Symbols and embodiment from the perspective of a neural modeler.

Andreas Knoblauch

2007

Preprint:

This is an accepted article published in Symbols, Embodiment and Meaning. The final authenticated version is available online at: https://doi.org/[DOI not available]

Symbols and embodiment from the perspective of a neural modeler *

Andreas Knoblauch

Honda Research Institute Europe GmbH
Carl-Legien-Str.30, D-63073 Offenbach/Main, Germany

Abstract. This paper contributes to the current debate about symbols and embodiment by pointing out the perspective of a neural modeler. I illustrate the default definitions of 'symbol', 'embodiment', 'meaning', and 'grounding' in the context of detailed neural network models, i.e., on a level more detailed than common connectionist approaches. My arguments are based on Hebbian cell assemblies and detailed models of the cortical microcircuitry. These models have been employed to implement a large-scale cortical architecture to enable a robot to perform simple tasks such as understanding and reacting to simple spoken commands. More generally, I finally discuss the relations between embodiment, grounding, anchoring, binding, and the invariant recognition in distributed hierarchical systems.

1 Introduction and definitions

1.1 Symbols

For a neural network modeler, one simple possible way to discern symbols from non-symbols is to look at the inner structure of the representational units. Sub-symbols have an inner structure which can be used to define a similarity metric relevant for the represented entity. In contrast, symbols have no relevant inner structure (i.e., symbols are abstract and arbitrary). For example, in simple object recognition systems, a non-symbol or sub-symbol may be a vector of sensory features, while a symbol may correspond to a single node representing an object category. These definitions are sufficient for a low-level (e.g., neural) description of a cognitive subsystem (e.g., for object recognition), but may not be adequate for the current debate which is about language and the representation of meaning. Here the discussion includes higher-level symbols employed by a cognitive system that is able to think, to reason, and to manipulate these symbols in a flexible way.

According to the workshop's default definition such a symbol is a "theoretical element that is arbitrary, abstract, and amodal" (Glenberg et al., 2007). Before we proceed by discussing and adapting that definition, it may be useful to be aware of the different contexts in which we will use the word "symbol". The situation is illustrated in Fig. 1. We live in a physical world W where systems or subjects S are part of that world and interact with the world. Some of the systems (namely we, the subjects) somehow are able to generate a usually

*to appear: in A. Glenberg, M. De Vega, A.C. Graesser, editors, *Symbols, Embodiment and Meaning*, Oxford University Press, Oxford, UK, 2007; send emails to: andreas.knoblauch@honda-ri.de

unique psychological or phenomenological space P, which we can employ, for example, to generate ideas or theories T about all kinds of issues on all levels W,S,P,T. In particular, we can make theories about S (predominantly done by biology, neuroscience, and AI), P (psychology), or T (metamathematics or logics). Ideas or theories T consist essentially of a set of symbols (as defined above) and additional rules determining how the symbols can be "manipulated".

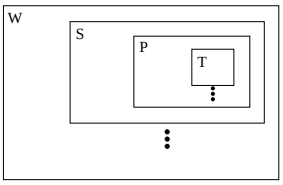

Fig. 1: Different modeling levels: We live in a physical world W. Systems (or subjects) S are part of that world and interact with the world. The subjects somehow are able to generate a (usually unique) psychological or phenomenological space P, which we can use, for example, to generate theories T about all kind of issues on all levels W,S,P,T.

Since symbols are part of theories or ideas this implies two aspects of a symbol, on the one hand in T, on the other hand the implementation of the symbol in S. As a neural modeler one is predominantly interested in the S-level implementation of P in W, i.e., in reducing the observable psychological and behavioral phenomena as good as possible to detailed neural and synaptic processes and finally to the physical laws. (The underlying preliminary naturalistic working hypothesis assumes that this goal is actually possible.) Since models about S are finally also T theories, the neural modeler (and any other kind of S-modeler such as many AI researcher) has obviously to discern between the two kind of symbols *within* his theory: Symbols to model the T-symbols of S, and symbols to model the implementation of the T-symbols of S. Thus, we will refer to these two kind of symbols as T-symbols and S-symbols, respectively. For example, for a cognitive system S capable of understanding language, a T-symbol is the representation of a word, while S-symbols are finer-grained entities used for implementing the word representation. For example, an S-symbol could be a node in a connectionist network (or alternatively, a state or band variable in a Turing machine) while a T-symbol could be implemented by a set of distributed S-symbols (and possibly further dynamic processes).

The current debate is about the question whether cognitive systems (such as we) are or have to be either symbolic or "embodied". Of course, any kind of cognitive system must be symbolic in the trivial sense that we have symbolic language and theories T. Thus, any cognitive system must be T-symbolic. Correspondingly, the critical question is not about the reality of T-symbols, but about the way T-symbols are implemented in S-symbols. Note that, depending on how we define "symbol" and "embodiment", it might be thinkable that T-symbols

could be implemented by non-symbolic processes on the S level (although this seems counterintuitive since we actually aim at developing a symbolic theory T about S).

1.2 Embodiment

Also embodiment comes in several different flavors (cf. Wilson, 2002). The strongest claim would be that embodiment could extend the qualitative or at least quantitative computational capabilities of a system S by exploiting the properties of W, e.g., described by the physical laws. If true, this would essentially negate the Church-Turing thesis that symbolic Turing machines can compute any "naturally" (or physically) computable function. For example, "embodied" analog computers might be better in simulating physical systems, or in computing real numbers with infinite precision (see also Brooks, 1990). Nevertheless, symbolic Turing computers can approximately simulate any physical system by numerically solving the differential equations of physics, although this may take much time and, for chaotic systems, infinite computing precision. Another example are computers exploiting the quantum properties of the physical world. It has been shown that such quantum computers, if physically possible, could compute certain functions much faster than conventional computers or Turing machines (DiVincenzo, 1995).

A less strong idea of embodiment is the dichotomy of embodied versus symbolic cognitive systems addressed by the current debate which attempts to classify cognitive systems according to the interface between the system S itself and the external world W. Obviously, any cognitive system S that deserves that name will have to interact with its environment (percept and act) and is therefore embodied in a trivial sense. Similarly, any cognitive system S must be symbolic in a trivial sense since it must explain our capabilities to use language and think in symbols (e.g., to develop theories T within our psychological space P). Thus, in this trivial sense any model of a cognitive system will be both embodied and symbolic.

Obviously, any cognitive system can be divided into sensors, actors, and internal machinery, such that the interaction with the environment is accomplished only via the sensors and actors. Strictly speaking such an interactive system cannot be adequately modeled by a Turing machine. The original "autistic" Turing machine has been suggested as a model for computation only. That scenario assumes separate phases for (1) providing the input from the environment to the Turing machine's band, (2) doing the computation independently of the environment possibly for a very long time, and (3) returning the output of the computation from the band to the environment. In contrast, we rather have to think of an "interactive" Turing machine that depends on the environment and can influence the environment at any time. Thus, it would be possible to define embodiment by the degree of interaction with the environment.

Our default definition of an "embodied" system goes in a similar direction by demanding that the meaning of a symbol must depend on activity in systems also used for perception, action, and that emotion and reasoning must require

the use of those systems (see Glenberg et al., 2007). This form of *weak embodiment* is stronger than the trivial version of embodiment, but obviously addresses merely the high-level structure of the internal machinery (e.g., in Marr's terms, the algorithmic or computational levels, but not the implementation level). As a consequence, any such "embodied" system can be translated into a purely (S-) "symbolic" system (e.g., a computer program or Turing machine) with the same sensor/actor interface and vice versa. This is true according to the Church-Turing thesis, at least as long as our "embodied" system does not exploit the physical world in a super-Turing manner as described before. We can also conclude that this form of embodiment will probably be neutral to such questions as whether "ideas are the sole province of biological systems" (as discussed in Glenberg et al., 2007). And, of course, the property of embodiment will be a gradual property. Nevertheless, the idea of embodiment might still prove useful, e.g., in building more efficient artificial cognitive systems, or in guiding the analysis of the brain.

1.3 Meaning

Our last definition of embodiment refers to the term "meaning" which may require an explicit definition. A simple definition states that meaning is the "content" of a sign or symbol. Here "content" refers to all the parts of an information processing systems's theories T that have a relation to the symbol. For example, the meaning of the word symbol "car" may include prototypical "icons" of cars, knowledge about the corresponding "consists-of" and "is-a" ontologies, knowledge about actions that can be done with, to, or by a car, and episodic knowledge about particular experiences with cars. This very general meaning of a symbol must usually be strongly constrained by context.

An alternative more behavioristic definition is the following: The meaning of a symbol or sign to a cognitive system is the sum of the potential or actual behavioral changes after receiving the sign. For example, the meaning of the traffic sign "stop" to a car driver is a propensity to apply the brakes to stop the car. Or the meaning of signs indicating bad politics to a voter is a propensity to no longer giving a vote to the responsible politician in future elections.

The second definition has several advantages: First, it does not refer to the internals of the system which may be difficult to observe and interpret, e.g., in animal experiments. Instead, it solely refers to the system's or agent's observable behavior or actions. Further, this direct reference to actions and agents is more relevant for our current debate about meaning and embodiment. Defining meaning in this way by particular actions then obviously implies (by definition) that establishing meaning is closely related to or even "depends on activity in action systems" which is essentially our definition of embodiment (see Glenberg et al., 2007).

Indeed, neurobiological experimenters investigating the brains of animals by correlating behavior and neural activity actually have to define categories (which can then be symbolized) by actions. For example, experimenters can find out which one of two possible interpretations of an ambiguous figure a monkey is

perceiving by training the monkey to move the right hand for interpretation 1 and the left hand for interpretation 2 (which already defines the meaning of the figures for the monkey). Similarly, it has been proposed that we humans as well as other animals can learn to discriminate two classes or categories only if the discrimination is behaviorally relevant. This is most obvious for lower animals which have only a very limited behavioral repertoire (and therefore only a limited way of developing abstract classes such as prey, predator, or mate corresponding to actions like feeding, flighting, or mating).

In summary, we may already conclude here that understanding the meaning of signs indeed requires embodiment defined as regular activation of action and perception related subsystems. However, this conclusion is not sufficient for understanding how the brain works, or to understand how to create intelligent artificial systems. This requires a detailed theory on how perception, action, and learning of abstract categories is accomplished by the brain. In the following we will deliver building blocks for this by having a closer look on neural models of symbols and meaning.

1.4 The easy and the hard problems

With this essay I do not intend to tackle what has been called the "hard problems" in the psychology of consciousness, for example, explaining how physical processes in the brain give rise to subjective experience (Chalmers, 1996; Jackendoff, 1987). Although I believe that, from a third-person perspective, symbols, embodiment, and meaning as defined above are easy problems (i.e., nothing "mystic"), and that we will probably soon be able to realize artificial systems that can be said to be embodied and represent meaning in a *similar* way to humans, I admit that assuming full identity between the first person subjective processes of humans (including feelings) and those of digital computers (Dennett, 1996; Metzinger, 2004) may have problematic consequences.

For example, if we would ascribe subjective feelings to a robot then we would also have to ascribe feelings to the robot rid of its sensory/motor interface (similarly as we ascribe feelings to quadriplegia and locked-in syndrome patients). Then we finally would have to ascribe feelings to a digital computer, i.e., finite-state-machine (FSM). More exactly, we would ascribe feelings to particular states of the FSM. Adapting Searle's Chinese room argument (Searle, 1980) to feelings (instead of meaning), this seems strange because the FSM's states are arbitrary, i.e., it is unclear why one particular physical state should be associated with pain (and not with joy or something else).

One could answer that meaning and feelings are not associated with a single state but with particular but still *recurring* state sequences (for example, realizing a kind of monitoring structure). However, this will probably not help us since we can always construct an "equivalent" FSM where such a sequence corresponds to a single state again (although this will require a large FSM with many states). We may accept this for the case of meaning defined in terms of "potential behavior" (see above) because the question "Is that machine understanding this?" can be resolved, in principle, by looking at the FSM's past or future

states. However, we are usually more reluctant in the case of feelings because the question "Is that machine feeling pain?" must be answered in the present (and proposing that past or future states would make a difference contradicts the state concept of classical physics).

2 A neural modeler's perspective

When words referring to actions or visual scenes are presented to humans, distributed neural networks including areas of the motor and visual systems of the cortex become active (e.g., Pulvermüller, 1999, 2003). The brain correlates of words and their referent actions and objects appear to be strongly coupled neuron ensembles in defined cortical areas. The theory of *cell assemblies* (Hebb, 1949; Braitenberg, 1978; Palm, 1982, 1990) provides one of the most promising frameworks for modeling and understanding the brain in terms of distributed neuronal activity. It is suggested that entities of the outside world (and also internal states) are coded in groups of neurons rather than in single ("grandmother") cells, and that a neuronal cell assembly is generated by Hebbian coincidence or correlation learning where the synaptic connections are strengthened between co-activated neurons. Models of neural (auto-) associative memory have been developed as abstract models for cell assemblies.

2.1 Local cell assemblies, associative memory, and neural S-symbols

The notion of cell assemblies as strongly coupled neurons leads to the concept of neural *(auto-) associative memory* (Willshaw et al., 1969; Palm, 1980; Hopfield, 1982). A particular simple model of neural associative memory has been proposed by Steinbuch and Willshaw (see Willshaw et al., 1969; Steinbuch, 1961; Palm, 1980; Knoblauch, 2005) consisting of McCulloch-Pitts type threshold units and recurrent binary synapses (Fig. 2). Here the activity pattern of the cell population can be described by a binary vector, and we identify stored activity patterns with the cell assemblies. After learning a number of cell assemblies the network can be described by a connection matrix A corresponding to a graph, where the nodes correspond to the neurons, and cell assemblies correspond to k-cliques of neurons (a k-clique is a subset of size k consisting of completely connected neurons). *Hetero-association* works similar to auto-association except that the "memory matrix" describes the synaptic connections between *two* different neuron populations. Hetero-associative connections can map assemblies of the first population (or sets or parts of them) to cell assemblies of the second population (or sets or parts of them).

The virtue of the binary model is that it is easy to understand, analyze, and implement, but the main results apply also to more realistic gradual and spiking models (Hopfield, 1984; Knoblauch and Palm, 2001). Neural associative memories have a couple of nice features. They achieve pattern completion, i.e., a cell assembly can be activated not only by the very same inputs that have been used for learning, but also by modified patterns that are "sufficiently" similar to the original address pattern. For example, assembly u2 in Fig. 2 will already

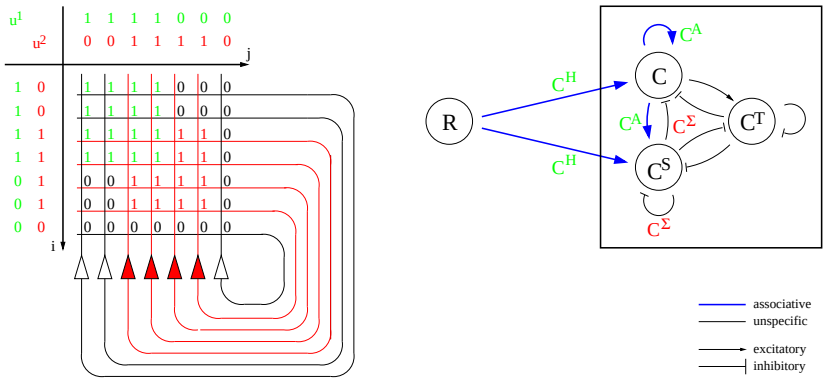

Fig. 2: Left: Neural (auto-)associative memory where two cell assemblies of size $k = 4$ have been stored (corresponding to the activation patterns u1 and u2) in the "memory matrix" of synaptic connections. Filled units indicate the active neurons of pattern u2. Right : A more realistic implementation of an associative memory modeling a small patch (about 1mm^3) of cortical tissue (Knoblauch and Palm, 2001). The model comprises several populations of excitatory and inhibitory spiking neurons of the integrate-and-fire type. Here C is the main excitatory population of pyramidal cells receiving input from another cortical patch R. C^S and C^T are inhibitory interneuron populations controling local excitation. Each neuron is modeled as a leaky integrator with excitatory and inhibitory conductances, where a spike is emitted as soon as the dendritic potential exceeds a threshold. The "memory matrices" are employed in several afferent and recurrent synaptic connections (cH,cA) from and to populations R,C,C^S. The remaining inhibitory feedback connections are unspecific (i.e., independent of the learned activity patterns).

be activated by addressing an arbitrary subset of size ≥ 3. It can be shown that the number of storable patterns scales almost with the number of synapses if the patterns are sparse and have random character (i.e., a population of n neurons can store almost n^2 sparse cell assemblies with $k \ll n$). Access time is essentially independent of the number of stored patterns. The overlaps of different cell assemblies can be used to express the similarities of the represented entities. Cell assemblies thereby provide a very natural associative way of grounding new representations in the sensory inputs by means of bidirectional associative connections (cf. Barsalou, 2003).

Associative memories have been used to model small volumes of cortical tissue (e.g., 1mm^3) corresponding to a macrocolumn or the range where dense local recurrent connections between any cell pairs are possible (Braitenberg and Schüz, 1991). A step towards biological realism is to replace the single McCulloch-Pitts population by more realistic spiking neuron models and to incorporate known

properties of cortical circuits (Fig. 2). These models can extend the computational abilities of the standard model, for example by making use of spike timings according to a latency code in that early spikes (relative to an external event or an underlying oscillation) are much more relevant than late spikes for activating an assembly (Knoblauch and Palm, 2001; Knoblauch, 2005).

Local cell assemblies can be seen as elementary neural (S-) symbols which can be "allocated" or learned to represent the inputs for further processing in down-stream target populations. The symbolic character is most apparent if the assembly is $k = 1$ corresponding to a localist code, or if the neurons that constitute a cell assembly are chosen at random, e.g., by noise. In the latter case the correlations (or overlaps) between two cells are minimal which is required to store a maximal number of different activity patterns. Due to their singular or random character the cell assemblies could be said to be abstract and arbitrary, whereas the property of amodality depends on the location of the neuron population, e.g., a local population receiving visual inputs will develop visual perceptual symbols (cf. Barsalou, 1999).

2.2 Global cell assemblies, language, and T-symbols

We have designed a large-scale brain model consisting of many interconnected cortical areas employing spiking associative memories. The model was implemented and tested on a robotic platform enabling the robot to understand and react to simple commands such as "Bot show plum!" (Knoblauch et al., 2004). The language part of the model is illustrated by Fig. 3, the action part by Fig. 4.

Each box in the figures corresponds to a spiking neural associative memory storing local cell assemblies as described above. For illustration purposes, each area has been labeled according to the current activity pattern. (In general a superposition of several stored local assemblies can be activated, e.g. to represent uncertainty or to represent new entities to be learned; here the labels correspond to the cell assembly most similar to the current activation pattern). The resulting *global assembly*, for example representing the T-symbol "plum", stretches over many cortical areas (involving visual, auditory, action, and goal-related areas) and changes dynamically during the process of "understanding" and reacting to the command. Thus, the global cell assembly as a whole works as a sign in Peirce's sense, i.e., as a mediator between the idea of a "plum" and the real plum in the external world. The global assembly consists of parts some of which can be attributed as "abstract" and "amodal", e.g., the lexical representation of "plum" in A3. But these symbolic parts are naturally grounded in the synaptically connected perceptual and action-related parts of the global assembly.

2.3 Cortical macrocolumns, prediction, and embodiment

Cognitive processes must be able to distinguish between different representational modi. For example, representational states may refer to present, future (or prediction), reality, wish, signal (detailed and concrete), or symbol (abstract, amodal), perception or action. Many cognitive architecture take a modular ap-

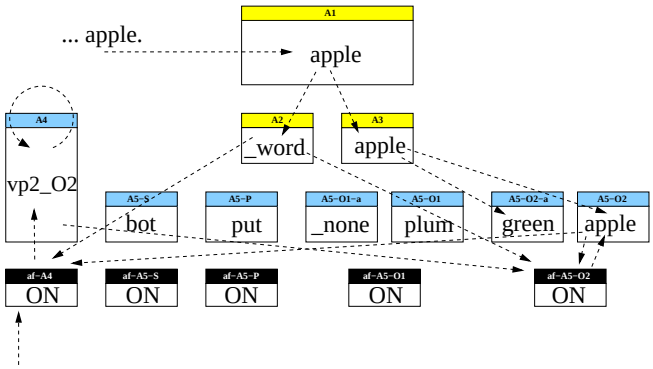

Fig. 3: The language part of an associative cortex model (see Knoblauch et al., 2004) at the end of processing the sentence "Bot put plum to green apple". Each box corresponds to a cortical (or subcortical) area modeled as a neural associative memory. The meaning of the sentence is represented by distributed cell assemblies comprising "slot areas" for different grammatical roles to implement elementary productivity and systematicity. Auditory input enters via areas A1 and A3 the central areas and is distributed across the grammatical slots according to a logic controlled by a grammatical sequence memory (A4) (where basic sentence types are stored) and subcortical "activation fields" (small boxes). Arrows indicate recently activated synaptic connections.

proach where these different representational modi are segregated into different cognitive subsystems or modules. (In general, an architecture can be said to be modular if it can be divided into subsystems such that there is much more communication between processes inside a subsystem than between processes of different subsystems.) For example, we could segregate a cognitive system into different modules for perceptions, actions, goals, memory, rule-based prediction systems, etc.

In the last section we have indeed argued how global cell assemblies can implement and ground T-symbols (e.g., words of a language) by distributed activation stretching over many sensory, motoric, and associative cortical areas (Pulvermüller, 1999). Thus, although the brain appears to have a modular character in that sense, hints to possible complementary strategies of grounding may be found when looking at the microstructure of a single cortical macrocolumn (Douglas and Martin, 2004).

Although, it is well known for a long time that neocortical anatomy exhibits a 6-layered structure, modelers have often neglected this fact when modeling a cortical patch by a single "monolithical" neuron population (e.g., Palm, 1982; Ritz et al., 1994; Knoblauch and Palm, 2001). This may be attributable to the wish to focus on a single layer or the lack of adequate computational resources to simulate more detailed models, but also to doubting or underestimating the

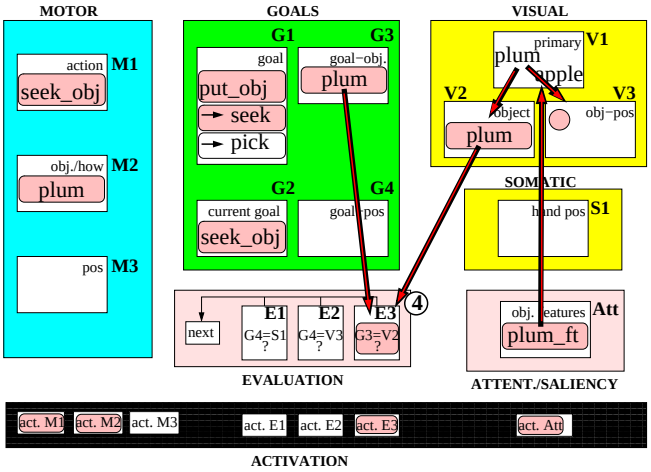

Fig. 4: Action part of an associative cortex model (see Knoblauch et al., 2004) during performing the command "Bot put plum to green apple!". The goal areas (G1-G4) received their inputs from the grammatical role areas (A5-S, A5-O1, etc. as illustrated in Fig. 3) and divide the goals into a sequence of subgoals (i.e., seek plum, pick plum, move to the apple, drop plum). (High-level) motor areas receive inputs from the goal areas in order to perform the current subgoal. The completion of subgoals and switching to the next subgoal is controlled by "evaluation fields" checking, for example, the consistency of visual perceptual activity patterns (e.g., in V4) with goal representations (e.g., in G3). At the shown system state the robot is about to finish the subgoal of seeking the plum.

functional significance of discrete within- or between-layer synaptic connections which appear to have a rather "fuzzy" character (Abeles, 1991; Braitenberg and Schüz, 1991).

In accordance with ideas developed earlier by Körner et al. (1999) (see also Hawkins, 2004; Rao and Ballard, 1999; Guillery, 2003), we assume as a working hypothesis that the *basic function of a cortical column* is to represent and *actively* predict its sensory inputs. To achieve this in a self-organizing, autonomous way, it is necessary to have access to (at least some of) the different representational modi described above. We propose that different representational modi of the same entity are located in different layers within the same macrocolumn rather than monolithically in different columns or areas.

Fig. 5 illustrates this functional model and our current implementation employing spiking associative networks similar as discussed in section 2.1. At each time the model must represent a state $v = (w, a)$ and use sensory input s to update the state v according to a function f. We found it meaningful to divide the state variable v into two rather independent entities, one variable w describing "external" entities from the outside world, and another variable a describing a

local "internal actor". In addition to updating a state, the system should also be able to predict a future state w' without accessing sensory input. Note that the proposed circuitry provides the basic ingredients for simulating (or predicting over) larger time intervals.

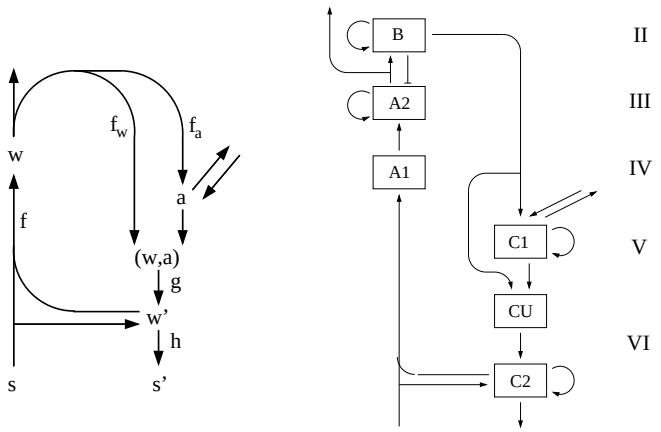

Fig. 5: **Left:** Basic functional circuit of a cortical column. Sensory input s is used to update the current world state w. This is used to choose an appropriate action a. World state and action can be used to predict the next world state w' and next sensory input s'. **Right:** Implementation of a cortical column by spiking associative memories (see Knoblauch et al., 2007, 2005; Körner et al., 1999). Sensory input activates signal-like representations (based on basis vectors) in the middle cortical layers (A1,A2), a symbol-like prototypical cell assembly in the upper layers (B), and finally action and prediction related representations in the lower layers (C1,C2).

By comparison with known anatomical facts we can match our functional model (Fig. 5) with the layered organization of neocortex (Körner et al., 1999; Guillery, 2003; Douglas and Martin, 2004; Braitenberg and Schüz, 1991; Felleman and Van Essen, 1991). For example, it is well established that feed-forward inputs to a cortical column target mainly layer IV neurons and that the feed-forward output to the next cortical stage leaves a cortical column via layer II/III neurons. In contrast to the feed-forward stream, feed-back inputs avoid layer IV and target mainly the upper layers (but also the lower layers). Another remarkable feature of the cortical microcircuitry is that layer V pyramidal neurons, at any cortical site, project to subcortical regions closely related to action and behavior (Guillery, 2003).

Based on these facts we believe that the forward recognition function f is located in the middle and upper layers, while the remaining functionality, related to behavior and predictions, is located in the lower layers V/VI. Furthermore, we believe that the recognition system of the middle and upper layers is split

up into two subsystems, one for fast bottom-up recognition (A system, layer IV and upper III) and another for refined recognition employing feedback (B system, layers II/III).

As an example we have implemented a model of several cortical and subcortical areas for learning saccadic object representations (including several visual cortical areas and the superior colliculus; see Knoblauch et al., 2007, 2005). In that particular case the representational world states are object views (e.g., the retinal image when fixating on a particular key feature of a visual object) and the actions correspond to saccades.

What does this has to do with embodiment and grounding? Our model suggests that actions are grounded in perceptions already at the level of a single cortical macrocolumn (cf. Hawkins, 2004; Körner et al., 1999; Guillery, 2003; Young, 1993). Thus, recognition of incoming signals will automatically induce an action related process in the same macrocolumn and related subcortical structures. In our model, the perceptual representation w is on the one side really symbolic in cortical layers II/III, it is grounded in sensory input and a signal-like (basis vector based) representation in layer IV, and it is also the result of a prediction $(w^{t-1}, a) \rightarrow w$ (layer VI) such that it is also grounded in action (layer V). Vice versa, for the same reason the produced action will be grounded in perception. Furthermore, the proposed circuitry provides the basic ingredients for simulating (or predicting) the represented state (w) over larger time intervals (at a microlevel), which has been suggested to be required for understanding meaning on the (P,T-) macro-levels (cf. Barsalou, 1999).

2.4 Hierarchies, invariances, binding, and grounding

The areas of the brain and particularly the cerebral cortex are organized in a distributed and hierarchically ordered manner. For example, the visual system of primates consists of about 50 areas at about 15 processing stages, where each area is dedicated to a specific task (Felleman and Van Essen, 1991). On the top of the hierarchy there are neurons in certain areas representing quite abstract entities coming close to (T-)symbols. For example, in the medio-temporal lobe cells have been found that respond in a specific and invariant way to a particular person, regardless of the stimulus being the person's face, a cartoon, or the written name (Quiroga et al., 2005).

There are many neural models (and AI systems) claiming to do object recognition in a similar way (e.g., Riesenhuber and Poggio, 1999; Wersing and Körner, 2003), although performance can not yet be compared to that of the real brain. The basic principle of a hierarchical object recognition system is illustrated in Fig. 6, left panel. Processing along the hierarchy usually changes gradually in two ways: On the one hand, the represented feature quality becomes more and more specific with increasing hierarchical level (e.g., from basic features to particular objects). On the other hand, the representations become more and more invariant against transformations in the sensory space (e.g., translation, scale, rotation, color, lighting conditions). The first aspect is essentially an AND operation (e.g., an corner consists of a vertical edge AND a horizontal edge), the

second aspect an OR operation (e.g., a feature configuration may occur at one OR another location as indicated by the pooling in Fig. 6, left panel).

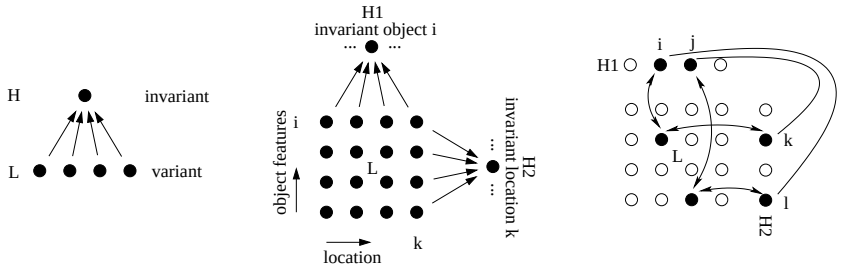

Fig. 6: Illustration of sign grounding, anchoring or binding. Left : The development of invariant abstract (e.g., symbolic) representations implies a non-invertible mapping between a lower processing stage L and higher stage H (e.g., mapping feature configurations at different spatial locations to the same object label). This complicates adequate grounding, in particular if several objects are processed at the same time. Middle : To become invertible one needs further (possibly independent) mappings, e.g., an explicit representation of the object location. Right : In the presence of several objects, correct bindings between the higher-level symbols (e.g., object label and location) and grounding to the lower level features may require additional mechanisms (e.g., temporary associations).

Thus, such an object recognition system must somehow map or bind, *on each level of the hierarchy*, a more abstract representation to particular configurations of lower level features which may vary significantly. Realizing this in a bottom-up fashion is straight-forward, for example by means of a multi-layer perceptron (MLP). However, because of the invariance property (by the OR operation), this mapping between the two levels is necessarily non-invertible which complicates binding between the two processing levels, in particular for more realistic scenarios requiring feed-back processing (e.g., dynamical ambiguous scenes with multiple objects, occlusions, clutter, etc., as well as particular queries referring to low level sub-symbolic properties of the objects).

The mapping can be made again (almost) invertible by further independent processing paths explicitely representing the varying factors (e.g., object type and object location in Fig. 6, middle panel). However, this may lead to a further binding problem if several objects are processed at the same time: For example in Fig. 6 (right panel) it is problematic to ground the two objects i and j of layer H1 in the feature layer L because it may be unclear which object occurs at which of the two locations k and l represented in layer H2. Implicit binding over the lower layer L is possible but non-trivial (note that the segregated nodes of layer L may correspond in fact to distributed overlapping feature *configurations*). Neural experimenters and theoreticians have spent a lot of work to investigate this form of the binding problem, and several explicit solutions have been sug-

gested, for example by spike synchronization and oscillations, rapid reversible synaptic plasticity, dynamic routing, attention, coarse conjunctive coding (e.g., see von der Malsburg, 1999; Shastri and Ajjanagadde, 1993; Treisman, 1998; Mel and Fiser, 2000; Knoblauch and Palm, 2002).

In summary, it appears that embodiment is very closely linked (if not identical) to the problems of symbol grounding (Harnad, 1990), symbol anchoring (Coradeschi and Saffioti, 2003), binding, and invariant recognition in distributed hierarchical systems. The underlying problem is always the coordination between the bottom-up and top-down streams of information in a hierarchical system, which still lacks a satisfactory solution in current artifical systems and theoretical models.

3 Conclusions

In general I am inclined to accept the major claims of the embodiment proponents, for example that understanding the meaning of a sentence may require the ability to simulate or predict the situation described by the sentence, or that meaning is closely related to action and perception (which may cause interference effects as observed in experiments, see Glenberg and Kaschak, 2002; Rizzolatti et al., 2001; Guillery, 2003). But in contrast to Searle's Chinese room argument I see no good reason why implementations on symbolic systems such as Turing machines or currently available digital computers should not do that job. Nevertheless, special hardware such as massively parallel neural networks, although in terms of price, speed, and flexibility still inferior to general purpose processors, could be of great advantage to this end (e.g., Hammerstrom et al., 2006).

Also I believe that many so-called symbolists would finally also agree with these positions, and that probably much of the remaining disagreement results rather from imprecise definitions about what exactly is a symbol and what is embodiment. As discussed in section 1.1, it is important here to distinguish between what I have called T-symbols and S-symbols. The T-symbols are used by the cognitive system for thinking and communicating, while the S-symbols may be used to implement the cognitive system (including the T-symbols) on a symbolic substrate such as a digital computer, Turing machine, or in a neural network architecture (see sections 2.1 and 2.2). For example, a T-symbol could be a word for an object, while an S-symbol could be a state of a Turing machine. Note that this distinction leads to the apparent contradiction that a system can be both (T-)embodied and (S-)symbolic. This is simply because embodiment in the sense of "employing systems used both for action and perception" is actually a high-level property and therefore independent of the implementation substrate. Thus, any implementation of an embodied system on a digital computing device must be called (S-) symbolic.

We have also seen that the dichotomy between symbolic and embodied systems can be problematic. As argued in section 1.2, any cognitive system must be necessarily both symbolic and embodied. Moreover, defining embodiment by

activity in systems for action and perception implies that being T-embodied is a rather gradual property: A high-level T-symbol may be embodied more or less deeply in a hierarchy of perception and action-related subsystems. At least in the brain there certainly exists such a hierarchy consisting of many different cortical areas (Felleman and Van Essen, 1991). Understanding a particular sentence of T-symbols will require some but probably not all of these areas. Thus, one could define the degree of embodiment by the number and hierarchical levels of such areas necessary for understanding.

For example, for understanding "the grass is yellow" a shallow embodiment is likely to be sufficient. Here T-symbolic processing may be enough for understanding which is merely the association of two well-known T-symbols. Of course, cell assembly theory would anyway predict that in the brain the T-symbol "yellow" is so familiar that there will be strong synaptic links from the (T-)symbolic "yellow" subassembly to low sensory neurons representing the color yellow. Thus, hearing the word "yellow" would inevitably activate those low level neurons to a certain degree. But I think that here this priming activity would be rather a side effect of cell assemblies and actually not necessary for understanding. In a potential experiment one could temporarily deactivate or disturb a number of low-level visual cortices (e.g., by TMS) while hearing "the grass is yellow" and one would predict that understanding is still possible.

In contrast, understanding or verifying the sentence "the second house on the right has seven floors" requires relating several T-symbols to a currently perceived visual scene which may be represented in a (T-)subsymbolic format in early visual areas. This process includes identifying the "second house on the right" among many other buildings and relating the number "seven" to the floors of the house. Here one would predict that understanding is much more vulnerable to the deactivation of low-level cortices.

Finally, on-the-fly understanding of sentences such as "the woman crutched the goalie the ball" containing innovative denominal verbs will probably require even deeper embodiment including mental simulation of the described situation at many processing levels (see Glenberg and Kaschak, 2002).

If we accept that embodiment is a gradual property within a processing hierarchy this idea becomes very closely related (if not identical) to long-discussed concepts and problems such as symbol grounding, symbol anchoring, and feature binding (see section 2.4). We could give the following (maybe too) simple definition: A sign or symbol of a higher processing stage H is embodied in (or, synonymously, grounded in, anchored in, or bound with) the signs or symbols of a lower processing stage L iff it is possible to establish a one-to-one mapping between the H signs and L signs.

This definition of embodiment would have several advantages: First, it is a unified definition for seemingly different but closely related concepts. The definition remains independent of particular modalities (such as particular perceptions, actions, or emotions). Instead it is sufficient to distinguish between higher and lower processing stages. But we can still say that a symbol X is grounded in particular action representations Y. And also we can still distinguish shallow

embodied systems from deeply embodied systems where the one-to-one mapping is established across several processing stages.

However, the term "mapping" may still be underspecified. Unfortunately, I cannot give a more specific answer, because in my opinion this is one of the unsolved core problems of current brain models and artificial cognitive systems as discussed in section 2.4: I.e., the problem of adequately integrating bottom-up information with top-down expectations is by no means satisfactorily understood in complex systems consisting of distributed processing hierarchies and capable of developing increasingly abstract invariant (and finally symbolic) representations. I believe that the circuitry of the cerebral cortex (see section 2.3) can tell us a lot of how this integration happens in the brain, and how learning of abstract categories is linked to action and behavior.

References

Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex.* Cambridge University Press, Cambridge UK.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–609.

Barsalou, L. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Science*, 7(2):84–91.

Braitenberg, V. (1978). Cell assemblies in the cerebral cortex. In Heim, R. and Palm, G., editors, *Lecture notes in biomathematics (21). Theoretical approaches to complex systems.*, pages 171–188. Springer-Verlag, Berlin Heidelberg New York.

Braitenberg, V. and Schüz, A. (1991). *Anatomy of the cortex. Statistics and geometry.* Springer-Verlag, Berlin.

Brooks, R. (1990). Elephants don't play chess. *Robotics and Autonomous Systems*, 6:3–15.

Chalmers, D. (1996). *The Conscious Mind.* Oxford University Press, Oxford.

Coradeschi, S. and Saffioti, A. (2003). An introduction to the anchoring problem. *Robotics and Autonomous Systems*, 43(2-3):85–96.

Dennett, D. (1996). *Kinds of Minds: Towards an Understanding of Consciousness.* Weidenfeld & Nicolson, London.

DiVincenzo, D. (1995). Quantum computation. *Science*, 270:255–261.

Douglas, R. and Martin, K. (2004). Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.*, 27:419–451.

Felleman, D. and Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47.

Glenberg, A., De Vega, M., and Graesser, A. (2007). Framing the debate. In Glenberg, A., De Vega, M., and Graesser, A., editors, *Symbols, Embodiment and Meaning*. Oxford University Press, Oxford, UK.

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

Guillery, R. (2003). Branching thalamic afferents link action and perception. *Journal of Neurophysiology*, 90:539–548.

Hammerstrom, D., Gao, C., Zhu, S., and Butts, M. (2006). FPGA implementation of very large associative memories. In Omondi, A. and Rajapakse, J., editors, *FPGA implementations of neural networks*, pages 167–195. Springer US.

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

Hawkins, J. (2004). *On Intelligence.* Times Books Henry Holt, New York.

Hebb, D. (1949). *The organization of behavior. A neuropsychological theory.* Wiley, New York.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, 79:2554–2558.

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science, USA*, 81(10):3088–3092.

Jackendoff, R. (1987). *Consciousness and the Computational Mind.* MIT Press/Bradford Books, Cambridge.

Knoblauch, A. (2005). Neural associative memory for brain modeling and information retrieval. *Information Processing Letters*, 95:537–544.

Knoblauch, A., Fay, R., Kaufmann, U., Markert, H., and Palm, G. (2004). Associating words to visually recognized objects. In Coradeschi, S. and Saffiotti, A., editors, *Anchoring symbols to sensor data. Papers from the AAAI Workshop. Technical Report WS-04-03*, pages 10–16. AAAI Press, Menlo Park, California.

Knoblauch, A., Kupper, R., Gewaltig, M.-O., Körner, U., and Körner, E. (2005). Design and simulation of a cortical control architecture for object recognition and representational learning. In Tsujino, H., Fujimura, K., and Sendhoff, B., editors, *Proceedings of the 3rd HRI International Workshop on Advances in Computational Intelligence*. Honda Research Institute, Wako, Japan.

Knoblauch, A., Kupper, R., Gewaltig, M.-O., Körner, U., and Körner, E. (2007). A cell assembly based model for the cortical microcircuitry. *Neurocomputing*, 70(10-12):1838–1842.

Knoblauch, A. and Palm, G. (2001). Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks*, 14:763–780.

Knoblauch, A. and Palm, G. (2002). Scene segmentation by spike synchronization in reciprocally connected visual areas. II. Global assemblies and synchronization on larger space and time scales. *Biological Cybernetics*, 87(3):168–184.

Körner, E., Gewaltig, M.-O., Körner, U., Richter, A., and Rodemann, T. (1999). A model of computation in neocortical architecture. *Neural Networks*, 12:989–1005.

Mel, B. and Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation*, 12:247–278.

Metzinger, T. (2004). *Being No One. The Self-Model Theory of Subjectivity.* The MIT Press/Bradford Book, Cambridge.

Palm, G. (1980). On associative memories. *Biological Cybernetics*, 36:19–31.

Palm, G. (1982). *Neural Assemblies. An Alternative Approach to Artificial Intelligence.* Springer, Berlin.

Palm, G. (1990). Cell assemblies as a guideline for brain research. *Concepts in Neuroscience*, 1:133–148.

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, 22:253–336.

Pulvermüller, F. (2003). *The neuroscience of language: on brain circuits of words and serial order.* Cambridge University Press, Cambridge, UK.

Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107.

Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025.

Ritz, R., Gerstner, W., Fuentes, U., and van Hemmen, J. (1994). A biologically motivated and analytically soluble model of collective oscillations in the cortex. II. Applications to binding and pattern segmentation. *Biol. Cybern.*, 71:349–358.

Rizzolatti, G., Fadiga, L., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2:661–670.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–57.

Shastri, L. and Ajjanagadde, V. (1993). From simple associations to systematic reasoning: a connectionistic representation of rules, variables and dynamic bindings. *Behavioral and Brain Sciences*, 16(3):417–494.

Steinbuch, K. (1961). Die Lernmatrix. *Kybernetik*, 1:36–45.

Treisman, A. (1998). Feature binding, attention and object perception. *Phil. Trans. R. Soc. London B*, 353:1295–1306.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, 24:95–104.

Wersing, H. and Körner, E. (2003). Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, 15:1559–1588.

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

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4):625–636.

Young, M. (1993). The organization of neural systems in the primate cerebral cortex. *Proceedings of the Royal Society: Biological Sciences*, 252:13–18.