Biological Cybernetics

A multimodal connectionist architecture for unsupervised grounding of spatial language

--Manuscript Draft--

Manuscript Number:	
Full Title:	A multimodal connectionist architecture for unsupervised grounding of spatial language
Article Type:	Original Paper
Keywords:	symbol grounding; unsupervised learning; multimodal representation; visual pathways; SOM
Corresponding Author:	Michal Vavrecka CZECH REPUBLIC
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Michal Vavrecka
First Author Secondary Information:	
Order of Authors:	Michal Vavrecka
	Igor Farkas
Order of Authors Secondary Information:	
Abstract:	We developed a biologically inspired unsupervised connectionist architecture for grounding the spatial terms. This two-layer architecture integrates information from visual and auditory inputs. In the first layer, it employs separate visual 'what' and 'where' subsystems to represent spatial relations of two objects in 2D space. The images are presented to an artificial retina and the phonologically encoded five-word sentences describing the image serve as auditory inputs. The visual scene is represented by several self-organizing maps (SOMs) and the auditory description is processed by a Recursive SOM that learns to topographically represent sequences. Primary representations from the first layer are unambiguously integrated in a multimodal module (implemented by SOM or "neural gas" algorithms) in the second layer. The simulations reveal that separate processing and representation of spatial location and object shape significantly significantly improves the performance of the model. The system is able to bind proper lexical and visual features without any prior knowledge. The results confirm theoretical assumptions about the different nature of visual and auditory coding that become efficiently integrated at the multimodal layer.
Suggested Reviewers:	Angelo Cangelosi A.Cangelosi@plymouth.ac.uk expert in symbol grounding and cognitive modeling
	Alberto Greco greco@unige.it expert in symbol grounding and compositionality
	Vadim Tikhanoff Vadim.Tikhanoff@iit.it He works in the area of cognitive robotics and knowledge representation

Noname manuscript No.
(will be inserted by the editor)

A multimodal connectionist architecture for unsupervised grounding of spatial language

Michal Vavrečka · Igor Farkaš

Received: date / Accepted: date

Abstract We developed a biologically inspired unsupervised connectionist architecture for grounding the spatial terms. This two-layer architecture integrates information from visual and auditory inputs. In the first layer, it employs separate visual what and where subsystems to represent spatial relations of two objects in 2D space. The images are presented to an artificial retina and the phonologically encoded five-word sentences describing the image serve as auditory inputs. The visual scene is represented by several self-organizing maps (SOMs) and the auditory description is processed by a Recursive SOM that learns to topographically represent sequences. Primary representations from the first layer are unambiguously integrated in a multimodal module (implemented by SOM or "neural gas" algorithms) in the second layer. The simulations reveal that separate processing and representation of spatial location and object shape significantly improves the performance of the model. The system is able to bind proper lexical and visual features without any prior knowledge. The results confirm theoretical assumptions about the different nature of visual and auditory coding that become efficiently integrated at the multimodal layer.

Keywords unsupervised learning · symbol grounding · spatial phrases

M. Vavrečka
Faculty of Electrical Engineering, Czech Technical University, Karlovo náměstí 13, Prague
Tel.: +420224357609
E-mail: vavrecka@fel.cvut.cz

I. Farkaš
Department of Applied Informatics, Comenius University, Mlynská dolina, 84248 Bratislava

1 Introduction

The question of how to acquire, represent and use knowledge is fundamental in the artificial intelligence and cognitive science research. Within the modern perspective, fueled by growing empirical evidence, we are looking for a system that interacts with the environment and is able to understand it, by forming its internal representations that result from this interaction. The representations in the system should preserve constant attributes and regularities of the environment, represent them as concepts, and connect these to the symbolic level. This approach to the representation of meaning differs from the classical symbolic approach based on formal principles (Newell & Simon, 1972; Pylyshyn, 1984). The formal approach has been generally criticized for its inherent difficulty to describe the whole world by logic relations (Zieliger, 2010). Formal semantics is sufficient only for the formal languages, but natural language and events captured from the environment are too complex and fuzzy for such semantics.

Hence we need a different strategy to create a system of representations and mechanisms for manipulating them. Barsalou (1999) proposed a theory based on perceptual symbol systems (PSS) that provide schematic neural representations spread across multiple sensory modalities. The advanced version of Barsalou's approach is expressed in the LASS theory describing the interaction between the linguistic system and the conceptual system (Barsalou, 2008). Similar ideas stem from the cognitive semantics (Lakoff, 1987), based on a perceptually created conceptual level and a grounded symbolic level. These approaches are closely related to the symbol grounding problem (Harnad, 1990). The basic task for symbol grounding is to find the function and an internal mechanism to create representations which are

intrinsic to the system, and hence do not need an external observer for their interpretation (Ziemke, 1999). Harnad (1990) proposed a hybrid architecture based on discrimination and identification. Discrimination is a subsymbolic (nonarbitrary) representation of perceptual inputs, and identification assigns (nonarbitrary) concepts to (arbitrary) symbols. Harnad used neural networks for the subsymbolic representations and the classical architecture for symbol operations. In the overview of grounding architectures, Taddeo and Floridi (2005) introduced zero semantical commitment condition as a criterion for valid solution to the symbol grounding problem, completely avoiding the designer's approach. This criterion, however, appears unsatisfiable in artificial systems (Vavrečka, 2006).

In the past decades, there was a number of different approaches and models of the symbol grounding (e.g. Dorffner, Hentze & Thurner, 1996; Cangelosi & Riga, 2006; Fontanari, Tikhanoff, Cangelosi, Ilin & Perlovsky, 2009; Greco, 2010). Sugita and Tani (2005) implemented a connectionist model, consisting of the interconnected modules that were trained to make the mobile robot execute motor commands and comment on them. The system was also able to understand verbal instructions, by demonstrating their execution, which can be interpreted as grounded linguistic knowledge. In all these models, the neural networks provide a natural computational framework for grounding knowledge acquisition. Computational models of grounding, presented so far, mainly focused on grounding nouns in sensorimotor object representations and verbs in actions directly performed by the agent (Marocco, Cangelosi, Fischer & Belpaeme, 2010). Roy and Pentland (2002) developed a system able to segment words from utterances and to associate the proper words with objects. Consequently, Roy et al. (2003) extended this architecture for perceptual, procedural and affordance representations to ground the meaning of words in conversational robots. These works do not care much about biologically inspired features of the models, though.

Our approach is based on biology-inspired modeling (Vavrečka, Farkaš & Lhotská, 2011). The architecture implements the multimodal representations in the framework of the PSS. There are symbolic inputs (sentences) processed by a separate auditory subsystem, perceptual inputs processed by a visual system and the multimodal layer that incorporates the process of identification of symbols with concepts by the integration of auditory and visual information. Pezzulo and Calvi (2011) also implemented Barsalous PSS as a computational model. Their architecture learns perceptual symbols and assembles them in simulators for perceptual and abstract categories.

With respect to learning paradigms, we can distinguish two types of connectionist models that link subsymbolic (conceptual) knowledge with (linguistic) symbols. The supervised approach is based on the error correction learning in which input patterns are linked with symbolic targets (labels). For instance, Cangelosi and Harnad (2000) developed a supervised computational model able to ground proper features to particular words. They distinguish two types of category learning, namely sensorimotor toil standing for concept acquisition from sensory inputs, and the symbolic theft based on the sharing of already grounded linguistic descriptions between agents. In the follow-up model (Cangelosi & Riga, 2006), linguistic inputs are linked with sensorimotor outputs using back-propagation algorithm. Both inputs and outputs are assumed to come from the environment, and the trained model not only provides an account for grounding linguistic symbols (sensorimotor toil), but also for the grounding transfer to novel linguistic expressions (symbolic theft). This mechanism allows to learn new words without a visual demonstration, which is closely related to the PSS. Tikhanoff, Cangelosi, Fitzpatrick, Metta, Natale, and Nori (2008) extended this supervised architecture and implemented it into a humanoid robot. The robot was able ground words and to understand sentences such as "put red sphere into container".

The unsupervised approach treats both perceptual stimuli and symbols equally as inputs, to be associated (typically) by Hebbian-like learning. This implies a different way of incorporating the symbolic (lexical) level. The target signal only functions as an additional input rather than being the source for error-based learning. The unsupervised models are typically based on self-organizing maps (SOM; Kohonen, 1990) that organize (high-dimensional) input vectors according to their similarities (topographically). For instance, DevLex model (Li, Farkaš, & MacWhinney, 2004) also consists of two self-organizing networks, one for lexical symbols and the other for conceptual (semantic) representations, that are bidirectionally connected. They can activate each other but there is no additional layer for multimodal representations. Within unsupervised approaches there emerged an alternative to link both perceptual and symbolic information with multimodal representations at the output. The example of this architecture is the unsupervised feature-based model, that was used to account for early category formation in young infants (Gliozzi, Mayor, Hu & Plunkett, 2009). This approach postulates the unsupervisory role of linguistic labels that can effect categorization during the acquisition process, which has also been supported by experimental evidence.

Our model is conceptually similar to that of Gliozzi et al (2009) but architecturally it is more complex, since it was tested in another domain of human cognition. We test our model in the area of spatial cognition, similarly to Regier (1996), who created a supervised neural network model consisting of several modules to ground the spatial terms. The architecture was able to ground both static spatial relations (e.g. left, right) and dynamic relations (e.g. around, through). However, in Regiers model the symbolic representational level was considered to be prior and fixed. On the contrary, we focus on unsupervised learning of spatial relations of two objects in 2D space, by linking the perceptual information and the linguistic description. The neural architecture we propose satisfies the requirement that the artificial system (agent) should learn its own functions and representations (Ziemke, 1999). In contrast to the classical top-down approach, our bottom-up approach restricts the designer's intervention in the representational system to a minimum. All representations are learned from the external environmental inputs.

In the basic model (Vavrečka, 2007) we simplified inputs for the unimodal layers to Cartesian coordinates of stimuli (i.e. 2D spatial positions) instead of the (high-dimensional) retinal images as visual inputs, and only two phonetic features instead of more detailed phonological representations as an auditory input. The multimodal layer consisted of 5 neurons representing basic spatial locations. The system reached 87% accuracy, that quantifies the degree of unambiguity of output representations. From the mapping perspective, the system learned to unambiguously map the pairs of output vectors (from the unimodal layers) to single units. The system was able to create perceptually grounded representations. The extended version of the model for the representation of dynamic spatial terms (Vavrečka, 2008) based on the RecSOM (Voegtlin, 2002) and the growing-when-required networks (Marsland, Shapiro & Nehmzow, 2002) was able to process visual sequences (around, through, outside, over, under) and it reached 88% accuracy.

The models described in this paper are the most recent implementations of our architecture. Our aim was to develop a biologically inspired model, so there are some changes in the architecture compared to the previous versions. From the neuroanatomic point of view, the information about the location and identification of an object in space are processed separately in what and where pathways (Ungerleider & Mishkin, 1982). The dorsal where pathway is assumed to be responsible for spatial representation of the object location, while the ventral what pathways are involved in object recognition and form representation. We incorporate this feature in our model. The neurological and psychological evidence also suggests that the brain must somehow integrate various features into a coherent whole. This problem was coined in literature as the (visual) binding problem, that is, a process of linking together the attributes (color, form, motion, size, and location) of a perceptual object. Our model proposes the unsupervised solution of the visual binding based on the integration of what and where pathways (see the review of connectionist approaches to the binding problem in van der Velde & de Kamps, 2006). One of the motivations for our model was to test, whether it is possible to bind location, color and shape of two objects without any prior knowledge and without external information. The model also provides a solution to the (unsupervised) symbol grounding that can be considered as a lexical binding. The sequences of symbols (words) processed in the auditory layer are grounded (bound) to proper features from the visual subsystem (shape, color, location).

The rest of the paper is organized as follows. In Section 2, we introduce the architecture in greater detail. Section 3 presents results from four series of simulations. Section 4 covers the discussion and the relation of our model to other models. Section 5 concludes the paper.

2 The Models

In our model, the representation process takes advantage of the unimodal layers of units. The auditory layer represents sentences and the visual layers represents spatial location, shape and color of objects. The multimodal level integrates the outputs of these unimodal layers. In contrast to the classical approaches that postulate the abstract symbolic level as fixed and prior (defined by the designer), in our model it is possible to learn and modify the auditory layer, visual layer and consequently the multimodal level. The schema of the system is depicted in Fig. 1.

In the three simulations, we compare different versions of the visual subsystem, analyzing the distinction between what and where pathways. The results help us to decide, whether this simplification is important for enhancing the overall model performance. The visual system of our model is tested in three different configurations: a single SOM that learns to capture both what and where information (Model 1), two separate SOMs for what and where information (Model 2), and two separate SOMs with reduced where representations (Model 3).

In Model 4, we compare different types of multimodal integration. Inspired by the biological evidence

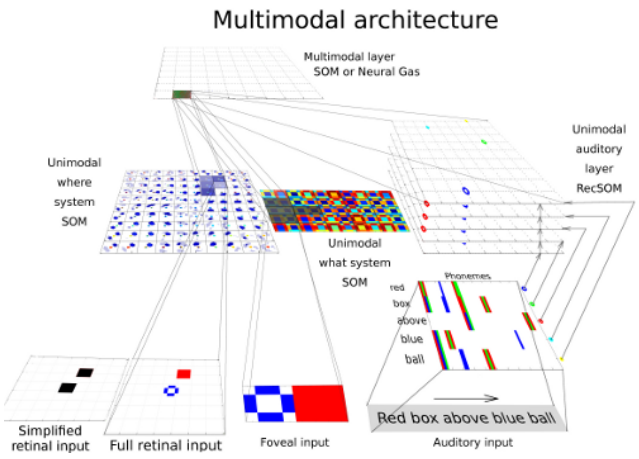

Fig. 1 Multimodal connectionist architecture for grounding spatial terms. The auditory layer represents sentences and the visual layers represent spatial location, shape and color of objects. The multimodal level integrates the outputs of these unimodal layers.

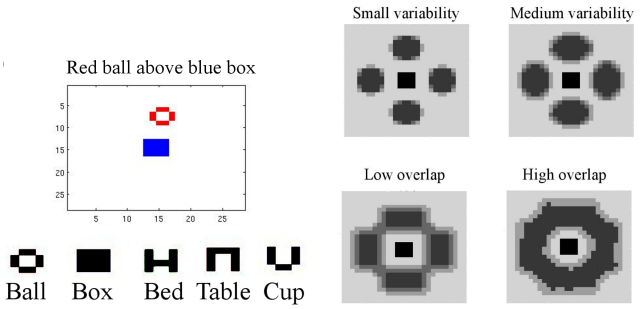

Fig. 2 (a) Example of a visual input scene and the monochrome visual vocabulary, (b) Simplified visual inputs with varying levels of spatial fuzziness.

about topographic organization of sensory and motor brain areas, we assume that primary unimodal layers are topographically organized. Although there exist examples of this organizing principle also in higher areas (Malach, Levy, & Hasson, 2002), it remains an empirical question, whether topographically organized responses are a general principle of the brain also at higher levels of organization. In the multimodal layer, we compare the SOM and neural gas (NG; Martinetz, Berkovich, & Schulten, 1993) algorithms as representatives of both approaches. Both algorithms are unsupervised, based on the competition among units, but NG uses a flexible neighborhood function, as opposed to the fixed neighborhood in SOM (that enforces topography). The goal was to experimentally investigate the effect of the neighborhood function in the multimodal layer. We used the modified SOM Toolbox (Vesanto, Himberg, Alhoniemi, & Parhankangas, 2000) for all simulations.

2.1 Input data

The visual scenes consist of the trajector and the base object in different spatial configurations. The base position is fixed in the center of the scene (the center of retina) and the trajector position is located in one (or at the boundary between two) of the spatial quadrants relative to the base. The positions along the main semi-axes are linguistically referred to as up, down, left, and right, but perceptually, the trajector position is fuzzy and random. The scene size (artificial retina) is 28×28 pixels and both objects consist of 4×4 pixels (Fig. 2a). We trained various models with an increasing combinatorial complexity, starting with simple inputs with 2 colors, 2 object types and 4 spatial relations, up to more complex inputs consisting of 3 colors (red, green, blue), 5 object types (box, ball, table, cup, bed) and 4 spatial relations (above, below, left, right).

The most complex scenario with 2 different objects in the scene amounts to 840 combinations. The corresponding training set resulted in 42000 examples (50 instances per spatial configuration). We also presented stimuli with increasing fuzziness in the spatial location to investigate the relation between fuzziness and the error in the visual and multimodal layer. The two conditions with the highest degree of fuzziness yield overlapping inputs (as seen in Fig. 2b).

2.2 Visual layer

The sensory input of the visual subsystem is captured by an artificial retina that serves as an input to the primary visual layer. Visual layer consists of the SOM(s) that learn the nonlinear mapping of input vectors to output units in the topography-preserving manner (i.e. similar inputs are mapped to neighboring units in the map). The SOM performs standard computations in each iteration. After presentation of a randomly chosen (rescaled) input vector $\mathbf{x}$, the output y_i of a unit i in the SOM is first computed as

$$y_i = 1 - \|[\mathbf{x}(t) - \mathbf{w}_i(t)\|$$

where $\|\|$ denotes the Euclidean norm (also in forthcoming equations), and then the k-WTA (winner-take-all) rule is applied. According to k-WTA, k most active units are proportionally kept active (with the activity of the best matching unit scaled to 1), and all other units are clamped to 0. In the models, we used $k = 6$. The motivation for this type of output representation consists in introducing some overlaps between similar patterns to facilitate generalization.

The output vectors of all unimodal modules are concatenated and serve as the input vector to the multimodal layer. For all visual maps, standard computations are performed regarding the weight update. After the best matching unit (winner) c has been found according to

$$c = \arg\min_i \{\|\mathbf{x}(t) - \mathbf{w}_i(t)\|\},$$

the weights in the winners neighborhood are updated as

$$\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + \mu\ h_{ci}(t)\ [\mathbf{x}(t) - \mathbf{w}_i(t)]$$

where μ is the learning rate, and $h_{ci}(t)$ is the neighborhood kernel around the winner c, with neighborhood radius linearly shrinking over time. Let us now take a more detailed look at these layers and their inputs.

In Model 1, the single SOM was tested whether it could learn to differentiate various positions of two objects, as well as object types and their color. In Model 2, we used separate SOMs for spatial locations (abstraction of where system) and separate SOM for color and shape of objects (abstraction of what system). The what system incorporates a simple attentional mechanism and represents the foveal input of two consequently observed objects. Two visual fields (each with 4×4 receptors) project simultaneously visual information about the trajector and the base in a fixed position to the unimodal what system. The color of each pixel is encoded by the activity level, rescaled to values between 0 and 1. Model 3 employs the same what and where systems as Model 2, but uses different inputs to the where system consisting of two monochromatic boxes (rather than concrete object shapes in color) in the particular spatial position. The size of all visual layers was fixed for all models, namely 30×30 neurons for the where system and 25×25 neurons for the what system. The sizes were estimated from previous simulations and they also stem from the number of combinations in the most complex scenario (840 combinations in where system and 210 in what system). All SOM maps have a hexagonal neighborhood function and the lattices with a toroid topology.

2.3 Auditory layer

Auditory inputs (English sentences) were encoded as high-dimensional patterns representing word forms using PatPho, a generic phonological pattern generator that fits every word (up to three syllables) onto a template according to its vowel-consonant structure (Li, 2002). PatPho uses the concept of a syllabic template: a word representation is formed by combinations of syllables in a metrical grid, and the slots in each grid are made up by bundles of features that correspond to consonants and vowels.

In our case of 5-word sentences, each sentence consists of five 54-dimensional vectors with component values in the interval (0,1), representing particular words. These vectors are sequentially fed (one at a time) to the RecSOM (Voegtlin, 2002), a recurrent SOM architecture, that uses a detailed representation of the context information (the whole output map activation) and has been demonstrated to be able to learn to represent much richer dynamical behavior (Tiňo, Farkaš, & van Mourik, 2006), compared to other recurrent SOM models (Hammer, Micheli, Sperduti, & Strickert, 2004). RecSOM learns to represent inputs (words) in the temporal context (hence capturing sequential information). RecSOM output, in terms of map activation, feeds to the multimodal layer, to be integrated with the visual pathway. Like SOM, RecSOM is trained by competitive, Hebbian-like algorithm. As a property of RecSOM, its units become sequence detectors after training, topographically organized according to the suffix (most recent words).

Formally, each neuron $i \in \{1, 2, ..., N\}$ in RecSOM has two associated weight vectors: $\mathbf{w}_i \in \mathcal{R}^n$ - linked with an n-dimensional input $\mathbf{s}(t)$ (in our case, the current word, with dimension $n = 54$) feeding the network at time t, and the weight vector $\mathbf{c}_i \in \mathcal{R}^N$ – linked with the context $\mathbf{y}(t-1) = [y_1(t-1), y_2(t-1), ..., y_N(t-1)]$ containing the unit activations $\mathbf{y}_i(t-1)$ from the previous time step. The output of a unit i at time t is $y_i(t) = \exp(-d_i(t))$, where

$$d_i(t) = \alpha\ \|\mathbf{s}(t) - \mathbf{w}_i\|^2 + \beta\ \|\mathbf{y}(t-1) - \mathbf{c}_i\|^2.$$

Here, $\alpha > 0$ and $\beta > 0$ are model parameters that respectively influence the effect of the input and the context upon neurons profile. Their suitable values are usually found heuristically (in our model, we use $\alpha = \beta = 0.1$). Both weight vectors are updated using the same form of a SOM learning rule

$$\Delta\mathbf{w}_i = \gamma\ h_{ci}\ (\mathbf{s}(t) - \mathbf{w}_i),$$
$$\Delta\mathbf{c}_i = \gamma\ h_{ci}\ (\mathbf{y}(t-1) - \mathbf{c}_i),$$

where $c = \arg\min_i\{d_i(t)\}$, is the winner index at time t, and $0 < \gamma < 1$ is the learning rate. (The winner can be equivalently defined as the unit c with the highest activation $y_c(t) : c = \arg\max_i\{y_i(t)\}$). Neighborhood function h_{ci} is a Gaussian (of width σ) on the distance $d(i, c)$ of units i and c in the map: $h_{ci} = \exp(-d(c, i)^2/\sigma^2)$. The neighborhood width σ linearly decreases in time to allow formation of topographic representation of input sequences. After training, all RecSOM units be-

come sensitive to particular sentences, ordered topographically according to sentence endings. The output vector is composed of five winners representing particular words in the sentence. The activations of winning units are slowly decayed in time (decreased by value 0.1 at each step) towards the end of sentence. This allows to represent the order of winners in the sequence, hence differentiating between similar base and the trajector phonetic features in a scene (e.g. "red ball above red ball").

2.4 Multimodal layer

The multimodal layer is the core of the system, since it learns to identify unique categories and represent them. In agreement with the theory of perceptual symbols systems (Barsalou, 1999), the main task for this layer is to process the output from the unimodal layers and to find and learn the categories by mapping different sources of information (visual and auditory) that refer to the same objects in the external world. Inputs for the multimodal layer are taken as concatenated unimodal activation vectors (from both modalities) using the above mentioned k-WTA mechanism, explained in Section 2.2. Unlike sparse localized output codes ($k = 6$) used at the unimodal layer (to facilitate generalization), the output representation in the multimodal layer with WTA mechanism is chosen to be localist ($k = 1$) for better interpretation of results and the error calculation.

We tested two unsupervised algorithms in the multimodal layer, SOM and NG, that differ in the neighborhood function. The size of the multimodal layer was set to allow a distinct localist representation of all 840 object combinations in the most complex data set, so we used 841 neurons (arranged in a 29×29 grid in case of SOM).

For clarity, we explain the NG algorithm briefly here. NG shares with SOM a number of features. In each iteration t, an input vector $\mathbf{m}(t)$ is randomly chosen from the training dataset. Subsequently, for all units we compute $d_i(t) = \|\mathbf{m}(t) - \mathbf{z}_i\|$ and then sort the units according to their increasing distances d_i, using indices $l = 0, 1, ...$ (where $l(0)$ corresponds to the current winner). Then we update all weight vectors $\mathbf{z}_i$ according to

$$\Delta\mathbf{z}_i = \eta \ \exp(-l(i)/\lambda) \ (\mathbf{m}(t) - \mathbf{z}_i)$$

with η being the learning rate and λ the so-called neighborhood range. We used $\eta = 0.5$ and $\lambda = n/2$ where n is the number of neurons. Both parameters are reduced with increasing t. It is known that after sufficiently many adaptation steps the feature vectors cover the data space with minimum representation error (Martinetz, Berkovich, & Schulten, 1993). Mathematically, the adaptation step of the NG can be interpreted as gradient descent on a cost function.

3 Results

We present results corresponding to the four models as described above in Section 2, tracking our "experimental trajectory", along which we eventually converged to Model 4. We trained each model for 100 epochs and tested it with a novel set of inputs. For each run, the data set was randomly split to training and testing subsets using the 70:30 ratio.

3.1 Quantification of the model accuracy

To quantify the model accuracy, we designed the following evaluation procedure for each trained model. After the model has been trained, we again ran once through the training set, in order to label all neurons, reflecting their responsiveness to each of the five input features (base color, base shape, spatial location, trajector color, and trajector shape). We attach to each neuron five counter arrays, initialized to zeros, each consisting of $n(f)$ slots, with $n(f)$ being the number of different (possible) values of feature f (depending on the task complexity). For each training input pattern, we find the winner (as in the SOM algorithm) whose five counter values are increased by one (i.e. for each current feature value). After the sweep through the training set, we assign unique feature labels to all neurons by applying the "maximum response principle," according to which each neuron becomes a representative of only the most frequent value of the given feature (for which that neuron became the winner most often). Then we measure the model accuracy, as the percentage of correctly classified test inputs. The feature of the testing pattern is assumed to be correctly classified, if it matches the winners representative feature. We first calculate the error for each feature separately, and then also the overall error for the whole scene/sentence that requires that all features in the testing sentence be correctly classified.

3.2 Model 1

In Model 1, the single SOM in the visual system is tested whether it can learn to represent all visual features simultaneously. We observe a high error in the where system for the trajector features, because trajector positions overlap in the specific area. Although

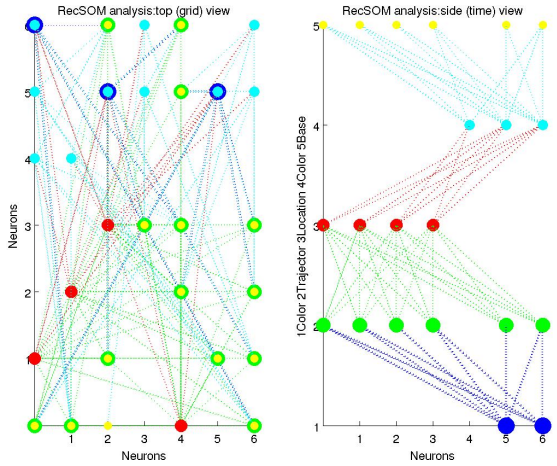

Fig. 3 The unit responses in the auditory layer of Model 1. If the same neuron responds to the same feature (e.g. shape) of the trajector and the base (overlapping dots), it will increase the error for the whole scene/sentence as well. (a) Visualization of the RecSOM grid (time is represented as a size of the dot, shown bottom-up 4b), (b) Timeline of the sentence processing (y-axis) bottom-up.

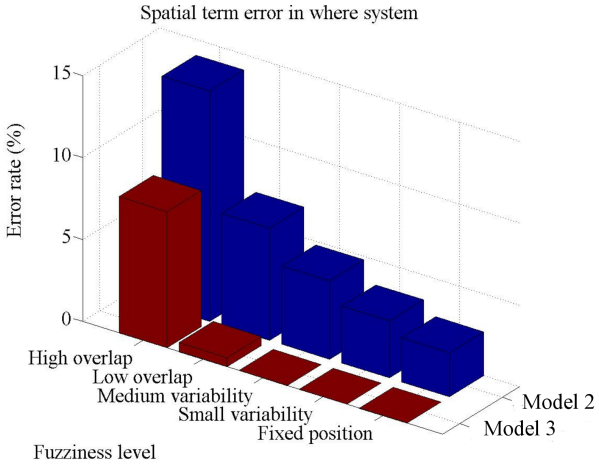

Fig. 4 Visualization of spatial term errors in the where layer for full retinal inputs (blue) and for bounding box inputs (red) as a function of the fuzziness level of trajector spatial location.

the spatial location of the trajector is fuzzy, the error for this feature in the test set is the lowest (14%). Low errors also result for base color (18%) and base shape (28%). Errors for trajector color (37%) and trajector shape (65%) are rather high. We also test whether the level of fuzziness (shown in Fig. 2b) affects the error in the SOM map. All features but the spatial location are not sensitive to the fuzziness level, as the errors vary within 3% range. On the other hand, the error for spatial location correlates with the fuzziness starting from 3% for fixed position of the trajector to 14% for highly overlapping spatial locations.

The auditory RecSOM layer performs better compared to the visual layer because the phonetic features, being sequentially fed to the system, are not fuzzy. There are 0% errors for base color, base shape, trajector color and spatial term. Error for the trajector shape is 1%. On the other hand, there is a 22% confusion error. The RecSOM architecture allows a neuron to become sensitive for multiple instances of the same word in the sentence, because it represents each sentence (sequence) in 2D grid. This neuron is sensitive to e.g. both red color of the trajector and the base. It results in the confusion of the neuron response (see Fig. 3) and increases the error in multimodal layer. This problem should be partially eliminated by decayed activation of winning neurons (see Auditory layer).

The performance of the multimodal layer heavily depends on the effectiveness of unimodal layers. The errors for the representation of trajector color (8%), base color (1%) and base shape (2%) are low. On the other hand, there are high errors for both the trajector shape (46%) and spatial term (25%). This is due to bad performance of the visual layer. The overall error of the system reaches 68%.

3.3 Model 2

Model 2 processes what and where information using separate SOMs, and we identify a difference in accuracy between the two systems. The what system outperforms the where system, as documented by low errors for base color (1%), base shape (8%), trajector color (0%) and trajector shape (5%). We did not test the performance of the what system for the spatial term simply because that information was not made available to the what system. The errors for the where and auditory systems are identical to Model 1, because these layers receive the same inputs as in Model 1. Notably, the additional what layer changed the performance of the multimodal layer. Errors for base color (2%) and base shape (4%) in the multimodal layer remain the same as in Model 1, but lower errors are observed for trajector color (1%) and trajector shape (5%). On the other hand, the system exhibits a much higher error for the spatial term (71%) compared to Model 1 (25%). The multimodal SOM layer is probably not able to merge the information from three unimodal layers. The overall error is 75%, caused by the problem with the representation of the spatial term. The more detailed analysis is postponed to Discussion.

3.4 Model 3

The simplification of inputs to the where system is achieved by using monochromatic bounding boxes in-

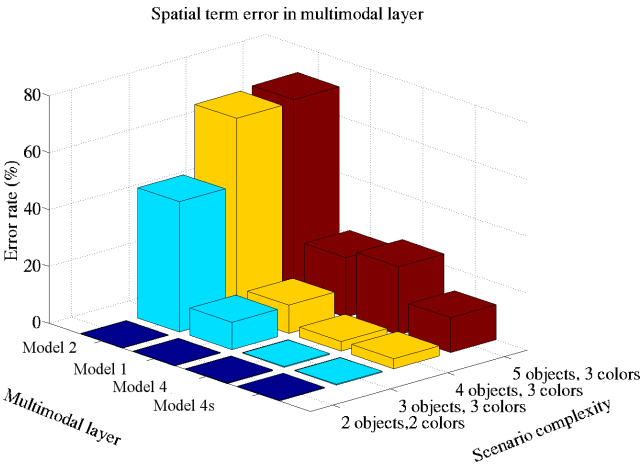

Fig. 5 Comparison of the errors in the multimodal layer for the representation of the spatial term. Model 4s (NG in the multimodal layer and a single SOM in the visual system) performs best.

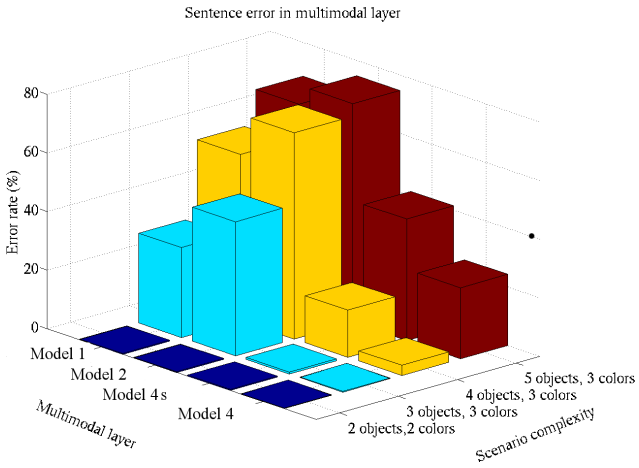

Fig. 6 Errors in the multimodal layer for whole scene (sentence) representation. Model 4 based on what and where visual system and NG in multimodal layer performs best.

stead of object shapes and colors. This expectedly led to lower errors compared to full retinal images (see Fig. 4). We do not compare the results for object features (shape and color), because in this model there is no information about them provided to the where system. The analysis of the SOM structure revealed a better organization of specific clusters in favour of bounding box inputs for the spatial term representation. These results lead us to the conclusion that it is possible to simplify the information projected to the where system to optimize the speed and effectiveness of our architecture. However, the simplification of the where inputs does not affect the performance of the multimodal layer. There are similar results for the object features, spatial term (70%) and also overall error (74%). So we tested the NG algorithm in the multimodal layer in further simulation trying to improve the performance.

3.5 Model 4

We compare the effectiveness of the SOM and NG algorithms in the multimodal layer. We observe a different type of clustering in the unimodal layers that are transferred to the multimodal layer, where the SOM is not able to adapt to the joint outputs from unimodal layers, apparently due to neighborhood constraints (Model 1 and 2). The results of the NG algorithm (Model 4 and 4s) for the same input data confirm this hypothesis. The multimodal layer based on NG is able to correctly map all the object features without any problem. There is a 0% error for both simplified inputs (Model 3) and also for full retinal projections to the where system (Model 2). The errors for multimodal NG module and the single SOM in the visual layer (Model 4s) are as follows: 1% for base color, 2% for base shape, 6% for trajector color and 26% for trajector shape. These results are significantly better than those for the multimodal SOM. Surprisingly, we observe the lowest error for the representation of the spatial term in the multimodal layer for NG algorithm and a single SOM visual layer (Model 4s). There is a 12% error compared to 24% for Model 4 (see Fig. 5). The SOM algorithm leads to higher errors of the spatial term for both models, namely 25% (Model 1) and 70% (Model 2). These results are contradictory, because Model 2 (and also Model 3) with separate what and where systems performs better for all features except the spatial term. This could be attributed to the missing information about the spatial term in the what system (see Discussion).

The comparison of the overall accuracy (whole sentence error) is shown in Fig. 6. The best results are obtained for what and where subsystems and the NG algorithm in the multimodal layer (Model 4). There is a 25% error compared to 70% overall error for multimodal SOM in the most complex scenario. Hence, the better, albeit not perfect, results are achieved with NG by sacrificing the topographicity of responses in the multimodal layer.

The last analysis is dedicated to the comparison of SOM (Model 3) and NG (Model 4) algorithms in the multimodal layer that have to process different levels of spatial fuzziness. Fig. 7 reveals a lower error for NG at all levels of fuzziness and the high errors for SOM regardless of the fuzziness level (70%). Hence, the multimodal SOM is unable to represent neither fuzzy nor distinct inputs.

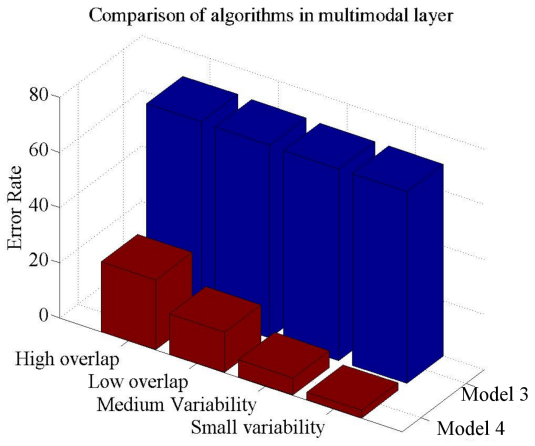

Fig. 7 Errors in the multimodal layer for SOM (Model 3) and NG (Model 4) algorithms as a function of the fuzziness level of the trajectors spatial location (see Fig. 2b).

4 Discussion

We analyze the presented models in the context of theoretical assumptions, especially the perceptual theory of cognition and conceptual approaches to knowledge representation. We also discuss various aspects of our model, its relation to other models and the features of lexical and visual binding.

4.1 Architecture

In our model, the representations take advantage of the two or three unimodal layers of units. The auditory layer represents unique labels (linguistic terms), whereas the where part of the visual system represents fuzzy information about the spatial locations of objects in the external world and what system captures shapes and colors of objects in a fixed foveal position. The multimodal level integrates the outputs of these unimodal layers. The grounded meaning is simultaneously represented by all layers (auditory, visual and multimodal), making this approach resemble the theory of Peirce (1931) who defined three components of a sign – representamen, interpretant and the sign itself. Our model represents the sign hierarchically which guarantees better processing and storing of representations, because the sign (the multimodal level) is modifiable from both modalities (the sequential "representamen" auditory level and the parallel "interpretant" visual level). This feature makes the units in the higher layer bimodal (i.e. they can be stimulated by any of the primary layers) and their activation can be forwarded for further processing. Bimodal (and multimodal) neurons are known to be ubiquitous in the association areas of the brain (Stein & Meredith, 1993). The multimodal layer is formed by exploiting the concept of self-organized conjunctive representations that have been hypothesized to exist in the brain with the purpose of binding the features such as various perceptual properties of objects (Mel & Fiser, 2000). Here we extend the concept of grounding by linking the subsymbolic and symbolic information. Hence, each output unit learns to represent a unique combination of perceptual and symbolic information (that could be forwarded to another, higher module).

Interestingly, the bimodal layer with conjunctive units is also used in recent generative probabilistic models that can be designed to link information from two (or more) sources (e.g. modalities). For example, the deep belief net (DBN) is a stochastic generative model (a multi-layer neural network with bidirectional connections) that learns to approximate the complex joint probability distribution of high-dimensional data in a hierarchical way. For instance, DBN was trained to classify the isolated hand-written digits into 10 categories, so the visual inputs (28×28 pixel images) were linked with categorical labels (Hinton, Osindero, & Teh, 2006). The linking was established via the training on image-label pairs (treated as inputs), using the higher (bimodal) layer (with 2000 units) that learned the joint distribution of those input pairs. DBN was shown to be superior to various other (discriminatory) models in this digit classification task. From the perspective of the representations formed in the multimodal units, Hinton et als goal was the same as ours (although our units are deterministic rather than stochastic). The separate multimodal level provides a platform for the development of subsequent stages of information processing (e.g. inference mechanisms). Further tests of this approach should also focus on scaling up our model to more complex mappings.

The architecture of our model shares some similarities with, but also differs from the DevLex model of early lexical acquisition (Li et al., 2004). DevLex, originally inspired by DISLEX model (Miikkulainen, 1997) also consists of two (growing) self-organizing maps, but these are directly interconnected. DevLex was proposed to learn the form-meaning associations (phonological word forms and meanings) via Hebbian updating the (bidirectional) connection links, aiming to model the processes of lexical comprehension and production. DevLex, however, does not contain a higher (e.g. multimodal) layer that integrates the modalities, as do other grounding models (Riga, Cangelosi & Greco, 2004; Roy, 2005). Instead, the overall representation of the meaning is taken as the joint coactivation of the two maps.

Our model is very similar to the connectionist model of Dorffner et al. (1996) that consists of two primary

(symbolic and conceptual) layers connected to one central layer. There is a linking layer (the counterpart of our multimodal layer) interconnecting the two primary layers via localist units that link both representations (i.e. one unit connects one word-concept pair of primary representations). First, one set of links (weights to the linking layer) is trained using a competitive mechanism exploiting the WTA approach. Then, the winners weights towards the other layer are updated according to the outstar rule (Grossberg, 1987). Hence, the purpose is to learn form-concept mapping, mediated by the linking layer. In both models, these mappings were aimed at simulating the word comprehension (form to meaning) and the word production (meaning to form), but our model is also able to bind visual features and also bind proper lexical units in the sentence to the visual counterparts (lexical binding or extended symbol grounding).

4.2 Visual binding

Our model proposes the unsupervised solution to the visual binding, based on the integration of what and where pathways. With respect to the visual binding problem, the model is based on convergent hierarchical coding, also called *combination coding* (Riesenhuber & Poggio, 2002). The neurons react only to combinations of features, that is, to an object of a particular shape and color at a particular retinal position (localist representation). The hierarchical processing implies that increasingly complex features are represented by higher levels in the hierarchy. Complex objects and situations are constructed by combining simpler elements. On the other hand, the convergent hierarchical coding requires as many binding units as there are distinguishable objects. It should result in a combinatorial explosion for large-scale simulations. Our model is able to represent 840 combinations, but it can also suffer from combinatorial explosion because we represent pairs of objects instead of separate entities in the primary layers. In case of 10 objects, 5 colors in 4 spatial locations we would need to represent 2450 object pairs in a primary what system, instead of 50 separate objects. It is also possible to add a separate layer for the color processing, in which case there will only be 10 objects presented in the what system (we plan to test this architecture in the future). Alternatively, we could represent the features in the activity of a population of neurons distributed within and across levels of the cortical hierarchy as distributed representation (Goldstein, 2002), although some authors have raised the question whether the combinatorial explosion is really a problem (Ghose & Maunsell, 1999). It is estimated that the number of objects, scenarios, colors and other features in the brain is approximately 10 million items. It is obviously beyond the limits of recent cognitive systems, but it is below the number of neurons in the mammalian visual cortex, so the combination coding could be a sufficient method. We could also adopt Neural Modeling Fields (Perlovsky, 2001), the unsupervised learning method based on Gaussian mixture models that arguably does not suffer from combinatorial complexity. The application of this theory to the area of symbol grounding resulted is 95% accuracy of the system that learned repertoire of 112 actions (Cangelosi, Tikhanoff, Fontanari, & Hourdakis, 2007).

4.3 Lexical binding

Our model is able to map the words in the sentence with the fixed grammar to the objects in the environment without any prior knowledge (lexical binding). Previous models of symbol grounding (Cangelosi, Greco & Harnad, 2000; Cangelosi & Parisi, 2004; Cangelosi & Riga, 2006; Cangelosi et al, 2007) deal with the lexical level but our model goes beyond words because it can represent sentences in RecSOM. The ability of lexical binding should be considered as an extension of the symbol grounding. Cangelosi et al. (2000) recommend to ground specific words (sensorimotor toil) at the first stage and then compositionally chain them in the grounded language level (symbolic theft). There are separate objects presented to their system within a training phase, grounding basic object features. Our approach can be considered an alternative to this theory. We also ground words at the first stage, but unlike the mentioned approach, we present sentences as linguistic inputs to be bound with proper features from the visual subsystem (shape, color, location). Compared to the classic sensorimotor toil experiments based on the grounding of two features, our system is able to ground 5 features simultaneously that speeds up the process of symbol grounding (faster acquisition of the grounded lexicon). Tikhanoff (2009) proposed the architecture (and implemented it in iCub robot) that was able to understand basic sentences but it was based on supervised learning. Our model is a proof of concept that also unsupervised architectures can find proper mapping between visual and lexical features. We are able to build representations solely from the sensory inputs, arguing that the co-occurrence of inputs from the environment is a sufficient source of information to create an intrinsic representational system.

4.4 Performance

The analysis of model behavior revealed that the trajector shape and the spatial term representations are the most difficult subtasks for visual unimodal systems. The difficulty is caused by the variability and fuzziness of these inputs. The correct representation of the trajector shape requires a separate unimodal what system. The errors for Model 1 and Model 2 confirm the necessity of the what system in the complex environment because we observe a 60% increase of errors in the model without a separate what system. On the other hand, the error for the spatial term in Model 4 reflects some problems with an increasing number of inputs to the multimodal layer, because there is a lower error for Model 1 compared to Model 2. The problem could reside in the number of dimensions. The multimodal module receives a 1300-dimensional input in Model 1 and a 1925-dimensional input in Model 2. The increase of dimensionality together with a localist unimodal output function may decrease the effectiveness for the spatial term representation, although other features are represented better in high-dimensional space. This contradiction has to be investigated in a greater detail.

The results for specific algorithms in the multimodal layer confirm our hypothesis that the SOM algorithm based on the fixed neighborhood function is not able to adapt to the distribution of the joint outputs from unimodal layers. The SOM-based models show a topology-preserving property for the input data, but they are weak with regard to properly represent clusters with different non-uniform data distributions (Kim, Sang-Woo & Minho, 2011). Our results are also in line with Pezzulo and Calvi (2011) who conclude that perceptual symbols may not be topographically organized, although some parts of the perceptual and motor areas show topological hierarchical organization. There also exist grounding models based on topologically organized connectionist networks (e.g. Joyce, Richards, Cangelosi & Coventry, 2003) to simulate the perceptual symbol system, but our results do not confirm this assumption.

The mapping in our models is actually a clustering process that makes the system also vulnerable to errors in the input space. If (at least) one perceptual input creates discrete clusters, successful learning can be achieved. In case of all fuzzy sources of information, it is difficult to create a system that is able to provide (without any additional information or supervision) a successful mapping, e.g. to learn a new meaning of spatial position at the boundary of two spatial areas (e.g. below and right) and the auditory information (beright). In other words, the successful clustering presumes that at least one modality provides distinct activation vectors for different classes to drive the clustering process (i.e. the classes are well separable in the corresponding input subspace). On the other hand, the occurrence of both auditory and visual fuzzy inputs is rare in the real world, so our system could be considered a step towards the solution for symbol grounding problem (at least at this small scale).

5 Conclusion

We have created a system that is able to extract constant attributes and regularities of the environment and identify them with abstract symbols. The meaning is nonarbitrarily represented at the conceptual level that guarantees the correspondence of the internal representational system with the external environment. We can also conclude that it is advantageous to follow the biologically inspired hypothesis about processing of visual information in separate subsystems. The question for the future research is to find a proper way of output coding from unimodal layers to increase system accuracy and to scale up the model. The main advantage of our model is the hierarchical representation of the sign components.

Acknowledgements This work has been supported by research program MSM 6840770012 of the CTU in Prague, SAIA scholarship and GAČR grant P407/11/P696 (M.V.) and by VEGA grant 1/0439/11 (I.F.).

References

1. Barsalou, L.W. Perceptual symbol systems. Behavioral and Brain Sciences, 22, 577–609 (1999).
2. Barsalou, L.W., Santos, A., Simmons, W.K., & Wilson, C.D. Language and simulation in conceptual processing. In: M. de Vega, A.M. Glenberg & A.C Graesser (eds), Symbols and Embodiment: Debates on Meaning and Cognition, Oxford University Press, 245–283 (2008).
3. Dorffner, G., Hentze, M. & Thurner, G. A connectionist model of categorization and grounded word learning, in Koster C., Wijnen F. (eds.): Proceedings of the Groningen Assembly on Language Acquisition (GALA '95) (1996).
4. Cangelosi, A, Greco, A., & Harnad, S. From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. Connection Science, 12(2), 143–62 (2000).
5. Cangelosi, A. & Parisi, D. The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. Brain and Language, 89(2), 401–8 (2004).
6. Cangelosi A, Riga, T. An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. Cognitive Science, 30(4), 673–89 (2006).

7. Cangelosi, A., Tikhanoff, V., Fontanari, J.F., & Hourdakis, E. Integrating language and cognition: A cognitive robotics approach. IEEE Computational Intelligence Magazine, 2(3), 65–70 (2007).
8. Farivar, R. Dorsal-ventral integration in object recognition. Brain Research Reviews, 61(2), 144–153 (2009).
9. Fontanari, J.F., Tikhanoff, V., Cangelosi, A., Ilin, R., & Perlovsky, L.I. Cross-situational learning of object-word mapping using Neural Modeling Fields. Neural Networks, 22, 579–585 (2009).
10. Goldstein, E.B. Wahrnehmungspsychologie. Spektrum Akademischer Verlag (2002).
11. Gliozzi, V., Mayor, J., Hu, J-F., & Plunkett, K. Labels as features (not names) for infant categorization: A neurocomputational approach. Cognitive Science, 33(4), 709–738 (2009).
12. Ghose, G.M. & Maunsell, J. Specialized representations in visual cortex: A role for binding? Neuron, 24, 79–85 (1999).
13. Greco A., Caneva C. Compositional symbol grounding for motor patterns. Frontiers in Neurorobotics, 4, 111. doi: 10.3389/fnbot.2010.00111 (2010).
14. Grossberg, S. Competitive learning: From interactive activation to adaptive resonance. Cognitive Science, 11(1), 23-63 (1987).
15. Hammer, B., Micheli, A., Sperduti, A., & Strickert, M. Recursive self-organizing network models. Neural Networks, 17(8-9), 1061–1085 (2004).
16. Hinton, G., Osindero, S., & Teh, Y. A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527–1554 (2006).
17. Joyce D., Richards L., Cangelosi A., Coventry K.R. On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In F. Detje, D. Drner, H. Schaub (Eds.), The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modeling, pp. 147–152, Universitaetsverlag Bamberg (2003).
18. Kim, B, Sang-Woo B., & Minho, L. Growing fuzzy topology adaptive resonance theory models with a push-pull learning algorithm, Neurocomputing, 74(4), 646–655 (2011).
19. Kohonen, T. Self-Organizing Maps. Springer. (3rd edition) (2001).
20. Lakoff, G. Women, Fire, and Dangerous Things. University of Chicago Press (1987).
21. Li, P., Farkaš, I., & MacWhinney, B. Early lexical development in a self-organizing neural network. Neural Networks, 17(8–9), 1345–1362 (2004).
22. Malach, R., Levy, I. & Hasson, U. The topography of high-order human object areas. Trends in Cognitive Sciences, 6(4), 176–184 (2002).
23. Marocco, D., Cangelosi, A., Fischer, K., & Belpaeme, T. Grounding action words in the sensorimotor interaction with the world: Experiments with a simulated iCub humanoid robot. Frontiers in Neurorobotics, 4(7), doi: 10.3389/fnbot.2010.00007 (2010).
24. Marsland, S., Shapiro, J., & Nehmzow, U. A self-organising network that grows when required. Neural Networks, 15(8-9), 1041–1058 (2002).
25. Martinetz, T., Berkovich S., & Schulten, K. "Neural-gas" network for vector quantization and its application to time-series prediction. IEEE Transactions on Neural Networks, 4(4), 558–569 (1993).
26. Mel, B. & Fiser, J. Minimizing binding errors using learned conjunctive features. Neural Computation, 12, 247–278 (2000).
27. Newell, A. & Simon, H. A. Human Problem Solving. Englewood Cliffs, NJ: Prentice-Hall (1972).
28. Peirce, C.S. Collected papers of Charles Sanders Peirce (C. Hartshorne, Ed.). Harvard University Press (1931).
29. Perlovsky, L.I. Neural Networks and Intellect: using model-based concepts. Oxford University Press, New York, NY (2001).
30. Pezzulo, G. & Calvi, G. Computational explorations of perceptual symbol systems theory. New Ideas in Psychology, 29, 275–297 (2011).
31. Regier, T. The Human Semantic Potential: Spatial Language and Constrained Connectionism. Cambridge, MA: MIT Press (1996).
32. Riesenhuber, M. & Poggio, T. Neural mechanisms of object recognition. Current Opinion in Neurobiology, 12, 162–168 (2002).
33. Roy, D., Pentland, A. Learning words from sights and sounds: a computational model. Cognitive Science, 26,113–146 (2002).
34. Roy, D., Hsiao, K., Mavridis, N. Conversational robots: Building blocks for grounding word meaning. Proceedings of the HLT-NAACL Workshop on Learning Word Meaning from Non-Linguistic Data, pages 70–77 (2003).
35. Rogers, T.T. & McClelland, J.L. Semantic Cognition: A Parallel Distributed Processing Approach. MIT Press, Cambridge (MA) (2006).
36. Stein, B. & Meredith, M. Merging of the Senses. Cambridge, MA: MIT Press (1993).
37. Taddeo, M. & Floridi, L. The symbol grounding problem: A critical review of fifteen years of research. Journal of Experimental and Theoretical Artificial Intelligence, 17(4), 419–445 (2005).
38. Voegtlin, T. Recursive self-organizing maps. Neural Networks, 15(8-9), 979–91 (2002).
39. Van der Velde, F. & de Kamps, M. Neural blackboard architectures of combinatorial structures in cognition. Behavioral and Brain Sciences, 29, 37–70 (2006).
40. Tikhanoff, V, Cangelosi, A., Fitzpatrick, P., Metta, G., Natale, L., & Nori, F. An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator. In: Performance Metrics for Intelligent Systems (PerMIS) Workshop, pp. 57–61 (2008).
41. Tikhanoff, V. Development of cognitive capabilities in humanoid robots. PhD thesis. School of Computing, Communications & Electronics, University of Plymouth (2009).
42. Tio, P., Farkaš, I., & van Mourik, J. Dynamics and topographic organization in recursive self-organizing map. Neural Computation, 18, 2529–2567 (2006).
43. Vavrečka, M. Symbol grounding in context of zero semantic commitment (in Czech) In: J. Kelemen, V. Kvasnička (eds.). Kognice a umelý život VII. 1. vyd. Opava : Slezsk univerzita, pp. 365–377 (2006).
44. Vavrečka, M. Grounding of spatial terms (in Czech). In: J. Kelemen, V. Kvasnička (eds.). Kognice a umelý život VII, Opava : Slezsk univerzita, pp. 365–377 (2007).
45. Vavrečka, M. Multimodal representations for symbol grounding (in Czech). In: V. Kvasnička, P. Trebatick (Eds.), Kognice a umlý život VIII, VšE Praha (2008).
46. Vavrečka, M., Farkaš, I., Lhotská, L. Bio-inspired Model of Spatial Cognition. In Lecture Notes in Computer Science 7062 LNCS (PART 1), pp. 443–450 (2011).
47. Vesanto, J., Himberg, J., Alhoniemi, E., & Parhankangas, J. Self-Organizing Map in Matlab: the SOM Toolbox. Proceedings of the Matlab DSP Conference, 35–40 (2000).

48. Zeilinger, H., Perner, A., Kohlhauser, S. Bionically inspired information representation module 3rd International Conference on Human System Interaction, HSI'2010 - Conference Proceedings , art. no. 5514490 , pp. 708–714 (2010).
49. Zimmer, H.D., Mecklinger, A. & Lindenberge, U. Handbook of Binding and Memory. Perspectives from Cognitive Neuroscience. Oxford University Press (2006).

Generated PDF
Click here to download Supplementary Material: paper.pdf