various behaviors can be modeled through the manipulation of the factors mentioned above (i.e., precedence threshold, number of internal comparisons) and as a matter of how global consonance is computed: a tower vs. a nest; a pot containing all cups vs. two or three cups nested. Figure 13.14 illustrates an imitative behavior corresponding to the nesting of three cups using the pot strategy.

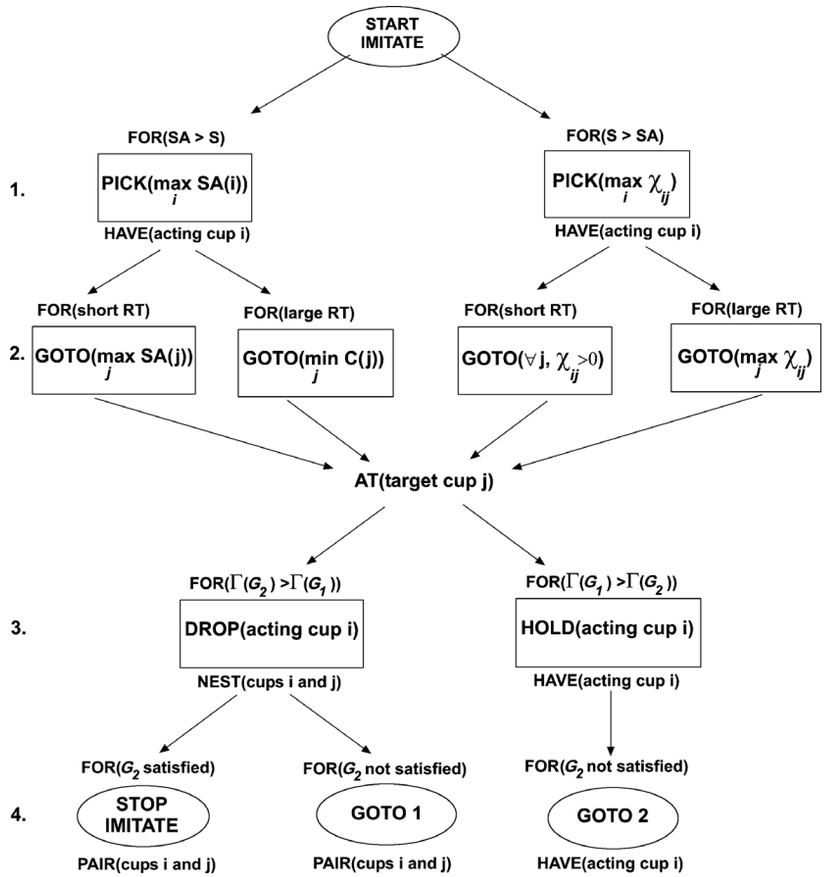

Figure 13.13 Planning graph showing the states of the system during the reconstruction of the demonstrated actions and the parameters that determine the transition from one state to another. At top of each state, the precondition is shown (i.e., FOR) and at bottom the post condition (i.e., HAVE). Actions at level 1 correspond to the choice of the acting cup as a function of the saliency (SA) and size constraints (S) applied. Actions at level 2 describe the criteria considered in the choice of the target cup. The system can react faster (short reaction time RT) or slower (large RT) and conservation (C) constraints can also be integrated. If the adaptive value of goal G_2 "place cup into cup" is higher then that of goal G_1, the system forms one pair (level 3), after which it can continue with the pairing behavior or stop the imitation (level 4). If the adaptive value of goal G_1, "hand manipulates cup" is higher then that of goal G_2, the system holds the acting cup, and the resulting behavior corresponds to the transfer one cup from a target position to another.

Figure 13.14 Imitative behavior of the simulated agent corresponding to the formation of a pot at the position of cup4. The first cup embedded is cup2, followed by the smallest cup from position 1, followed by the cup from position 3. Size constraints are satisfied with respect to the acting cup and the recipient pot cup. A pile may result (not shown), because consonance is computed between visible cups, and the size of the already nested cups is not taken into account.

13.6 From nesting cups model to an action–language model

Our task now is to provide a computational framework that extends the action system towards the integration of a language system, and to discuss the possibility that previously developed structures for goal-directed imitation and object manipulation can provide a substrate for language bootstrapping. The working hypothesis is that the capacity for language and the ability to combine objects and to imitate novel behavior follow a common developmental path, and may share a common neural architecture, where they influence each other.

13.7 Lexical grounding on subsymbolic knowledge

An action–language computational model has at its foundation a mechanism for grounding the conceptual information on subsymbolic structures. Our work builds on previous modeling attempts on the world-to-word mapping.

Regier (1995) described a system capable of learning the meaning of spatial relations such as "on" and "under" from pictorial examples, each labeled with a spatial relation between two of the objects in the scene. Learning spatial knowledge was based on a tripartite trajectory representation of type *source–path–destination*, aimed at grasping the event logic in both motion and language. Bailey (1997) developed the VerbLearn model, where actions are represented using executing schemas, while words are encoded using structures of features, given by the parameters of the motor system: acceleration, posture/shape of the hand, elbow joints, target object. The interface between language and action levels is played by a linking feature structure, which binds the words bidirectionally with the actions. Siskind (1995, 2001) proposed a conceptual framework for grounding the semantics of events for verb learning, using visual primitives which encode notions of support, contact, and attachment. 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 previous work on learning a synthetic protolanguage in an autonomous robot (Billard, 2002) used a time-delay associative network, consisting of a Willshaw network (Willshaw *et al*., 1969), to which self-recurrent connections have been added, to provide a short-term memory of activation of units. Sensor and actuator information is memorized for a fixed duration to allow association to be made between time-delayed presentations of two inputs. For the work described here, we used a similar architecture consisting of two time-delay networks, one feeding in sensorimotor information and the other linguistic input, connected through a set of bidirectional, associative links. The model takes as input pictures of the environment (i.e., an agent demonstrates how an object can be moved on a table) and short sentences about them. The aim is to ground, using the set of associative, bidirectional links, the meaning of a small lexicon in the structure of sensorimotor features available to the simulated agent.

There are a number of challenges that a system faces in learning word meanings. Among these are: (a) multi-word utterances, where the agent has to figure out which words are responsible for which parts of the meaning, and (b) bootstrapping, i.e., when first words are learned, the system cannot use the meaning of words in an utterance to narrow down the range of possible interpretations. To solve the former problem, we followed the approach described in Siskind (1996) by applying cross-situational learning to the symbols that make up the meaning of a word. Cross-situational learning refers to the idea that children narrow the meaning of words by retaining only those elements of meaning that are consistently plausible across all the situations in which the word is used.

Hebbian and anti-Hebbian learning rules adapt the associative links between the sensorial and linguistic time-delay layers, in such a way that each word develops a set of associative links with the feature units. Words can compete for labeling an associated feature. The winner is the unit with the highest connection strength that passes an arbitrary confidence threshold. Autonomous bootstrapping, consisting in extracting tiny bits of linguistic knowledge and using them for further analysis of the inputs, is applied to simplify and urge the learning process.

As a result of the specifics of Hebbian learning, frequently used words (e.g., the verb "move") develop stronger weights, and the systematic association between a word and a set of features is reflected in the development of a rich and strong set of feature weights. Anti-Hebbian learning is meant to decrease the weights for variant, unsystematic associations between a word and world features (e.g., for the word "look").

13.7.1 Attention-directing interactions during the seriated cups task

The preliminary experiments described above have been conducted with an artificially generated lexicon. The next step in the development of the action–language computational framework will be to train the model with input from natural language. We are currently running a set of experimental studies for the systematic observation of the interaction between a human caregiver and the child during the seriated nesting cups task (see Fig. 13.15). In this respect, the original experiment of the seriated nesting cups task is modified in several ways. The demonstrator is replaced by the child caregiver, who interacts with the infant in three different conditions. During the first condition the infant and the caregiver are playing with several toys and nested cups. The second condition is similar to the experimental setting of the original seriated nesting cups task (Fig. 13.15a). The caregiver demonstrates the seriation task and also interacts with the child during the imitation phase, by directing her attention and providing feedback on the success of the task. In the third condition, the caregiver is allowed to freely interact with the infant during the demonstration and imitation periods, using linguistic imperatives and gestures, in order to help the infant build the seriated structure using the most advanced strategy (Fig. 13.16b). Infants between 12 and 36 months of age are observed and all sessions are recorded on video and audio.

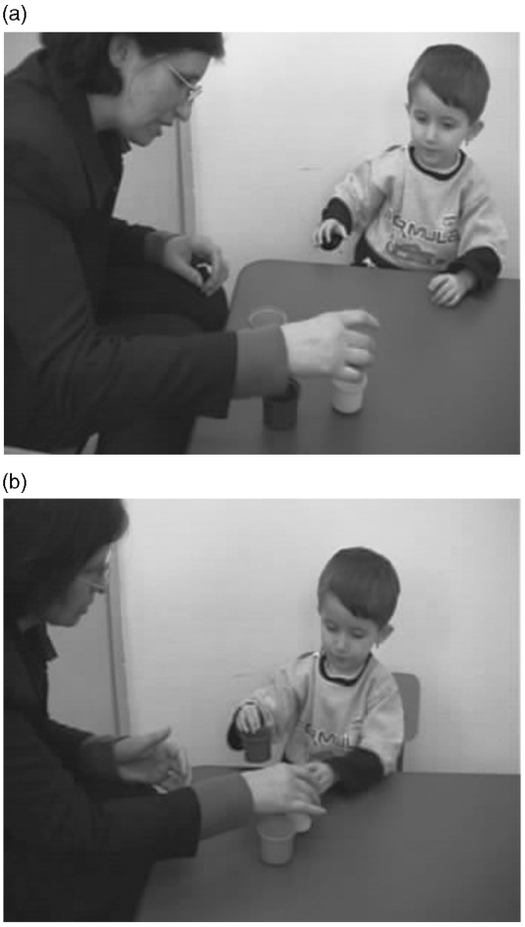

Figure 13.15 Human child–caregiver experimental setting. (a) The caregiver demonstrates the seriated nesting cups task. (b) The child imitates the caregiver, while the latter/guides the child's attention using linguistic cues and gestures.

The video and audio recordings will be transcribed using CHAT system (Codes for Human Analysis of Transcripts) (MacWhinney, 2000). CHAT was developed and used as the standard transcription system for the CHILDES project (Child Language Data Exchange System) aimed at researching infants' directed speech. The transcripts of the records will be analyzed and segmented, and relevant linguistic sequences from the human child–caregiver interaction will be selected and used as training input for an action–language computational model. We envisage the existence of two stages in the development of the linguistic function in robot infants. During the first stage, the model will be trained with simple descriptive sentences in order to learn the meaning of a basic set of words. This will enable the imitator to communicate with the demonstrator during the seriated cups task.

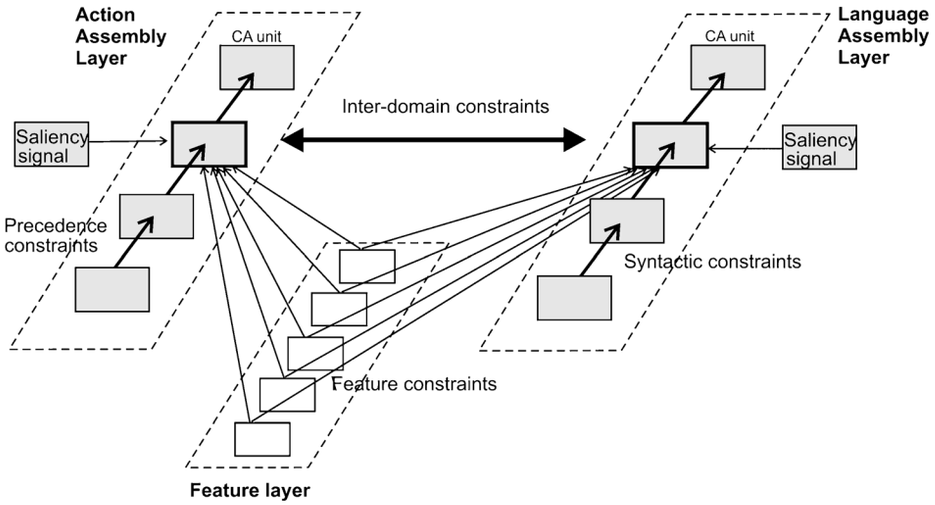

Figure 13.16 Integration of the language cell assembly layer within the action model developed for the simulation of the seriated nesting cups task. Each cell assembly layer is grounded on the internal mapping of the external world from the object recognition network. For simplicity of presentation, only the connectivity of one cell in each layer is shown. The activity of the action and linguistic cell assembly is modulated by the saliency signal received from the attention module (not shown in figure). The action and language networks are connected through a set of bidirectional, associative links, subject to learning through asymmetric cooperation and competition.

When the agent becomes capable of naming the objects and the events occurring in the environment, the input provided to the imitator during the demonstration task can become more complex. Relevant attention-directing sequences from the human child–caregiver interaction will be fed into the model, by the demonstrator agent. The demonstrator robot will accompany the seriation actions with linguistic cues and short descriptions of what it is doing, in order to direct the imitator's attention and to explain the goal of the demonstration. In the current model of seriated nesting cups task, the joint attention mechanism is meant to spotlight objects and events being attended to, through the integration of a number of bottom–up and top–down saliency cues (e.g., color contrast, motion contrast, gaze follow, and pointing). In the action–language model we intend to enhance the power of the joint attention mechanism, and to investigate the role of linguistic and gestural attention cues in scaffolding the robot's early language and goal-directed imitation development.

13.7.2 A computational view on the action–language reuse hypothesis

Having established a basic action and imitation system, we want to go beyond the "mirror hypothesis" and to investigate how the acquisition of language capability

may interact with previously constructed structures for goal-directed imitation and object manipulation. Further development of the action model is towards the integration of the language network within the cell assembly architecture developed for the seriated nesting cups task. The model will be trained with sequences of gestural and symbolic attention-directing interactions between the human child and caregiver, as described above.

The architecture of the envisaged action–language model is shown in Fig. 13.16. The language conceptual layer develops in a similar manner with the construction of the action cell assembly layer. The adaptation of the language network's sets of weights is governed by similar learning rules. The basic categorization process may be responsible for the mapping of words into linguistic cell assemblies. The vigilance parameter can be used to tune the generality of the linguistic categories formed. Each linguistic cell assembly is grounded in a subsymbolic sensorial representation from the object recognition network. In this way, the semantic features of a word are grounded on the feature constraints of a linguistic cell assembly. Short serial-order dependencies between words may be learned in the precedence constraints at the cell assembly level. The cell assembly layers are connected through bidirectional, associative weights, which play a major role in applying the neural patterns of activity established in the action layer to the developing processes taking place in the language layer. To give an account of how this process of interdomain bootstrapping may take place, we introduce the reader to the theory of Cortical Software Re-Use (CSRU) (Reilly, 2001; Reilly and Marian, 2002).

The CSRU theory states that the sensorimotor areas of the brain provide the computational building-blocks for higher-level functions. A computational account of how motor programs developed for object manipulation might be reused for language syntax was provided by Reilly (1997) in responding to the work of Greenfield *et al*. (1972). Reilly (1997) selected the seriated nesting cups task and an equivalent speech task and implemented them within a connectionist framework. The hierarchical structures of the motric and linguistic representations were first encoded using a recursive auto-associative memory (Pollack, 1990), then the goal representations were fed into a simple recurrent network, which had to generate the appropriate sequence of actions as output. The reusability hypothesis was consistent with the finding of a training advantage when a recurrent network was pretrained on the object assembly task prior to learning a language task.

The neural building-block proposed by CSRU theory is the *collaborative cell assembly* (CCA). Re-use is operationalized through the operation of different types of computational collaborations between cell assemblies. *Structural isomorphic reuse* occurs from exploiting approximately the same cortical region (see Greenfield's (1991) hypothesis on the action–language homology on the Broca's area). *Symmetric collaboration* occurs between equally developed cell assemblies, and employs an "indexing" operation of a neural population by another neural population (Reilly and Marian, 2002). An example can be a system that learns to foveate a visual target and then "reuse" the saccade map to

achieve ballistic reaching (Scassellati, 1999). Finally, *asymmetric collaboration* occurs where a less well-developed cell assembly exploits the functionality of a more developed one. We believe that the latter form of collaboration can play an important role in explaining the neural patterns of competition and cooperation, responsible for the integration of proprioceptive, sensorimotor and linguistic information.

Our approach to the modeling of neural grammars is inspired by the concept of *word-sensitive sequence detectors* proposed by Pulvermüller (2003). These can be assumed to operate on *pairs* of elementary units, for the detection of frequently occurring sequences of two words. Sequence detectors could operate on webs representing words from given lexical categories, and could support generalization of syntactic rules to novel word strings. For instance, separate sequence detectors could respond to pairs of the types: pronoun–verb, verb–verb suffix, and preposition–noun. By operating on short rather than long temporal dependencies, the word-sequence detectors are scalable and composable. They allow temporal processing of complex structures through the assemblage of simple sequence detectors in a hierarchical, recursive, and compositional fashion.

Operationalization of the reuse hypothesis comes naturally in our model. First, the semantics of a word is grounded in the sensorial representation of the associated object/event. This enables us to investigate the means by which word understanding and production can be affected by the operation of previously constructed sensorimotor structures. Second, the robot must process, store, and retrieve syntactical information from the transcribed sequences of the human child–caregiver interactions. These processes will take place during the simulation of a social interaction protocol, when the infant robot must also execute different sequences of movements, according to its internal model developed.

The operation of the action schema for goal-directed imitation introduces a number of constraints on learning by the robot infant how to generate correct and meaningful sentences. We intend to investigate the dynamic computational means for cooperation and competition between the neural patterns established in the action network and those developing in the language network. The goal is to study not only how the linguistic function can make use of the sequence detectors developed for the seriation ability, but also how goal-directed action can use top–down linguistic cues to discover "what" to imitate.

13.8 Discussion

In this chapter we have presented ongoing work on modeling the action–language analogy hypothesis. We have used the seriated nesting cups task (Greenfield *et al*., 1972) as a developmental framework for investigating the interaction between goal-directed imitation, object manipulation, and language acquisition. A conceptual framework drawing on neurobiological and psychological experimental data was proposed to account for the development of the goal structure that lies at the foundation of object manipulation and

language usage in infants and robots. In modeling the seriated nesting cups behavior, we investigated the effect of varying a number of parameters of the computational model, as a way of accounting for systematic differences in child behavior.

A central tenet of our model is that development of goal-directed imitation and object manipulation skill is supported by joint attention mechanisms. These mechanisms play an important role in coordinating the behavior and sharing a similar perceptual context necessary for the learner to infer the demonstrator's intentions and goals (Tomasello, 1998). During the last decade, several researchers have attempted to build mechanisms of joint attention in robots (Scassellati, 1999; Kozima and Yano, 2001; Nagai *et al*., 2003). Our approach is unique in that it focuses on the integration of multiple attention constraints in order to obtain more complex behavior for the agents.

Our model also contributes to the theory of goal-directed imitation, and the experimental data showing that infants learn *what* to imitate before learning *how* to imitate (Bekkering *et al*., 2000). Modeling of the pairing and pot strategies in seriating cups was based on the assumption that the robot infant learns first about objects, then about the relations between them, and finally about the events and movements. The internal model developed during the first stages reflects the primacy of object related information. By learning *how* to imitate, we mean learning how the goal of the imitation task (i.e., the pot of cups) is achieved using a certain strategy (i.e., subassembly strategy) by imitating the way the hand movements are sequenced relative to the acting objects.

Modeling of the seriated cups task was challenging in that the resulting model had to reproduce the consistency of infant strategic behavior, as well as the variety of human behavior. The systematic differences between infants' strategies were accounted for through a learning process whose parameters are developmentally constrained (i.e., basic categorization, spatiotemporal object representation, joint attention mechanism). In the model presented here, the basic categorization process plays a crucial role in modeling the pairing and pot strategies.

With the increasing resources available to the learning process, the system is able to extract, store, and remember more information from the sequence of demonstrated actions. As pointed out in the methodological approach outlined in Section 13.1, the system gradually develops to accommodate more complex behaviors, by integrating in a cumulative manner prior structures that provide competencies to be reused by later developed skills.

The developmental model of the seriated nesting cups task is to our knowledge the first computational proposal of this type. A central tenet of this approach is that modeling of the seriated cups task can best be addressed in a multiple constraints satisfaction framework. The variety of imitative behaviors was replicated through a process of probabilistic satisfaction of three types of constraints (conservation, size, and saliency). The choice of the constraints was inspired by the experimental observations of Greenfield and colleagues. They describe the existence of three criteria in the infants' choice of the objects for action: proximity, size, and contiguity. In our model, conservation C stands for the

satisfaction of proximity and contiguity. Saliency was introduced to account partially for the variability of the behavior observed in the group of human infants.

At this point, the model lacks the simulation of the third and most advanced sub-assembly strategy for nesting cups. We envisage modeling of this stage along the lines described above, based on several developmental assumptions: (a) the focus of attention shifts from the objects to the movements executed by the hand; (b) the increase of the vigilance parameter leads to the specialization of the goal "hand manipulates cup"; and (c) during retrieval other constraints than those given by the size weights may be taken into account in computing the internal consonance of the model (see Fig. 13.10b).

As a drawback, the model does not have a consistent way to handle noise in the environment.

13.8.1 Relevance to the Mirror Neuron System hypothesis

Computational modeling can significantly contribute to the understanding of the wider impact that the mirror neuron system has on brain computation. Evidence of the mirror neuron system and its role in driving action, language, and imitation is still new (Rizzolatti and Arbib, 1998; Iacoboni *et al*., 1999). With respect to this, computational modeling significantly contributes to the understanding of the wider impact that the mirror neuron system has on brain computation (Oztop and Arbib, 2002; Oztop *et al*., this volume). The work presented here can bring computational evidence for the hypothesis that the abilities to imitate and to manipulate objects lie at the foundation of language development in humans.

Greenfield *et al*. (1972) suggested that the internal model of seriated cups was used by the youngest children mainly to get the nesting behavior going, whereas it was used by the older children as a basis for terminating activity as well. For them, the final stage of the demonstration appeared to function as a precise goal and signal for termination. Similarly, during the first developmental stage, the robot infant extracts a generalized goal state of the form "put the cups inside each other," while during the second stage, the goal representation is restructured to "put several/all cups into the largest cup." Only during the third developmental stage is the agent capable of extracting detailed information on the order in which the hand manipulates the cups, and to infer the demonstrator's goal "put all cups in the largest cup by placing each cup into the next largest cup."

Acknowledgments

This research was carried out during the visit of Ioana (Marian) Goga to the Autonomous System Laboratory, L'Ecole Polytechnique Fédérale de Lausanne, and was supported by the Swiss National Science Foundation through grant 620-066127 of the SNF Professorships Program. The authors are very grateful to Stefan Schaal for providing access to the Xanim simulation environment for the experiments presented here.

References

Arbib, M., 2003. The evolving mirror system: a neural basis for language readiness. In M.H. Christiansen and S. Kirby (eds.) *Language Evolution: The States of the Art*. Oxford, UK: Oxford University Press, pp. 182–200.

Bailey, D.R., 1997. When push comes to shove: a computational model of the role of motor control in the acquisition of action verbs. Ph.D. dissertation, University of California, Berkeley, CA.

Bekkering, H., Wohlschlager, A., and Gattis, M., 2000. Imitation is goal-directed. *Q.J. Exp. Psychol.* **53A**: 153–164.

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*. Cambridge, MA: MIT Press, pp. 281–311.

Billard, A., and Dautenhahn, K., 2000. Experiments in social robotics: grounding and use of communication in autonomous agents, *Adapt. Behav.* **7**: 411– 434.

Billard, A., and Hayes, G., 1999. DRAMA, a connectionist architecture for control and learning in autonomous robots. *Adapt. Behav. J.* **7**: 35–64.

Billard, A., and Mataric, M., 2001. Learning human arm movements by imitation: evaluation of a biologically-inspired connectionist architecture. *Robot. Auton. Syst.* **941**: 1–16.

Billard, A., Epars, Y., Schaal, S., and Cheng, G., 2003. Discovering imitation strategies through categorization of multi-dimensional data. *Proceedings Int. Conference on Intelligent Robots and Systems*, Las Vegas, NV, pp. 2398–2403.

Breazeal (Ferrell), C., and Scassellati, B., 2002. Challenges in building robots that imitate people. In K. Dautenhahn and C.L. Nehaniv (eds.) *Imitation in Animals and Artifacts*. Cambridge, MA: MIT Press, pp. 363–390.

Brooks, R.A., Breazeal, C., Irie, R., *et al*., 1998. Alternative essences of intelligence. *Proceedings 15th National Conference on Artificial Intelligence*, Madison, WI, pp. 961–968.

Byrne, R.W., and Russon, A.B., 1998. Learning by imitation: a hierarchical approach. *Behav. Brain Sci.* **21**: 667–721.

Calinon, S. and Billard, A., in press. Learning of gestures by imitation in a humanoid robot. In K. Dautenhahn and C.L. Nehaniv (eds.) *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*. Cambridge, UK: Cambridge University Press.

Carey, S., and Xu, F., 2001. Infant's knowledge of objects: beyond object files and object tracking? *Cognition* **80**: 179–213.

Carpenter, G. A., and Grossberg, S., 1987. ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics* **26**: 4919–4930.

Chappelier, J.C., Gori, M., and Grumbach, A., 2001. Time in connectionist models. In R. Sun and C.L. Giles (eds.) *Sequence Learning: Paradigms, Algorithms, and Applications*. New York: Springer-Verlag, pp. 105–134.

Clark, A., 1997. *Being There: Putting Brain, Body and World Together Again*. Cambridge, MA: MIT Press.

Demiris, J., and Hayes, G.M., 2002. Imitation as a dual-route process featuring predictive and learning components: a biologically plausible computational model. In K. Dautenhahn and C.L. Nehaniv (eds.) *Imitation in Animals and Artifacts*. Cambridge, MA: MIT Press, pp. 321–361.

Dominey, P.F., 2003. Learning grammatical constructions from narrated video events for human–robot interaction. *Proceedings IEEE Humanoid Robotics Conference*, Karlsruhe, Germany

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

Festinger, L.A., 1957. *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Frezza-Buet, H., and Alexandre, F., 2002. From a biological to a computational model for the autonomous behavior of an animal. *Inform. Sci*. **144**: 1–43.

Glenberg, A.M., and Kaschak, M.P., 2002. Grounding language in action. *Psychonom. Bull. Rev*. **9**: 558–565.

Goga, I., and Billard, A., 2004. *A Computational Framework for the Study of Parallel Development of Manipulatory and Linguistic Skills*, Technical Report. Lausanne, Switzerland: Autonomous Systems Laboratory, Swiss Institute of Technology.

Goodson, B.D., and Greenfield, P.M., 1975. The search for structural principles in children's manipulative play. *Child Devel*. **46**: 734–746.

Greenfield, P., 1991. Language, tool and brain: the ontogeny and phylogeny of hierarchically organized sequential behavior. *Behav. Brain Sci*. **14**: 531–550.

Greenfield, P.M., and Schneider, L., 1977. Building a tree structure: the development of hierarchical complexity and interrupted strategies in children's construction activity. *Devel. Psychol*. **13**: 299–313.

Greenfield, P., Nelson, K., and Saltzman, E., 1972. The development of rulebound strategies for manipulating seriated cups: a parallel between action and grammar. *Cogn. Psychol*. **3**: 291–310.

Grossberg, S., 1976. Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, and illusions. *Biol. Cybernet*. **23**: 187–202.

Iacoboni, M., Woods, R.P., Brass, M., *et al*., 1999. Cortical mechanisms of human imitation. *Science* **286**: 2526–2528.

Itti, L., and Koch, C., 2001. Computational modeling of visual attention. *Nature Rev. Neurosci*. **2**: 194–203.

Iverson, J.M., and Thelen, E., 1999. Hand, mouth, and brain: the dynamic emergence of speech and gesture. *J. Consci. Stud*. **6**: 19–40.

Harnad, S., 1990. The symbol grounding problem. *Physica D: Nonlin. Phenom*. **42**: 335–346.

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

Hebb, D.O., 1949. *The Organization of Behavior: A Neuropsychological Theory*. New York: John Wiley.

Hochreiter, S., and Schmidhuber, J., 1997. Long short-term memory. *Neur. Comput*. **9**: 1735–1780.

Hubel, D., 1995. *Eye, Brain, and Vision*, 2nd edn. New York: Scientific American Library.

Kozima, H., and Yano, H., 2001. A robot that learns to communicate with human caregivers. *Proceedings 1st Int. Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, Lund, Sweden, p. 85.

Levy, W.B., and Desmond, N.L., 1985. The rules of elemental synaptic plasticity. In W. Levy and J. Anderson (eds.) *Synaptic Modifications, Neuron Selectivity and Nervous System Organization*. Hillsdale, NJ: Lawrence Erlbaum, pp. 105–121.

Lieberman, P., 2000. *Human Language and Our Reptilian Brain: The Subcortical Bases of Speech, Syntax, and Thought*. Cambridge, MA: Harvard University Press.

Maas, W., and Bishop, C.M. (eds.) 1999. *Pulsed Neural Networks*. Cambridge, MA: MIT Press.

MacWhinney, B., 2000. *The CHILDES Project*, 3rd edn., vol. 1, *Tools for Analyzing Talk: Transcription Format and Programs*. Mahwah, NJ: Lawrence Erlbaum.

Marian (Goga), I., Reilly, R.G., and Mackey, D., 2002. Efficient event-driven simulation of spiking neural networks. *Proceedings 3rd WSEAS International Conference on Neural Networks and Applications*, Interlaken, Switzerland.

Mervis, C.B., 1987. Child-basic objects categories and early lexical development. In U. Neisser (ed.) *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge, UK: Cambridge University Press, pp. 201–233.

Metta, G., Sandini, G., and Konczak, J., 1999. A developmental approach to visually guided reaching in artificial systems. *Neur. Networks* **12**: 1413–1427.

Nagai, Y, Hosoda, K., and Asada, M., 2003. How does an infant acquire the ability of joint attention? A constructive approach. *Proceedings 3rd Int Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pp. 91–98.

Nothdurft, H.C., 2000. Salience from feature contrast: additivity across dimensions. *Vision Res*. **40**: 3181–3200.

Oztop, E., and Arbib, M., 2002. Schema design and implementation of the grasp-related mirror neuron system. *Biol. Cybernet*. **87**: 116–140.

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

Pollack, J.B., 1990. Recursive distributed representations. *Artif. Intell*. **46**: 77–105.

Pulvermüller, F., 1999. Words in the brain's language. *Behav. Brain Sci*. **22**: 253–336.

2002. A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. *Prog. Neurobiol*. **67**: 85–111.

2003. *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge, UK: Cambridge University Press.

Regier, T., 1995. A model of the human capacity for categorizing spatial relations. *Cogn. Ling*. **6**: 63–88.

Reilly, R.G., 2001. Collaborative cell assemblies: building blocks of cortical computation. In *Emergent Neural Computational Architectures Based on Neuroscience: Towards Neuroscience-Inspired Computing*. New York: Springer-Verlag, pp. 161–173.

2002. The relationship between object manipulation and language development in Broca's area: a connectionist simulation of Greenfield's hypothesis. *Behav. Brain Sci*. **25**: 145–153.

Reilly, R.G., and Marian (Goga), I., 2002. Cortical software re-use: a computational principle for cognitive development in robots. *Proceedings 2nd Int. Conference on Development and Learning*.

Rizzolatti, G., and Arbib, M.A., 1998. Language within our grasp. *Trends Neurosci*. **21**: 188–194.

Rose, S.A., Feldman, J.F., and Jankowski, J.J., 2001. Visual short-term memory in the first year of life: capacity and recency effects. *Devel. Psychol*. **37**: 539–549.

Sauser, E., and Billard, A., 2005. Three-dimensional frames of references transformations using recurrent populations of neurons. *Neurocomputing* **64**: 5–24.

Scassellati, B., 1999. Imitation and mechanisms of shared attention: a developmental structure for building social skills. In C. Nehaniv (ed.) *Computation for Metaphors, Analogy, and Agents*. New York: Springer-Verlag, pp. 176–195.

Schaal, S., 1999. Is imitation learning the route to humanoid robots? *Trends Cogn. Sci.* **3**: 223–231.

2001. *The SL Simulation and Real-Time Control Software Package*, Computer Science Technical Report. Los Angeles, CA: University of Southern California.

Schultz, T., and Lepper, M., 1992. A constraint satisfaction model of cognitive dissonance phenomena. *Proceedings 14th Annual Conference of the Cognitive Science Society*, Bloomington, IN, pp. 462–467.

Seidenberg, M.S., and MacDonald, M.C., 1999. A probabilistic constraints approach to language acquisition and processing. *Cogn. Sci.* **23**: 569–588.

Siskind, J.M., 1995. Grounding language in perception. *Artif. Intell. Rev.* **8**: 371–391.

1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* **61**: 1–38.

2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Artif. Intell. Rev.* **15**: 31–90.

Smith, C., 1970. An experimental approach to children's linguistic competence. In J.R. Hayes (ed.) *Cognition and the Development of Language*. New York: John Wiley, pp. 109–135.

Steels, L., 2003. Evolving grounded communication for robots. *Trends Cogn. Sci.* **7**: 308–312.

Sutcliffe, R.F.E., 1992. Representing meaning using microfeatures. In R. Reilly and N.E. Sharkey (eds.) *Connectionist Approaches to Natural Language Processing* Englewood Cliffs, NJ: Lawrence Erlbaum, pp. 49–73.

Tan, A., and Soon, H., 1996. Concept hierarchy memory model: a neural architecture for conceptual knowledge representation, learning, and commonsense reasoning. *Int. J. Neur. Syst.* **7**: 305–319.

Tomasello, M., 1988. The role of joint attentional processes in early language development. *Lang. Sci.* **1**: 69–88.

Wang, D., 2002. Temporal pattern processing. In M. Arbib (ed.) *The Handbook of Brain Theory and Neural Networks*, 2nd edn. Cambridge, MA: MIT Press, pp. 1163–1167.

Weng, J., McClelland, J., Pentland, A., *et al*., 2001. Autonomous mental development by robots and animals. *Science* **291**: 599–600.

Willshaw, D., Buneman, O., and Longuet-Higgins, H., 1969. Non-holographic associative memory. *Nature* **222**: 960–962.

Ziemke, T., 1999. Rethinking grounding. In A. Riegler, M. Peschl and A. von Stein (eds.) *Understanding Representation in the Cognitive Sciences*. New York: Plenum Press, pp. 177–190.

Zlatev, J., and Balkenius, C., 2001. Introduction: Why epigenetic robotics? *Proceedings 1st Workshop on Epigenetic Robotics*, Lund, Sweden, pp. 1–4.