# Self-organisation of Language Instruction for Robot Action Control

Mark Elshaw, Stefan Wermter and Peter Watt
Centre for Hybrid Intelligent Systems
School of Computing and Technology
University of Sunderland
St Peter's Way, Sunderland, SR6 0DD
UNITED KINGDOM
http://www.his.sunderland.ac.uk

*Abstract* – **Most current approaches for robot control do not make use of language and ignore neural learning. However, our robot control approach uses language instruction and draws from the concepts of regional distributed modularity, mirror neuron theory and neural assemblies. We describe a self-organising model that clusters action verbs into different locations of the output layer dependent on the body part they are associated with. In doing so we build on our previous work by using actual sensor readings from the MIRA robot that incorporate semantic features of the action verbs. Furthermore, we outline a hierarchical computational model for a neurally inspired self-organising robot action control system using language for instruction.**

## I. INTRODUCTION

Recently there has been growing interest in learning in robotics. However these approaches rarely use neural networks or language instruction. Furthermore, they are restricted in their general autonomous behaviour and only learn what has been pre-specified and coded. Even the "Talking Heads" approach that incorporates the emergence of language in robots [18] gives little consideration to neuroscience-inspired learning in humans.

Some robots like the tour-guide robot Rhino [5] have been quite robust in terms of their localization and navigation behaviour, however they do not interact via language. Although the conversation office robot jjj-2 [1] can be instructed to navigate to certain landmarks and the Minerva tour-guide [19] interacts by using simply preprogrammed speech, they are restricted in their ability to learn. Furthermore, the Kismet interactive robot [4] can recognise and represent emotions using a static sophisticated head but does not understand or generate real language.

Our approach to robot control using language incorporates some neuroscience evidence related to the architectural and processing characteristics of the brain [20]. In particular it focuses on the neurocognitive evidence on cortical assemblies of Pulvermüller et al., regional modularity in the brain and mirror neuron theory. By building on our previous work this paper describes how these concepts are the basis of a robotic control system using language inputs.

## II. MODULARITY IN CELL ASSEMBLIES

Regional distributed modularity in the brain is based on various distributed neural networks in diverse regions that carry out processing in a parallel distributed manner to perform specific cognitive functions [15]. The brain consists of a group of collaborating networks, none of which can deal with a complex task alone [20]. Brain imaging techniques have identified to a significant extent the distributed regional modularity organisation of language processing in the brain [9]. The early model of language processing based on two cortical regions linked via the arcuate fasciculus [2] has been extended to include additional brain regions. For instance, speech comprehension and information recollection has been observed to involve four regions in the left hemisphere of the cerebral cortex [3] and semantic language operations involve the superior temporal sulcus, middle temporal gyrus, angular gyrus and lateral frontal lobe [7]. Recently, cortical assemblies have been identified in the cortex that activate in response to the performance of motor tasks at a semantic level [13, 16]. This evidence supports that these neurons are involved in actions, observing actions and communicating actions.

The neurocognitive evidence of Pulvermüller, 1999 [13] and Pulvermüller, 2002 [12] supports that cell assemblies are activated in different regions of the brain dependent on the word type being processed. This evidence offers the basis for our approach. Pulvermüller, 1999 [13] noted that activation was found in both hemispheres of the brain for content words and for vision words in the perisylvian and in the parietal, temporal and/or occipital lobes. For content words the cell assemblies that were activated depended on semantic features that come from various modalities and include the complexity of activity performed, the number of muscles used, the colour of the stimulus, the tool used, the smell or taste of the object, and whether the person can see herself doing this activity.

Pulvermüller et al., 2000 [14] when examining the processing of action verbs that relate to the leg, face and arm found that this was done in the brain by activating cell assemblies that are associated through semantic information with the appropriate body part. They found that the average response times for lexical decisions was faster for face-associated words than for arm-associated words and the arm-associated were faster than leg ones. There was a significant difference for the prefrontal region and occipital regions and above the motor and premotor cortex. The prefrontal area was found to be associated mainly with arm verbs and the occipital visual areas for face verbs.

With regards to the mirror neuron theory Rizzolatti and Arbib, 1998 [16] found that neurons located in the F5 area of a primate's brain were activated by both the performance of the action and its observation. The recognition of motor actions comes from the presence of a goal and so the motor system does not solely control movement [8]. The role of these mirror neurons is to depict actions so they are understood or can be imitated. The mirror neuron system was a critical discovery as it shows the role played by the motor cortex in action depiction [17].

## III. SELF-ORGANISING NETWORK

Our robot control approach makes use of self-organising networks that offer an unsupervised associative memory approach. Self-organising networks consist of an input and an output layer, with every input neuron linked to all the neurons in the output layer [10]. The output layer creates a topographical representation that clusters similar inputs together by creating patterns of activation (see Fig. 1).

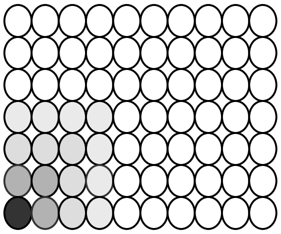

Fig. 1 A representation of the activity output layer of a self-organising network - The darker the neuron the greater the activation.

A typical self-organising network algorithm has an input vector represented as $i = [i_1, i_2, \ldots, i_n]$. The input vector is presented to every output unit of the network, the weights between the links in the network are provided by

$$w_j = [w_{j1}, w_{j2}, \ldots, w_{jn}] \tag{1}$$

where j identifies unit j in the output layer and n is the *nth* element of the input. The output $o_j$ of unit j is established by determining the weighted sum of its inputs, given by:

$$o_j = \sum w_{jk} i_k = w_j \bullet i \tag{2}$$

The weights are initalised randomly and hence a unit of the network will react more strongly than others to a specific input representation. The weight vector of this unit as well as the eight neighbouring units are altered based on the following:

$$\Delta w_{jk} = \alpha(i_k - w_{jk}) \text{ and } w_{jk}(t+1) = w_{jk}(t) + \Delta w_{jk} \tag{3}$$

where α is the learning rate parameter that is usually set between 0.2 and 0.5.

## IV. SELF-ORGANISATION EXPERIMENT FOR ROBOT CONTROL

Through our previous study [6] it was possible to identify that a self-organising network was able to recreate the findings of Pulvermüller et al. on action verb processing related to body parts. However, this approach relied upon subjective interpretations as to the features that are applicable for the action verbs and their values. In order to have greater objectivity and to allow the incorporation of self-organising maps into a robot control system, sensor readings were taken from the MIRA robot (see Fig. 2). Such sensor readings incorporate semantic features to describe the action verbs such as the degree of motion and object manipulation.

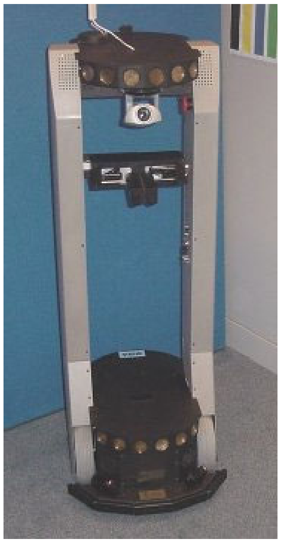

Fig. 2 The MIRA Robot.

### A. Experimental Method

The MIRA robot has a PC, microphone and speakers and a PC104 audio board. Wireless communication between the robot and a computer is used. The robot has an adjustable camera, IR table sensors and a 2-degree gripper

that contains break-beam sensors to detect objects. MIRA can perform neural network based behaviour.
This robot was programmed to perform various actions that are associated in humans with the leg, head or hand, and take sensor readings. The leg verb actions were go, turn left, turn right, forward and backwards; head action verbs were head up, head down, head right and head left; and finally the hand verbs were pick, put, lift, drop and touch. For instance, the hand verb action 'pick' included the following subactions (i) slowly move forward to the table; (ii) tilt camera downwards to see table, (iii) lift gripper to table height; (iv) open gripper; (v) close gripper on object; (vi) stop forward motion; and (vii) lift gripper.

In order to provide sufficient and varied action verb training and test data the actions were repeated 20 times under diverse conditions. For instance, the speed the robot was traveling at, the height of the table the object was on and the angle that the camera was tilted or panned to were varied. The sensor readings were taken 10 times a second while MIRA performed these actions including the state of the gripper, the velocity of the wheels and the angle that the robot's camera was at. The full list of the sensor readings taken that were to act as the semantic features of the action verbs are given in Table I.

To reduce the size of the input to the self-organising network to a manageable level 10 sets of the readings were taken over time to represent the action verb. This was achieved by taking the first, last and eight equi-distant sets of readings and combining them to create a single input for a sample. Various preprocessing activities were performed on the data to make it suitable for introduction into the neural networks. As self-organising networks require the input values to be represented numerically 'yes' was represented as 1 and 'no' 0. The gripper break-beam state values were represented as 'no beams broken' 0.25, 'inner broken' 0.5, 'outer broken' 0.75 and 'both broken' 1. In the case of gripper state 'between open and closed' was given the value 0.3, 'gripper open' 0.6 and 'gripper closed' 0.9.

As self-organising networks typically perform better when input values are between 0 and 1 there was a need to normalise the sensor readings for such variables as velocity of left wheel, velocity of right wheel, x co-ordinate of robot, y co-ordinate of robot, and the pan and tilt of the camera. In the case of x co-ordinates the values varied from –1235 and +1380. Normalisation was done by taking the sensor readings for the specific feature for all samples across the ten sets of readings and positioning the value between 0 and 1 dependent on its relative size. For example, the x co-ordinate values were normalised using the equation (4).

$$\frac{x - \min(x)}{\max(x) - \min(x)} \quad \text{for all x} \qquad (4)$$

TABLE I
SENSOR READINGS TAKEN BY ROBOT DURING ACTIONS.

| Sensor Reading | Value |
|---|---|
| Velocity of left wheel | Real number |
| Velocity of right wheel | Real number |
| X co-ordinate of robot | Real number |
| Y co-ordinate of robot | Real number |
| Break-beam state of gripper | No beams broken, inner broken, outer broken, both broken |
| Gripper state | Gripper fully open, closed, between open and closed |
| Gripper at highest or lowest position | No Yes |
| Gripper moving upwards or downwards | No Yes |
| Table sensors activated | No Yes |
| Gripper opening or closing | No Yes |
| Pan of camera | integer |
| Tilt of camera | integer |

### B. Unsupervised Learning

In the experiment the input layer to the self-organising networks had 120 units, one for each of the preprocessed sensor readings. The output layers had various sizes (from 8 by 8 units to 13 by 13 units) and the networks were trained for between 50 to 500 epochs at intervals of 50 epochs. Fig. 3 provides an example self-organising network showing the input and output for a 'pick' action verb training sample [11]. The number of training and test samples for each action were 15 and 5 respectively. The location of each of the training and test samples on the self-organising output layers were identified based on the units that had the highest activation.

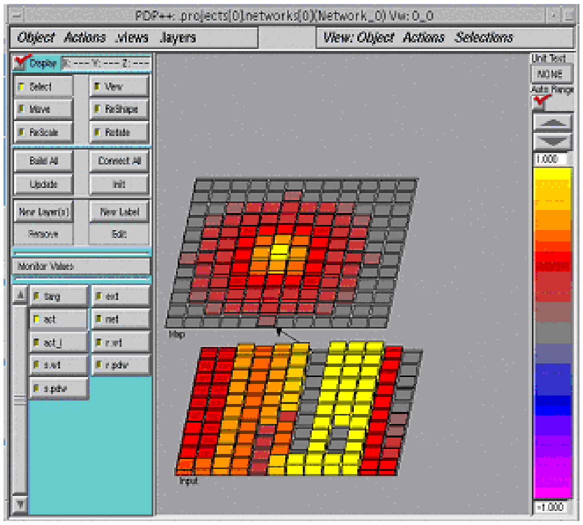

Fig. 3 Example self-organising network showing the input and output for a pick training sample.

# V. RESULTS AND DISCUSSION

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | |
| | | | | | Forward<br>T. Left<br>T. Right<br>Go | | | | | | |
| | | | | | | | | | | | H. Right |
| | | | | | | Go | | | | | |
| | | | | | H. Down | T. Right | | | Backwards | H Left | |
| | | | | | H. Up | | | | | H Up | |
| | | | | | | | | | H. Down<br>H. Up<br>H. Left<br>H Right<br>T. Left<br>T. Right | | |
| | | | | | | | | | | Touch | |
| | | | | | | | | Lift | | | |
| | | | | | | | | | | | |
| | Lift<br>Drop | | | | | | | | | | |
| | Put | | | | | | | | | Pick | Pick |

Fig. 4 The units on the training data for a 12 by 12 network with a training time of 50 epochs that had 10% or more samples for specific action verbs whose activation is the greatest.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | |
| | | | | | Forward<br>T. Left<br>T. Right<br>Go | | | | | | |
| | | | | | | | | | | | H. Right |
| | | | | | Go | | | | | | |
| | | | | | | T. Right | | | Backwards | | |
| | | | | | | Forward | | | | H Up | |
| | | | | | | | | | H. Down<br>H. Up<br>H. Left<br>H Right<br>T. Left<br>T. Right | | |
| | | | | | | | | | | Touch | |
| | | | | | | | | Lift | | | |
| | | | | | | | | | | | |
| | Drop | | | | | | | | | | |
| | Put<br>Lift | | | | | | | | | | Pick |

Fig. 5 The units on the test data for a 12 by 12 network with a training time of 50 epochs that had 10% or more samples for specific action verbs whose activation is the greatest.

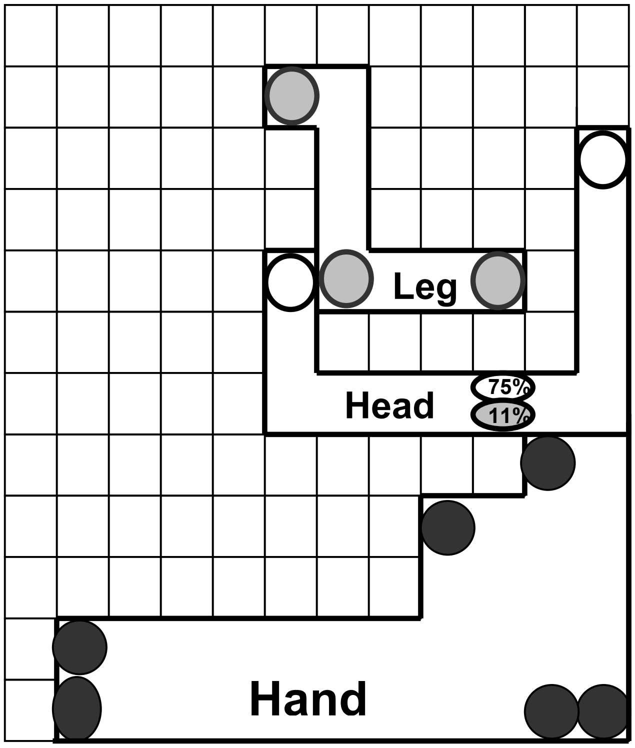

Fig. 6 The units on the training data for a 12 by 12 units network with a training time of 50 epochs that had 5% or more samples for specific body parts whose activation value is the greatest.

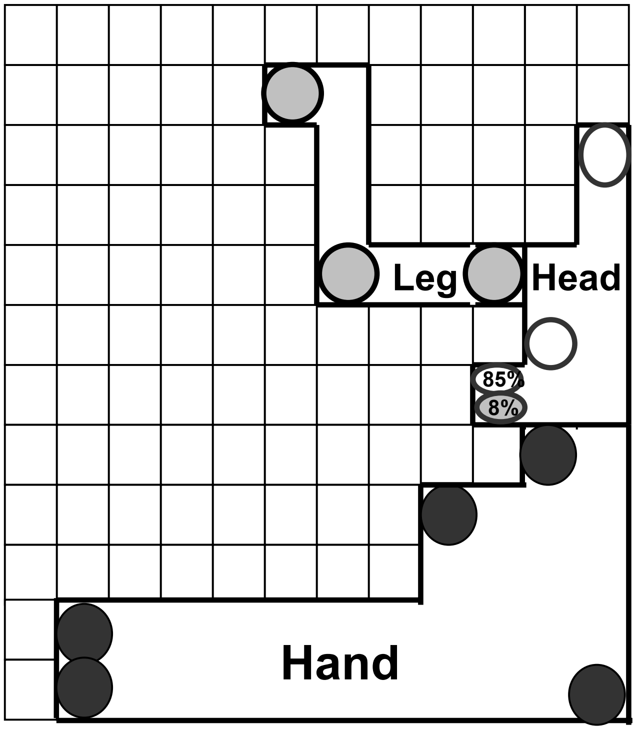

Fig. 7 The units on the test data for a 12 by 12 units network with a training time of 50 epochs that had 5% or more samples for specific body parts whose activation value is the greatest.

When considering self-organising networks with output layers of between 8 by 8 and 11 by 11 units the networks were only able to produce a split between hand action verbs and the other two classes. These networks were not able to cluster the leg and head actions in different regions of the output layer. However, it did indicate an ability to produce a split between simple actions such as 'forward' or 'head right', and more complex actions such as 'put' or 'pick'. It seems that these network output layers were too small to allow a clear split between the three body part classes.

However, for the network architecture that has more memory with 12 by 12 units in the output layer at a training time of 50 epochs there was clear clustering into the three body parts (see Fig. 4 to 7). The hand actions words such as 'pick', 'touch', 'lift' were at the bottom of the training and test output layers in the hand body part region, with the head action verbs like 'head up' and 'head down' slightly below and to the right of the leg region containing action verbs such as 'turn right' and 'go'. Although one unit within the head region contained both head and leg action verb samples with the highest activation, the percentage for head samples was much higher on both test and training data. For the training and test data the percentage of head verb samples with the highest activation for that unit was 75% and 85% respectively compared with 11% and 8%.

For the training data 100% of the hand and head fell in the appropriate region and 89% of the leg data. For test data the percentage was even better with 100% for hand and head and 91% for leg. It is interesting to note that within the hand verb region there was a good division into the actual action verb classes. From Figs 5 and 6 'pick' was located in the lower right of the region, 'put' in the lower left, 'drop' in the unit above 'pick', 'touch' at the top of the hand region and most of the 'lift' samples were located in a unit just below 'touch'.

Hence a network of this size can in principle realise the findings on Pulvermüller et al. on the processing of action verbs with different cell assemblies representing the specific body parts. The network was able to identify the semantic features from the actual sensor readings for the individual action verb classes that were specific to the appropriate body part. For such an architecture on both training and test data the clusters were in very similar position on the output layer, which points to the ability of the network to generalise on data it has not seen before. When considering the percentage of test data that fell in the regions identified by the training data the percentages were very good. For the hand action verbs 100%, leg for 95% and head for 88% of the test data fell into the appropriate training region. Therefore, if the self-organising network was used in the control of a robot it may perform successfully in an on-line manner clustering semantic features of the action to the appropriate region of the output layer.

## V1. FUTURE WORK

The experiment performed supports that self-organising networks cluster action verbs using semantic features that come from objective sensor readings with the appropriate body part as suggested by Pulvermüller et al. Self-organising networks seem suitable for incorporation into a robot control system that uses language that combines brain-inspired modularity, the neurocognitive evidence on action verb processing and mirror neuron theory. This is to be achieved by a hierarchical structure of self-organising networks that learns to associate the semantic features that represent the action verbs with a representation of the word form.

As can be seen from Fig. 8 the approach firstly uses a self-organising network to associate the action verbs with the appropriate body part by clustering the verbs in different regions of the output layer. In the next processing level there is a self-organising network for each of the body parts that uses the input sensor reading vectors to associate the actual action verbs with different regions. Also at this level the word forms that are represented using a random number approach based on the phonemes in the words are clustered in a self-organising network. In the upper-most level self-organising network the action verbs and their appropriate word form are associated by using the clustering patterns from the networks of the previous level. Hence the association means we can give the robot the action verb semantic representation and get the robot to associate this with the word form and so state what the action is or give it the word form and get it to perform the action.

Hence, if we wanted the robot to tell us the required action is 'put', the 'put' action verb sensor reading representation would be introduced into the trained body part network, which would locate it in the hand region of the output layer. The hand self-organising network would then position the input in the 'put' region of the output layer. As the robot is determining the word form there would be no input from the word form self-organising network into the action and word form association self-organising network. However, as the network has learned to associate this action with the appropriate word form the 'put' region of the network is activated. The robot will then state that the action semantic features provided are those for 'put'.

This approach offers some brain-inspired regional modularity by having multiple self-organising networks each performing a subtask of the overall task. These networks are linked in a distributed overall memory organization. The approach also takes into account the neurocognitive evidence of Pulvermüller et al. in that cell assemblies in different regions are associated with specific action verbs as a functional unit, with the association being based on the action verbs relationship with the appropriate body part. Furthermore, by using the sensor readings as input the mirror neuron concept is included through the understanding of the action by gaining the representation that could come from either performing the action or a stored representation linked to observation that creates the same activation pattern in units.

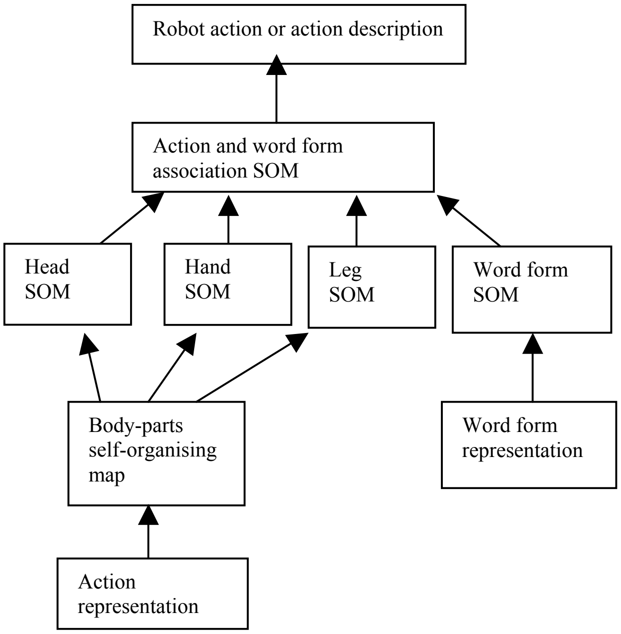
Robot action or action description
Action and word form association SOM
Head SOM
Hand SOM
Leg SOM
Word form SOM
Body-parts self-organising map
Word form representation
Action representation

Fig. 8 Hierarchical model for robot control system using language.

## VII. CONCLUSION

A model for a robot control system based on language instructions has been described that considers that cell assemblies in different regions of the brain are used to process action verbs based on their association with appropriate body parts. In doing so we expanded on previous work by including more objective sensor readings that incorporate semantic features to the process. Furthermore, this paper describes an hierarchical self-organising approach that controls a robot using language based on distributed regional modularity in the brain, mirror neuron theory and neurocognitive evidence on clustering action verbs.

## ACKNOWLEDGMENT

This work is part of the MirrorBot project supported by the EU in the FET-IST programme under grant IST- 2001-35282 coordinated by Prof. Wermter.

## REFERENCES

[1] Asoh, H., Huyamizu, S., Isao H., Motomura, Y. Akaho, S. and Matsu, T., Socially Embedded Learning of Office-Conversant Robot jjjo-2. *Proceedings of the International Joint Conference on Artificial Intelligence*, Nagoa, 1997.

[2] Bear, M., Connors, B. and Paradiso, M., *Neuroscience: Exploring the Brain*, 1996.

[3] Binder, J., Frost, J., Hammeke, T., Cox, R., Rao, S., and Prieto, T. Human Brain Language Areas Identified by Functional Magnetic Resonance Images. *The Journal of Neuroscience*, Vol. 17, No. 1, pp. 280-288, 1997.

[4] Breazeal, C. and Scassellati, B. A Context-Dependent Attention System for a Social Robot. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, Sweden, pp. 1146-1151, 1999.

[5] Burgard, W., Cremers, A.B., Fox, D., Hähnel, D., Lakemeyer, G., Schulz, D., Steiner, W. and Thrun. S. Experiences with an Interactive Museum Tour-Guide Robot. *Artificial Intelligence*, Vol. 114, No. 1-2, 2000.

[6] Elshaw, M. and Wermter, S. A Neurocognitive Approach to Self-organisation of Verb Actions. *Proceedings of the International Joint Conference on Neural Networks*. Honolulu, USA, pp. 24-29, 2002.

[7] Friedman, L., Kenny, J., Wise, A., Wu, D., Stuve, T., Miller, D., Jesberger J. and Lewin, J. Brain Activation During Silent Word Generation Evaluated with functional MRI. *Brain and Language*, Vol. 64, pp. 943-959, 1998.

[8] Gallese, V. and Goldman, A. Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Science*, Vol. 2, No. 12, pp. 493-501, 1998.

[9] Gazzaniga, M., Ivry, R. and Mangun, G. *Cognitive Neuroscience: The Biology of the Mind*, W.W. Norton & Company Ltd, 1998.

[10] Kohonen, T. *Self-Organizing Maps*. Springer Verlag, Heidelberg, Germany, 1997.

[11] McCelland, J. and Kawamoto, A. Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences. In McCelland, J. and Rumbelhart, D. editors, *Parallel Distributed Processing Vol. 2*, MIT Press, Cambridge, USA, pp. 272-331, 1986.

[12] Pulvermüller, F. A Brain Perspective on Language Mechanisms: From Discrete Neuronal Ensembles to Serial Order. *Progress in Neurobiology*, Vol. 67, pp. 85-111, 2002.

[13] Pulvermüller, F. Words in the Brain's Language. *Behavioral and Brain Sciences*, Vol. 22, No. 2, pp. 253-336, 1999.

[14] Pulvermüller, F., Hare, M. and Hummel, F. Neurophysiological Distinction of Verb Categories. *Cognitive Neuroscience*, Vol. 11, No. 12, pp. 2789-2793, 2000.

[15] Reggia, J., Shkuro, Y., and Shevtsova, N. Computational Investigation of Hemispheric Specialization and Interactions. In Wermter, S., Austin, J. and Willshaw, D. editors, Emergent *Neural Computational Architectures based on Neuroscience*, Springer-Verlag, Heidelberg, Germany, pp. 68-82, 2001.

[16] Rizzolatti, G. and Arbib, M. Language Within Our Grasp. *Trends in Neuroscience*, Vol. 21, No. 5, pp. 188-194, 1998

[17] Rizzolatti, G., Fogassi, L. and Gallese, V. Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action. Nature Review. Vol. 2, pp. 661-670, 2001.

[18] Steels. L. The Origins of Syntax in Visually Grounded Robotic Agents. Artificial Intelligence, Vol. 103, No. 1-2, pp. 133--156, 1998.

[19] Thrun, S., Bennewitz, M., Burgard, W. Dellaert, F. Fox, D., Haehnel, D., Rosenberg, C., Roy, N., Schulte, J. and Schulz, D. MINERVA: A Second Generation Mobile Tour-Guide Robot. International Conference on Robotics and Automation, 1999.

[20] Wermter, S., Austin, J., Willshaw, D., and Elshaw, M. Towards Novel Neuroscience-inspired Computing. *Emergent Neural Computational Architectures based on Neuroscience*, Springer-Verlag, Heidelberg, Germany pp. 1-19, 2001.