Autonomous Robots (2019) 43:1271–1290
https://doi.org/10.1007/s10514-018-9793-7

Sensorimotor input as a language generalisation tool: a neurorobotics model for generation and generalisation of noun-verb combinations with sensorimotor inputs

Junpei Zhong[1,2] · Martin Peniak[3] · Jun Tani[4] · Tetsuya Ogata[5] · Angelo Cangelosi[2]

Received: 9 May 2016 / Accepted: 27 July 2018 / Published online: 21 August 2018

Abstract
The paper presents a neurorobotics cognitive model explaining the understanding and generalisation of nouns and verbs combinations when a vocal command consisting of a verb-noun sentence is provided to a humanoid robot. The dataset used for training was obtained from object manipulation tasks with a humanoid robot platform; it includes 9 motor actions and 9 objects placing placed in 6 different locations), which enables the robot to learn to handle real-world objects and actions. Based on the multiple time-scale recurrent neural networks, this study demonstrates its generalisation capability using a large data-set, with which the robot was able to generalise semantic representation of novel combinations of noun-verb sentences, and therefore produce the corresponding motor behaviours. This generalisation process is done via the grounding process: different objects are being interacted, and associated, with different motor behaviours, following a learning approach inspired by developmental language acquisition in infants. Further analyses of the learned network dynamics and representations also demonstrate how the generalisation is possible via the exploitation of this functional hierarchical recurrent network.

Keywords Recurrent artificial neural networks · Language learning · Multiple time-scale recurrent neural network · Developmental robotics · Neurorobotics

1 Introduction

For the design of social robots (Breazeal 2004; Dautenhahn 2007), besides of building robots with human-like external morphology, the ability to process, to understand and generate language is one of the key factors to support human-robot interaction. However, to build a model to accomplish similar processes for social robotics, the design of the robot's abilities of understanding, generation and generalisation of natural language is still an open challenge. Particularly, natural language understanding for a social robotic system plays an essential role as it interfaces the vocal command from human users to an internal representation in the robot's own cognitive system. In this study, we will apply a developmental robotics approach to the design of language and communication abilities in robots, following an incremental and interactive process to language learning, inspired by language development in infants.

1.1 Language understanding for robot systems

Important recent developments in social robotics, such as robots performing human-like emotion expression (Zhong and Canamero 2014) and social attention for autonomous movement (Novianto 2014), have been accompanied by language understanding approaches focusing on the grounding of natural language into the agent's sensorimotor experience and its situated interaction (Cangelosi 2010a; Steels and Hild 2012).

✉ Junpei Zhong
zhong@junpei.eu

1 Artificial Intelligent Research Center, National Institute of Advanced Industrial Science and Technology, Aomi 2-3-26, Tokyo 135-0064, Japan

2 Centre for Robotics and Neural Systems, University of Plymouth, Plymouth PL4 8AA, UK

3 Cortexica Vision Systems, London, UK

4 Okinawa Institute of Science and Technology, Okinawa, Japan

5 Department of Intermedia Art and Science, Waseda University, Tokyo, Japan

For instance, in Tellex et al. (2011), Matuszek et al. (2013), syntactic parsing techniques are used to ground the language into primitive motor actions (e.g., pickup, move, place), which can be inferred within graph models. Similarly, Misra et al. (2014) developed a system for mobile robots which is able to learn to ground the language instructions from a corpus of pairs of natural language including both verbs and spatial information. In Yürüten et al. (2013), it was proposed that in order to understand the object affordance which can be described by adjectives, the most crucial property is the shape-related one.

Besides the direct modelling methods for robot language learning, an alternative approach to build a learning model for language is based on developmental robotics (Weng 2001; Asada 2009; Cangelosi and Schlesinger 2015). Taking inspiration from developmental psychology and developmental neuroscience studies, this approach emphasises the role of the environment and of the interactions that occur during learning, over a progression of learning stages. In the context of language understanding, the core of developmental robotics approaches to language learning is following a similar developmental pathway of infants acquiring grounded representations of natural language and forming a symbol system through embodied interaction with the physical environment (Cangelosi 2010b). Furthermore, via language learning an agent should also be able to generalise by inferring un-trained combinations of words within the lexical constructions acquired. One possibility to accomplish generalisation is to make good use of the semantic compositionality.

Various developmental robotics models have been developed that incrementally model the various stages of language acquisition in infants, from phoneme acquisition, to object and action names, to word combinations. For example, the cognitive model presented in Guenther (2006) outlines the cortical interactions in the syllable generation process which result in different developmental phenomena. This mimics the first stage of language development. The Elija model (Howard and Messum 2011) is a vocal apparatus which strictly follows detailed developmental stages. Working as an articulatory synthesizer, it firstly learns the production of sounds on its own. Then a caregiver is used to produce speech by using speech sounds for object names using reinforcement learning, where the reward is again given by the response of the caregiver. Likewise, a self-organizing map together with reinforcement learning was proposed in Warlaumont (2013), which demonstrated that the reinforcement learning based on the similarity of vocalization can improve the post-learning production of the sound of one's language.

From the models mentioned above, we can see that most of the methods for modelling the first stages of phonetics production do not tend to use robotic platforms. On the other hand, for the modelling of the later stages of lexical development, after assuming that phonetics skills are mastered, robotic systems are usually employed to establish the meta-knowledge about the association between vocal speech and the referents or the actions. Therefore, except studies focusing on the mental imagination of actions as in Golosio (2015), the mechanical morphology of a robot is particularly important when modelling the acquisition of words, especially those used to name the motor actions. For instance, the model from Mangin and Oudeyer (2012) gets as input dance-like combinations of human movement primitives plus ambiguous labels associated with these movements. Concentrating on the second and third stages of the associating lexicon, words and motor actions, the robot in Dominey et al. (2009) is able to acquire new motor behaviours in an on-line fashion by grounding the vocal commands on the pre-defined control motor primitives. Similarly, Siskind (2001) proposed a model which uses visual primitives to encode notions of different actions to ground the semantics of events for verb learning. Using structured connectionist models (SCMs), (Chang et al. 2005) built a layered connectionist model to connect embodied representations and simulative inference for verbs. In Cangelosi and Parisi (2004), the emergence of verb-noun separation is learned while the agents are interacting and manipulating the objects. Meanwhile, the tasks during of such interaction may be essential during learning too (Goodman and Frank 2016). Recent experiments (Rohlfing 2016; Andreas and Klein 2016) and also proposed that language learning should be posited in the context of task-directed behaviours.

In terms of the learning structure, Stramandinoli et al. (2012) developed a model about the grounding hierarchy of the verbs with more complex meanings (such as "keep", "reject", "accept" and "give") which related to the internal states of the caregivers and which were used to build a robotic model for the grounding of increasingly abstract motor concepts and words. As follow-up studies of Dominey et al. (2009), Dominey (2013), Hinaut and Dominey (2013) focused on the understanding of grammatical complexity. They used recurrent neural networks (RNN) to learn grammatical structure based on temporal series learning in artificial neural networks.

Also using RNN, Sugita and Tani (2005) reported experiments with a mobile robot implementing a two-level RNN architecture called Recurrent Neural Network with Parametric Bias Units (RNNPB). This allows the robot to map a linguistic command containing verbs and nouns into context-dependent behaviours corresponding to the verb and noun descriptions respectively. It was among the first to develop a robotic model of semantic compositionality based on the sensorimotor combinatory. With a cognitive robot experiment, the recurrent network models the emergence of compositional meanings and lexicons with no a priori

knowledge of any lexical or formal syntactic representation.

Comparing to RNNPB, another kind of RNN architecture called Multiple Timescale Neural Network (MTRNN) is able to ground different scales of sensorimotor information into the hierarchical structure of sentences, such as the spelling of words (Ogata and Okuno 2013) and words and sentences (Hinoshita 2011). The kind of recurrent models provides a memory to store the spatial and temporal structure of the environment and the lexical structures. Given the fact that RNN can learn the arbitrary length of the dependencies in statistical structures and their context, the storage ability of the RNN out-performs most of the language learning models.

1.2 Embodied symbolic emergence in a hierarchical structure

In the developmental psychology which studies focusing specifically on the emergence of nouns and verbs, there is still an open debate between the learning stages and their relative temporal acquisition order. For the early stages of the verb and noun learning, it is widely accepted that most of the common nouns are generally learned before verbs (Gentner 1982), by first connecting speech sounds (labels, nouns) to physical objects in view. However, some nouns which relate to context, such as "passenger", are learnt at a relatively later stage, only after "an extensive range of situations" (contexts or life phases) have been encountered (Hall and Waxman 1993), during which verbs may play a crucial role. The embodied learning of verbs and nouns is not correlated to one single modality in sensory percept's: experiments done in Kersten (1998) suggest that the nouns are grounded from the intrinsic properties of an object, even at different movements and orientations, while verbs are accounted for the movement path of an object. This distinction may be associated with the neuroanatomy distinction between the ventral and dorsal (what/where) visual streams, involved in the generation of nouns and verbs respectively. As Maguire et al. (2006) suggested, some nouns and verbs can be learnt more straightforward to learn because they can be accessed perceptually. On the other hand, some abstract words, either verbs or nouns, should only be learnt from a social and linguistic context.

For instance, while infants learn the word-gesture combination at the age of two, they associate the meaning of verbs with the meanings of the higher-order nouns (Bates and Dick 2002). Such verbs with complex meaning are obtained from both motor action and visual percept (Longobardi et al. 2015). As summarised in Cangelosi and Parisi (2001) and Cangelosi and Parisi (2004), comparing to the static object perception that associates with simple nouns, the early verb learning involves a temporal dynamic from motion perception. Indeed, we assert that the learning processes of nouns and verbs (especially for those with complex meanings) are not separated; there is a close relation between verb and noun development, during which the embodied sensorimotor information plays a crucial role.

During this embodied development, both the perceptual system and the motor system contribute to language comprehension (e.g. Pulvermüller 2002; Kaschak 2005; Pecher et al. 2003; Saygin 2010). This embodied development may contribute to the emergence of how compositional semantics of a sentence can be acquired by a language acquisition system without knowing any explicit representations about either the meaning of word or motor behaviours as a priori. In this way the system can refer a semantic compositionality by the sensorimotor combinatoriality. It also extends Piaget's proposal that language learning is a symbolised understanding process for dynamic actions, which is "a situated process, function of the content, the context, the activity and the goal of the learner" (Holzer 1994).

The sensorimotor information is not the only mechanism acting as a learning tool for language acquisition. Conversely, recent research also proposes that language is such a flexible and efficient system for symbolic manipulation which is more than a communication tool of our thoughts (e.g. Landy et al. 2014; Mirolli and Parisi 2009, 2011.) For the predictive effect from language to sensorimotor behaviours, vocal communication can be one of the sources that drive the visual attention to become predictive, by making inferences as to the source-inferences (Tomasello and Farrar 1986). In this process, language can trigger a predictive inference about the appearance of a visual percept, driving a predictive saccade (Eberhard 1995). Therefore, the sensorimotor system is affected by the inferences from the auditory modality or even from higher level cognitive processes.

We concluded this bidirectional relationship between language learning and sensorimotor system in a hierarchical cognitive framework proposed in Zhong (2015), in which the language understanding and grounding occurs during the dynamical process hierarchically from the neural processes on the (lower) receptor level to the higher level understanding which happens in the (higher) prefrontal cortex. As the review done by Tenenbaum (2011), the hierarchical framework can be detailed formulated in a probabilistic way, in which the abstract knowledge also acts as a prior to guide our learning and reasoning. The probabilistic based models have also been applied in acquiring abstract knowledge from robot-environment interaction (Konidaris et al. 2015), human-robot interaction (Iwahashi 2008) and multimodal living environment (Attamimi 2016). Additionally, the hierarchical architecture can also be implemented as connectionist models. For example, the hierarchical recurrent neural architectures can be found in Zhong et al. (2011), Zhong et al. (2012a), Zhong et al. (2012b), due to the fact that the learn-

ing modalities of visual perception and motor actions can be represented as both spatial and temporal sequences, so that the recurrent connections provide possibilities to intertwine these two modalities.

In this paper, due to our interests in the non-linear dynamics of the system and its contribution to the generalisation abilities, the recurrent neural models would be a proper model to model this process. Although similar RNNPB (Sugita and Tani 2005) or MTRNN (Heinrich et al. 2015) networks have been used to learn verbs and nouns features with motor actions and visual features, the model we will use is a *single* MTRNN model to learn both the sensory and motor information in a single set of sequences, because we regard the perception and action having inseparable links (e.g. Wolpert et al. 1995; Noë 2001) and should be encoded solely as similar data structures. Moreover, since the training of such a large MTRNN has become more and more feasible in recent years due to the accessibility and affordability of GPU computing, a large data-set from robotic experiment will be tried to be conceptualised towards abstract representations on the higher level of this hierarchy, similar to the developmental processes of language conceptualisation and categorisation.

To summarise, compared with the connectionist models on semantic compositionality (Sugita and Tani 2005; Heinrich et al. 2015), the novelties of our model and experiments are:

- Instead of using the neural binding methods on multiple RNNs, the hierarchical MTRNN provides another perspective to model the emergence of semantic compositionality over multi-modal data, which may be more parallel to the perception-action coupling of different levels of the nervous system (Sperry 1952): perception and action processes are functionally intertwined, which we represent in the recurrent connections from the low to the top layer in our hierarchical network.
- Technically, in our model, the multi-modal data (language, visual and proprioceptive) was implemented into a single hierarchical network. This uniformity can be discovered in the higher-level heteromodal representation in the multisensory neurons with continuous feedback and feed-forward connectivities (Ghazanfar and Schroeder 2006; Macaluso and Driver 2005). That is similar to the recurrent neural architecture we use. Furthermore, a single RNN network that incorporates multi-modal signals would be beneficial to improve the generalisation ability.
- Using a humanoid robot and a large-scale dataset, we can observe how the semantic dynamic is emerged on different levels with a similar learning process of the human morphology. The later experiments will also show how the semantic structures of verbs are self-organised on the higher-level of neurons, suggesting a similar neural representation may exist in the human brain activities.

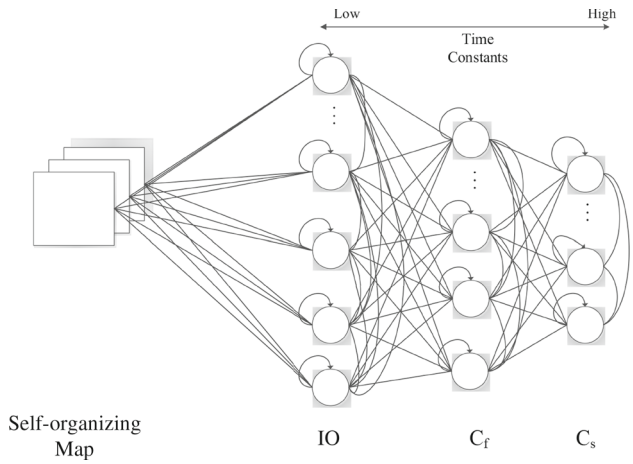


Fig. 1 Architecture of multiple time-scale recurrent neural network

2 The multiple timescale recurrent neural network model

Briefly, the motivations that we employ recurrent neural models, specifically, the MTRNN, to model the learning processes of the language learning from the sensorimotor interaction are:

- The hierarchical neuron distribution in a single MTRNN with multi-modal inputs is able to mimic the dynamical and bidirectional processes of the heteromodal neurons when human is learning the multi-sensory knowledge;
- Furthermore, such dynamical process in the RNNs is able to form the bifurcation functions in which the functional hierarchy is formed in a self-organized way in one network (Tani 2014);
- The MTRNN is able to be stacked in a hierarchical way which is also similar to the hierarchical organization of the brain areas (Zhong 2015);

Our language learning model is based on the combination of an MTRNN network with Self-Organizing Maps (SOMs) to control the humanoid robot iCub, being trained on the understanding of a set of noun-verb combinations to perform a variety of actions with different objects. Figure 1 shows the learning architecture incorporating a Multiple Timescale Recurrent Neural Network (MTRNN) (Yamashita and Tani 2008) and the self-organizing maps. The core module of the system is the MTRNN, which will learn sequences of verb-noun instructions and will control the movement of the robot in response to such instructions. The inputs to the MTRNN correspond to the language command inputs, to the visual inputs as well as the proprioceptive inputs. We regard these three modalities as a whole sensorimotor input because the MTRNN model is able to learn the relation between the verbs and nouns and seen objects within the context of the

non-linearity of the sensorimotor sequences in a hierarchical manner. This network will learn this non-linearity in the functional hierarchy in which the neural activities are self-organised, exploiting the spatiotemporal variations.

2.1 Using a self-organizing map as a sparse structure

The initial input data sets, consisting of speech, camera images, and proprioceptive (kinesthetic) states are pre-processed (see Eqs. 1–4) using three SOMs respectively for the linguistic, visual and motor input modalities.

Although the MTRNN could be trained with original data representation, we usually employ pre-processing modules for the MTRNN inputs, which result in a sparse structure of the weighting matrices in the network. Also the MTRNN outputs are decoded into the original data structures. The sparseness in weighting matrices has a similar concept of sparse coding in computational neuroscience (Olshausen and Field 1997): the weighting matrices are sparsely distributed, which is an analogous form of the sparse distributed representations that are used in our neural activities, such as in visual (Essen 1985) and auditory cortex (Reale and Imig 1980). Previous research on language learning in RNN (Awano et al. 2011) also showed that a sparse encoding results in robustness in training and a better generalisation results and improved robustness with noisy inputs.

Here the sparseness structure in the weight matrices is given by the SOMs (Kohonen 1998). During this process, the SOM performs as a dimensional mapping function, with an output space with higher dimensions than the input space. Having a discretised and distributed neural encoding in the output space, the pre-processed SOM modules are able to reduce the possible overlap of the original data within the original input space. Therefore, the topological homomorphism produced by the SOM guarantees that the training vectors between the raw training-sets and the input vectors are topologically similar with each other.

In the SOM training here, assuming the input vectors are

$$x = [x^1, x^2, \ldots, x^m]^\intercal \tag{1}$$

where m is the number of dimensions of the input vectors. These input vectors are mapped to an output space whose coordinates define the output topology of the SOM. Connecting between the input and output spaces, the weight vector is defined as

$$w_j = \left[w_j^1, w_j^2, \cdots, w_j^m\right]^\intercal, j = 1, 2, 3, \cdots, n \tag{2}$$

where neuron j is one of the input space vectors and n is the total number of those neurons. When a self-organising map receives an input vector, the algorithm finds a neuron associated with weights that are most similar to the input vector. The measure of similarity is usually done using the Euclidean distance metric, which is mathematically equivalent to finding a neuron with the largest inner product $w_j^\intercal x$. Thus the very neuron that is the most similar match for the input vector is referred to as the best matching unit (BMU) and it is defined as:

$$c = \arg min_j \|x - w_j\| \tag{3}$$

The dimensionality mapping is achieved when the BMU coordinates are used to update the weights of the neighbourhood neurons around neuron c by driving them closer to the input vector at iteration t:

$$w_j(t+1) = w_j(t) + \delta(x_j - w_j) \tag{4}$$

δ is a Gaussian neighbourhood function, which determines the adjusting rate for the weights.

Therefore, the output of the SOM which is encoded in a high-dimensional input space, is still able to preserve the topological properties of the input space due to the use of the neighbourhood function.

2.2 Multiple timescale recurrent network (MTRNN)

As shown in Fig. 2, the neurons in the MTRNN form three layers: an input-output layer (IO) and two context layers called Context fast (C_f) and Context slow (C_s). In the following text, we denote the indices of these neurons as:

$$I_{all} = I_{IO} \cup I_{C_f} \cup I_{C_s} \tag{5}$$

where I_{IO} represents the indices to the neurons at the input-output layer, I_{C_f} belongs to the neurons at the context fast layer and I_{C_s} belongs to the neurons at the context slow layer. The neurons on a layer own full connectivity to all neurons within the same and adjacent layers, as shown in Fig. 1. The difference between the fast and slow context layers as well as the input-output layer consists in having distinct time constants τ, which determine the speed of the adaptation given a time sequence with a specific length, when updating the neural activity. The larger the value of τ, the slower the neuron adaptation. The difference of adaptation rate of the neurons further assemble features of the input sequences in various timescales. Therefore, given the previous states $S(0), S(1), \ldots, S(t)$, their spatiotemporal features will be self-organised on different levels of the network. So the MTRNN is not only a continuous time recurrent neural network that can predict the next states $S(t+1)$ of the time sequence, but also its internal state acts as a hierarchical memory to preserve the temporal features of the non-linear dynamics in different timescales. In the embodied learning case, such memories, mostly in a set of of oscillatory patterns,

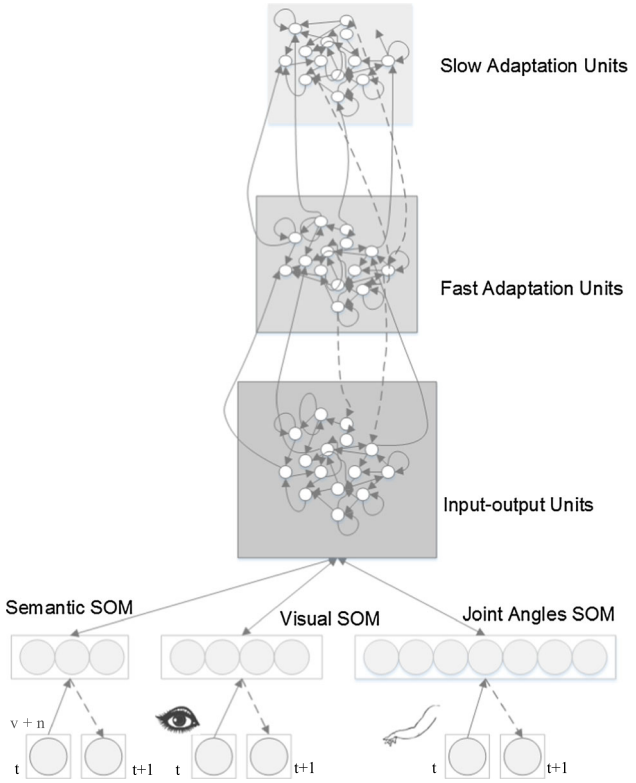


Fig. 2 Language learning model based on MTRNN

represent the verb/noun semantics during the robot interaction. Therefore, such patterns are learnt by self-organising as fixed points and limit cycle non-linear dynamics.

2.2.1 Learning

In general, the training of the MTRNN follows the updating rule of classical firing rate models, in which the activity of a neuron is determined by the average firing rate of all the connected neurons. Additionally, the neuronal activity is also decaying over time following an updating rule of the leaky integrator model. Therefore, when time-step $t > 0$, the current membrane potential status of a neuron is determined both by the previous activation as well as the current synaptic inputs, as shown in Eq. 6:

$$\tau_i u'_{i,t} = -u_{i,t} + \sum_j w_{i,j} x_{j,t} \tag{6}$$

where $u_{i,t}$ is the membrane potential, $x_{j,t}$ is the activity of j-th neuron at t-th time-step, $w_{i,j}$ represents the synaptic weight from the j-th neuron to the i-th neuron and τ is the time scale parameter which determines the decay rate of this neuron. One of the features that is similar to the generic continuous time recurrent neural networks (CTRNN) model is that a parameter τ is used to determine the decay rate of the neural activity; a larger τ means their activities change slowly over time compared with those with a smaller τ.

Assuming the i-th neuron has the number of N connections (i.e. the total number of the neurons in the network is N), Eq. 6 can be transformed into

$$u_{i,t+1} = \left(1 - \frac{1}{\tau_i}\right) u_{i,t} + \frac{1}{\tau_i} \left[\sum_{j \in N} w_{i,j} x_{j,t} \right] \quad (\text{if } t > 0) \tag{7}$$

When the time-step $t = 0$, the membrane potential of the IO neurons is set to 0 and the context neurons are set to initial states $C_{s_c}(i, 0)$:

$$u_{i,0} = \begin{cases} 0, & \text{if } \ t = 0 \text{ and } i \in I_{IO}, \\ C_{s_c(i,0)}, & \text{if } \ t = 0 \text{ and } i \notin I_{IO} \end{cases} \tag{8}$$

The neural activity of a neuron is calculated in two methods (the sigmoid function and the soft-max function), depending on which level the neuron belongs with:

$$y_{i,t} = \begin{cases} \dfrac{e^{u_{i,t}}}{\sum_{j \in Z} e^{u_{j,t}}}, & \text{if } i \in I_{IO}, \\ \dfrac{1}{1 + e^{-u_{i,t}}}, & \text{otherwise.} \end{cases} \tag{9}$$

Particularly, the soft-max activation function gives rise to the recovery of a similar probability distribution as the SOM pre-processing modules. Therefore, this activation function results in a faster convergence to the MTRNN network training.

During the training process, it is to minimize the error E defined by the Kullback-Leibler divergence:

$$E = \sum_t \sum_{i \in O} y^*_{i,t} log \left(\frac{y^*_{i,t}}{y_{i,t}} \right) \tag{10}$$

where $y^*_{i,t}$ is the desired neural activation of the i-th neuron at the t-th time-step, which acts as the target value for the actual output $y_{i,t}$. The target of the training is to minimize E by back-propagation through time (BPTT).

In the BPTT algorithm, the input of the IO neuron is calculated from a mixed partition value r (called the feedback rate) of the previous output value y and the desired value y^*. (Eq. 11)

$$x_{j,t+1} = (1 - r) \times y_{j,t} + r \times y^*_{j,t} \tag{11}$$

where we will use $r = 0.1$ during training, and $r = 0$ during generation, which means that the network is used to generate the sequences autonomously.

At the n-th iteration of training, the synaptic weights and the biases of the network of neuron i are updated according to Eq. 12.

$$w_{i,j}^{n+1} = w_{i,j}^{n} - \eta_{i,j} \frac{\partial E}{\partial w_{i,j}} = w_{i,j} - \frac{\eta_{i,j}}{\tau_i} \sum_t x_{j,t} \frac{\partial E}{\partial w_{i,t}} \tag{12}$$

$$b_i^{n+1} = b_i^n - \beta_i \frac{\partial E}{\partial b_i} = b_i - \beta_i \sum_t \frac{\partial E}{\partial u_{i,t}} \tag{13}$$

$$\frac{\partial E}{\partial u_{i,t}} = \begin{cases} y_{i,t+1} - y^*_{i,t+1} + \left(1 - \frac{1}{\tau_i}\right) \frac{\partial E}{\partial u_{i,t+1}}, \\ \qquad \text{if} \quad i \in I_{IO}, \\ \sum_{k \in I_{all}} \frac{\partial E}{\partial u_{k,t+1}} \left[\lambda_{i,k} \left(1 - \frac{1}{\tau_i}\right) + \frac{1}{\tau_k} w_{ki} f'(u_{i,t})\right], \\ \qquad \text{otherwise.} \end{cases} \tag{14}$$

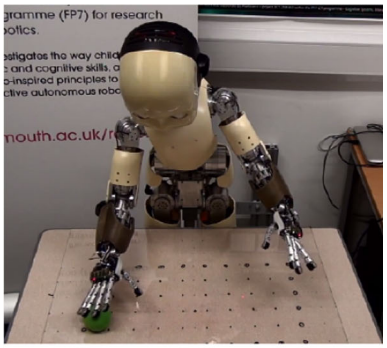
In Eqs. 12 and 13, the partial derivatives for w and b are the sums of weight and bias which determine the changes over the whole sequence respectively, and η and β denote the learning rates for the weight and bias changes. Particularly, the term $\partial E / \partial u_{k,t}$ can be calculated recursively as Eq. 14, where the $f'()$ is the derivative of the sigmoid Function defined by Eqs. 8 and 9. The term $\lambda_{i,k}$ is the Kronecker's Delta, whose output is 1 when $i = k$, otherwise, it is set to 0.

3 Experiments

To examine the network performance, we recorded the real world training data from object manipulation experiments based on an iCub robot (Metta et al. 2008). This is a child sized humanoid robot built as a testing platform for theories and models of cognitive science and neuroscience. Mimicking a two-year old infant, this unique robotic platform has 53 degrees of freedom. As such, using the iCub, we set a learning scenario in which a human instructor was teaching the robotic learner a set of language commands whilst providing kinaesthetic demonstration of the named actions. This setting is similar as the infant-directed action or motionese scenario (e.g. Brand et al. 2002; Brand 2007) where the mother modifies their actions when demonstrating objects to infants in order to assist infants' processing of human action. Duplicating the learning environment of the development process, the aim of these experiments was to evaluate the verb-noun generalisation with a large data-set using the MTRNN. We were also interested in how the mechanisms, especially the neural activities in the hierarchical architecture, result in such a generalisation.

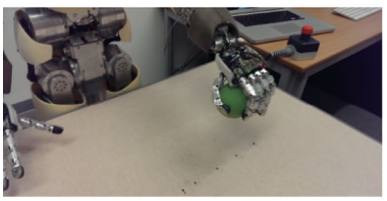
3.1 Experimental setup

Figure 3a shows the setup used in our experiments. During the training process, the data set was obtained using the following steps:



(a)

(b)



(c)

Fig. 3 Experimental scenario. **a** iCub Manipulation setting. **b** Objects used in the experiment. There are eight different objects shown in this image. The last object that is not present is a green ball, which is shown in Fig. 3c. **c** Example of a complex lifting action involving the coordination of the entire upper body actuated by 41 motors

1. Objects with significantly different colours and shapes were placed at 6 different locations along the same line in front of the iCub (i.e. the **objects** from perception).
2. A vocal command was spoken by an instructor according to the visual scene that was perceived by the iCub. A complete sentence of the vocal command is composed of a verb and a noun such as "lift [the] ball". This was recognised by the speech recognition software called Dragon dictate,[1] with which the corresponding verb and noun were recognised and then translated into two dedicated discrete values based on the verb and noun look-up table (Table 1) (i.e. a sentence includes **a verb and a noun**).

[1] The speech recognition is not always successful here. It is not the main research topic in this work. But for the sake of a more natural training process, we manually monitored to obtain current recognition results before the data sequences were recorded to ensure a better performance. Please also see http://www.nuance.co.uk/dragon/index.htm.

Table 1 Look-up table of verbs and nouns for the data sets: the instructor showed the robot with different combinations of the 9 actions and 9 objects

Actions	Slide left	Slide right	Touch	Reach	Push
Verb value	0.0	0.1	0.2	0.3	0.4
Actions	Pull	Point	Grasp	Lift	
Verb values	0.5	0.6	0.7	0.8	
Objects	Tractor	Hammer	Ball	Bus	Modi
Noun value	0.0	0.1	0.2	0.3	0.4
Objects	Car	Cup	Cubes	Spiky	
Noun values	0.5	0.6	0.7	0.8	

The actions and the objects are represented in two discretised values for semantic command inputs which range from 0 to 0.9. For instance, the command "lift [the] ball" is translated into values [0.8, 0.2]

3. Following the command "lift [the] ball", the built-in vision tracker of the iCub searches for a ball-shaped object and automatically locate it in the middle of the receptive field; in this way, the joint angles of head and neck measure the position of the object (for the purpose of generalisation of different locations).
4. Joint positions of the head and neck are recorded. The sequence recorder module of the iCub was used to record the sensorimotor trajectories while the instructor was guiding the robot by holding its arms to perform a certain action for each object (i.e. the **motor actions**).

During the testing process, all the objects are placed on the table. The vocal command from the instructor are acted before the action execution. The whole experimental setup used combinations of 9 actions and 9 objects. The objects and one example of the action can be found in Fig. 3b and 3c. From these combinations, both the vocal commands (i.e. a complete sentence includes verb and noun) and the sensorimotor sequences can be created. To the best of our knowledge, this 9×9 noun-verb scenario is one of the setups with the highest combination of verbs and nouns in grounded robot language experiments (e.g. Tani et al. 2004; Yamashita and Tani 2008). We used such a large number of data to test the combinatorial complexity and mechanical feasibility of this model, as well as to evaluate the generalisation ability and its internal non-linear dynamics when using such a large data-set. From an engineering point of view, after testing the feasibility of generalisation, it is also possible to apply this model in a real-world robot application.

As mentioned before, each speech command was recognised and translated into two semantic command units. Using 9 discretised values for verbs and 9 for nouns, the semantic commands have thus 81 possible combinations. This translation was done according to the verb and noun look-up table, as shown in Table 1. Since we used the visual object tracker in the iCub, the joints of neck and eyes automatically represent the location of the particular object which is presented in the vocal commands. Also the movements of the joint angles in the torso are recorded as the sequences of the motor actions. During the data recording, each recording sequence lasted 5 seconds and the encoder values of 41 joints were sampled at $50ms$ intervals. Thus, the complete input vector of the data set contains 100 steps of the discrete semantic command, location of visual attention and joint movement of the torso, as shown in Table 2.

Three experiments were carried out and are described in the next subsections: in the first experiment, given the 9 actions and 9 objects data set, we will search the parameter space and find the best parameters for the network training. In the second experiment, the training and generalisation performance will be shown given different types of manipulated data sets. For the third experiment, we will further analyse the generalisation ability of the MTRNN network. All these experiments were run using a modified version of the Aquila software (Peniak et al. 2011) in a GPU computer with one Tesla C2050 and two GeForce GTX 580 graphic cards.

3.2 Training performance

In this experiment, we used the data set consisting of the complete 9×9 combinations (i.e. number of verbs: $N_v = 9$, number of nouns: $N_n = 9$), which include information about 6 different object locations. The 6 locations were placed along the straight line on the table as shown in Fig. 3a. Thus the whole data-set contains $9 \times 9 \times 6 = 486$ sequences (teaching time took less than 1 hour totally), which were all used for training the network.

After a brief hyper-parameter search experiment shown in Table 3, we selected the best parameters for this data-set are (70, 3, 50, 120) in the parameter space ($\tau_s, \tau_f, N_{C_s}, N_{C_f}$). We then examined the training performance of the network under this parameter setting using different data-sets. To test the generalisation ability, these data-sets were manipulated: a subset of the combinations of actions and objects were removed from the training set, to be used as validation test sets when testing the generalisation ability of the network. The detailed information about the manipulated data-sets are shown in Table 4, where the coloured numbers N indicate the specific verb-noun combination removed in the specific N-th data-set. We can see that the number of removal sets was increasing from the first to the third test-set, indicating the difficulty of generalisation was increasing. Also at the second and the third data-sets, some of the removal sets were next to each other, which further increased the difficulty of generalisation.

Table 2 Structure of the training data

Description	Semantic commands	Object location (neck and eyes)	Torso joints
Dimension	2	6	3
Description	Left arm joints	Right arm joints	
Dimension	16	16	

Table 3 Training error with different parameter settings (C_s, C_f, N_{C_s}, N_{C_f})

Parameters	Error 1	Error 2	Error 3	Ave.
(70, 5, 20, 60)	0.084	0.081	0.085	0.0833
(70, 3, 20, 60)	0.084	0.085	0.082	0.0837
(70, 5, 30, 60)	0.084	0.086	0.083	0.0843
(70, 5, 30, 50)	0.082	0.079	0.080	0.0803
(70, 5, 30, 100)	0.079	0.079	0.078	0.0787
(70, 5, 60, 100)	0.078	0.078	0.077	0.0777
(70, 5, 40, 120)	0.079	0.079	0.078	0.0787
(70, 5, 50, 140)	0.075	0.075	0.077	0.0757
(70, 5, 60, 160)	0.072	0.071	0.074	0.0723
(70, 5, 50, 120)	0.071	0.070	0.071	0.0707
(70, 3, 50, 120)	0.071	0.071	0.070	0.0707
(70, 5, 70, 120)	0.070	0.071	0.072	0.0710

Table 4 Some of the sequences containing particular semantic combinations of verbs and nouns were removed during training

N. / V.	Tractor	Hammer	Ball	Bus	Modi	Car	Cup	Cube	Spiky
Slide left	1/2/3			3		2	3		
Slide Right	2/3	1		3		2	3		
Touch	3	2	1	3			2/3		
Reach		2/3		1	3		2	3	
Push		3	2		1/3			2/3	
Pull		3	2		3	1		2/3	
Point			3	2		3	1		2/3
Grasp			3	2		3		1	2/3
Lift			3		2	3			1/2/3

The number i in the cell indicates that such a combination was removed in the i-th training set for generalisation experiments

We used the parameter set of (50, 5, 70, 100). To further demonstrate the robustness of the generalisation ability given the un-trained sensorimotor sequences, the validation sets, which were not included in the training, were fed into the network. In this way, we aimed to test how the network responds to noun-verb combinations not used during training. Using the three MTRNNs we trained from three data-sets, we performed three generalisation experiments using the missing verb-noun combinations. In the experiments, only the first time step data in the sequence was provided (i.e. $r = 0$ in Eq. 11), which includes the initial position of the torso, head, and eye motors, as well as the vocal command. Then the network prediction was used as the input of the next time-step and formed a closed-loop to complete 100-step of the time sequence generation. The errors of the whole three training-sets, as well as those in different steps are shown in Table 5.

Table 5 RMS error of the generalisation tests

Test	1	2	3
RMS error (All)	0.0052	0.0069	0.0169
RMS error (Step 1-20)	0.0064	0.0082	0.0240
RMS error (Step 21-40)	0.0042	0.0075	0.0194
RMS error (Step 41-60)	0.0033	0.0069	0.0150
RMS error (Step 61-80)	0.0031	0.0062	0.0121
RMS error (Step 81-100)	0.0024	0.0052	0.0101

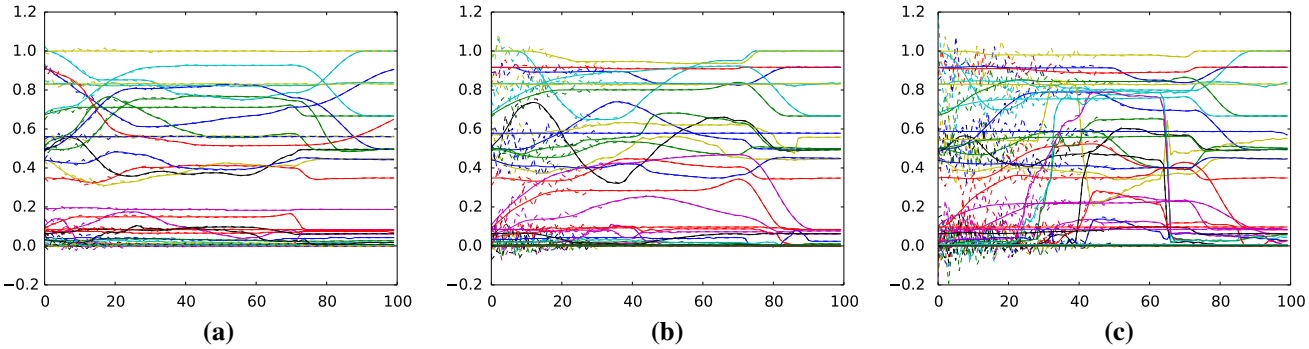


Fig. 4 Trajectory generation The generated trajectories (dotted) with 41 dimensions were plotted and compared with the original trajectories. Three test-sets were selected to validate the training performances with different training sets. Similar to our RMS error shown in Tab. 5, larger errors could be found at the beginning of the sequences. **a** Generated trajectory from MTRNN 1, Test-set 61 (v.-n.: 0.1–0.1). **b** Generated trajectory from MTRNN 2, Test-set 231 (v.-n.: 0.4–0.2). **c** Generated trajectory from MTRNN 3, Test-set 484 (v.-n.: 0.8–0.8)

A more straightforward visualisation of the network performance can be found in Fig. 4, which displays three examples of generated time sequences for motor actions from three MTRNNs. As we calculated in Table 5, the training error became larger when the number of training samples was smaller. In particular, a larger error could be found at the beginning of each time sequence, but the network became stable and generated a stable motor trajectory with less error as time elapsed. There were some errors displayed in the trajectories generation, so sometimes the generated robot behaviours based on the trajectories are biased with the original ones. However, in most of the cases, the generated robot behaviours correctly followed the semantic commands.[2]

4 Generalisation analyses

In this section, we focus on the problem of how the verb-noun generalisation ability of the MTRNN network is achieved. The experiments we showed in the previous section, while only part of the verb-and-noun combinations were presented in the training of the network, it was able to "understand" the un-trained verb-and-noun semantic compositionality. During the training and execution phrases, the iCub learnt and duplicated the actions that the verb instructor speaks with the object that specified in the noun. At the meanwhile, since we trained one object at 6 different locations on the table, the robot can "adjust its attention" toward the intended object at different random locations on the table during execution. For an experiment with a similar aim of generalisation, (Sugita and Tani 2005) reported combining two hierarchical recurrent neural networks which can also accomplish verb-noun generalisation for understanding semantic compositionality in a situated environment. The model they used, called recurrent neural networks with parametric biases units (RNNPB), had similar non-linear dynamics as the MTRNN: the non-linear dynamics are determined by a small number of neural units which act as bifurcation for the whole system.

However, in our case, the learning sequences contain a much larger dimension (35) of the motor joint angles for the iCub movements, compared with motor sequences that trained in Sugita and Tani (2005). Furthermore, while the object appeared at one location in Sugita and Tani (2005), the differences in location of our work also increases the complexity of learning. On the other hand, this complex setting results in the bifurcation which occurs hierarchically in the MTRNN structure, but not been discovered in RNNPB yet.

From this point, we hypothesise that the MTRNN, or any other hierarchical RNNs, results in the separation in the network dynamics about different modalities in a self-organised way associating the semantics with the robot behaviours and the object categories after training. This type of separation should depend on the different organisation of the training data structures, and occurs on different levels of the hierarchical architecture using different strategies. For instance, in Sugita and Tani (2005), such association learning occurring on the PB level binds the semantic and the behaviour representations. Similar association learning also can be found in Heinrich and Wermter (2018). On the other hand, the single RNN we use, although with more complexity in training, allows a higher generalisation abilities because all the modalities are learnt in a single dynamical system. As shown In our experiment setting, after enough training, the synaptic weights between a *basic* motor behaviour (e.g. concepts of "lift")[3] are strengthened about the verb input. And due to its complexity of iCub's (as well as human's) morphology, controlling its behaviours is difficult so it dominates a large portion of the spatio-temporal space in the sensorimotor

[2] https://youtu.be/FOgKbJ-iEhM.

[3] The *basic* motoric perspective of verbs here means that such kind of motor actions belong to general definitions such as "slide", "touch" and etc, without a specific goal for directing action.

Table 6 Removal of data in the 3 × 3 data-set

	Tractor	Hammer	Spikey
Slide left	1		2
Slide right		1/2	
Touch	2		1

The number i in the cell indicates that such a combination was removed in the i-th training set for generalisation experiments

sequences as well as in the neural dynamics. This is similar to the mechanism that the hearing of a verb causes neural firing in the primary motor and pre-motor cortices, corresponding to certain motor action fires when a particular verb is heard or said on the C_s layer. On the contrary, the noun also affects part of the sensorimotor outputs by offsetting the motor actions toward its interacting object, resulting in a specific goal-directed action. This appears to depend on somatotopically mapped parietal regions, parallel to our C_f layer.

In the following experiments, we will examine this hypothesis by means of manipulating data and visualising the training results.

4.1 Generalisation with partial inputs

In this subsection, we concentrate on the comparisons of the results after the removal of different modalities. These comparisons included two parts: i) Error of generalisation after removals; ii) Visualisation of weights after removals.

For the first part of the analysis, in order to obtain a more conclusive statement, we used two sets of data 9 × 9 and 3 × 3 of verb-noun combinations. The 3 × 3 data-set (Table 6) contains a subset of the data-set from previous experiment; it contains the combinations of three actions and three objects, which were placed in 6 different locations. We used a similar look-up table as Table 1 except that only 3 nouns and 3 verbs were used for the vocal command discretisation. For the second part of the experiment, the visualisation of weights was only done with the 3 × 3 data-sets, since its features are easier to observe and its basic principle can be easily extended to the 9 × 9 data-set.

For both parts of the experiment, in order to observe how different lexical categories and visual input affected the training results, especially within the output of the sequences of the motor behaviours, different parts of the input data were removed:

1. No modification (base-line)
2. Remove the noun input (i.e. the first input unit was reset to zero.)
3. Remove the verb input (i.e. the second input unit was reset to zero.)
4. Remove the location of the visual object (i.e. from the third to eighth units were reset to zero.)

During the generalisation tests, the full 3 × 3 or 9 × 9 datasets were placed into the network. The training error and generalisation error of the motor output was compared in Tables 7 and 8. From these two tables, we can see that the removal of the verb resulted in a larger generalisation error than the other two tests, while the removal of the object location resulted in the lowest generalisation error.

For the second part of the experiment, the main aim was to understand the effect of a particular input modality (presenting as semantic structures or visual input) in the whole network activities by observing the visualization of the weights. We conducted an experiment with a smaller data-set (3 × 3), due to the fact that smaller number of weights give a better presentation for the visualization. But a similar conclusion would be extended into the larger 9 × 9 data-set. Figure 5 visualises the weighting matrix, where the neurons from number 0 to number 703 were neurons on the IO layer, from number 704 to number 764 were neurons on the C_f layer and from number 765 to number 794 were neurons on the C_s layer. The weight matrices in Fig. 5a, Fig. 5c and Fig. 5d looked quite similar. But in Fig. 5b, without the verb input, we could easily notice that a large amount of weights from IO layer to C_f remain to be un-trained. To quantitatively evaluate this observation, Table 9 calculated the 2-norm to obtain the Euclidean distances from the manipulated weighting matrices to the base-line matrix. The 2-norm was calculated by:

$$d(\mathbf{W}^m - \mathbf{W}^b) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^m - d_{ij}^b)^2} \tag{15}$$

where $\mathbf{W}^m$ is the weighting matrix after data manipulation, $\mathbf{W}^b$ is the weighting matrix from the base-line experiment, d is the weight from the i-th neuron to j-th neuron. Here $n = 795$ which is the total number of neurons.

From the comparisons of weight matrices and the Euclidean distances, we further verified our hypothesis that the semantic compositionality of verbs represented as motor behaviours plays a significant role in the network since it is further grounded in the differences of motor action trajectories, which dominate a large spatio-temporal space of the sequences.

4.2 Internal dynamics

In the previous analysis, we have looked at the generalisation ability of the MTRNN. A preliminary conclusion suggests that the lexical structure of the verb plays a significant role in maintaining the convergence of the temporal sensorimotor sequences. In this section, we are particularly interested in how the generalisation capabilities are brought by the recurrent connected hierarchical structure. We believed that part of

Table 7 Errors: removal part of input (3 verbs and 3 nouns)

Error	Training: w/o v.	Generalisation: w/o v.	Training: w/o n.	Generalisation: w/o n.	Training: w/o visual	Generalisation: w/o visual
Test 1	0.0003	0.1041	0.0003	0.0594	0.0003	0.0868
Test 2	0.0003	0.1129	0.0003	0.0612	0.0003	0.0933

Table 8 Errors: removal part of input (9 verbs and 9 nouns)

Error	Training: w/o v.	Generalisation: w/o v.	Training: w/o n.	Generalisation: w/o n.	Training: w/o visual	Generalisation: w/o visual
Test 1	0.0003	0.5311	0.0003	0.5223	0.0003	0.0921
Test 2	0.0005	0.6623	0.0005	0.7473	0.0005	0.1379
Test 3	0.0006	0.8574	0.0006	0.7494	0.0006	0.1771

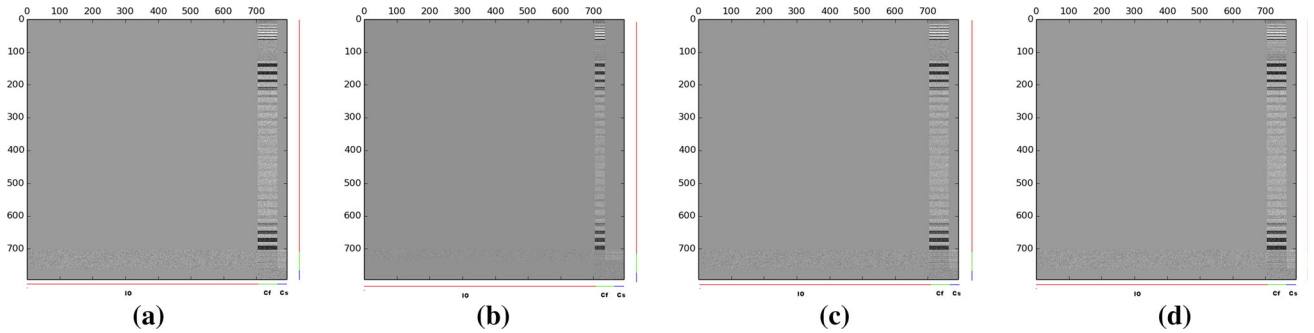


Fig. 5 Weight visualization by input removal: different colours along the axis represent different layers (red: *IO*, green: C_f, blue: C_s) Without the verb input, we could easily notice that a large number of weights from *IO* layer to C_f remain to be un-trained in Fig. 5b. And no big differences can be observed in Fig. 5a, c and d. **a** Weight matrix of normal training (base-line). **b** Weight matrix without verb input. **c** Weight matrix without noun input. **d** Weight matrix without visual input (Color figure online)

Table 9 Euclidean distances between partial input matrices and normal training matrix

	W/o verb	W/o noun	W/o location
Distance	8.9100	0.9450	0.6736

these answers can be found by observing the detailed neural activities on each context layer given the selection of different inputs. The neural activities were therefore examined using the 9×9 data-set, with a previously trained MTRNN with the parameter setting of (70, 3, 50, 120).

The following figures showed the PCA trajectories of the internal neural dynamics on the C_f (Fig. 6) and C_s (Fig. 7) layers. Since the complete 9×9 data-set contains 486 sequences, whose patterns can hardly be observed in one single figure, only a few samples were presented in the following figures to clearly show the PCA trajectories. Figures 6a and 7a showed the selected PCA trajectories on the C_f and C_s layers. These trajectories mainly concern combinations of verb inputs and a few noun inputs. We can see that the verbs mainly determine the patterns of the trajectories, which implies that the motor processing of verbs mainly affects the temporal dynamics in the MTRNN. Since perception and action are intertwined, we expect such neural phenomenon about motor execution exist during both the action execution and observation since the system needs a number of neural dynamics to maintain such motoric memories.

The following figures mainly show how the differences in lexical structures and visual information result in the differences in the PCA trajectories. Figures 6b and 7b show the PCA trajectories of the internal dynamics on C_f and C_s layers, with different noun inputs; Figs. 6c and 7c showed the PCA trajectories with different object location inputs. We could observe that the differences of nouns on the C_f (Fig. 6b) cause divergences at the beginning of the trajectories, but not at the end. From Fig. 6c comparisons show the differences of visual inputs produce even smaller divergences in the trajectories, and that the divergences mainly occurred at the middle of the trajectories. Comparatively, from the activities on the C_s layer (Fig. 7b and c), the divergences of the trajectories from nouns and visual inputs were even smaller: the C_s layer mainly encoded the information from the verbs.

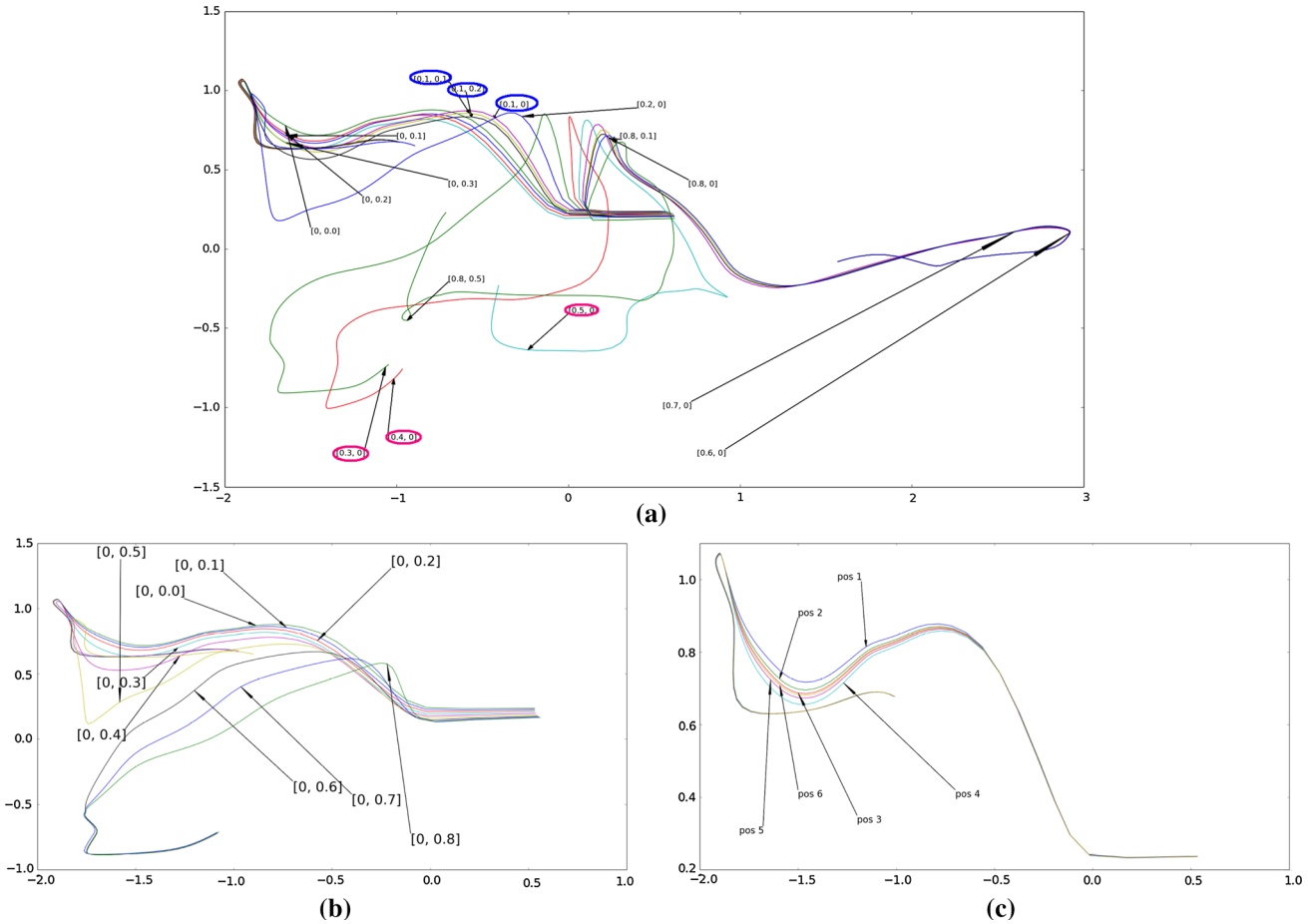


Fig. 6 Principle component analysis on the C_f neurons. With comparison, we can observe the differences in verbs (Fig. 6a) result in larger divergence than nouns and locations. **a** Neural activation C_f from selected sequences. It shows that the sequences with different nouns are clustered closer than those with different verbs. Particularly we can compare (verb-noun) combinations of $(0.3{-}0.5, 0)$ (red) and $(0.1, 0.0{-}0.2)$ (blue). **b** C_f with different nouns. **c** C_f With different object locations (Color figure online)

To summarise the MTRNN analysis, the model self-organises similar patterns on various levels for every sensorimotor sequence, reflecting the hierarchical structure for the vocal commands. Particularly, we can see that the difference between verb inputs results in larger divergence of the trajectories than noun and object-location differences. Due to the data structure of our input vectors, the IO layer represents a collection of each word. With a slower adaptation rate than the IO layer, the C_f represents the grounded meaning of each verb, noun, and visual information. This grounding process is learnt by all temporal sensorimotor sequences. Similarly, using slower changing neurons, the C_s layer represents the general motor behaviour (i.e. the verb) of the whole sensorimotor sequence.

Therefore, the C_f activation mainly represents the lexical structures (verbs and nouns). The visual location has a limited effect on the C_f activation, probably because the information of noun already has overlap with the object information about the visual location. As the main factor of the C_f layer, the same verbs are represented as a similar pattern on the fast context layer in all Fig. 6a–c. The difference from nouns can be observed at the beginning of the trajectories. It correspond to the difference of robot behaviours at the beginning of the time sequences, caused by the neck and eye tracking before the actual hand movement starts. Comparing with the C_f layer, the C_s activation changes even slower. It generally represents the motor behaviours; only the verbs are represented in different patterns.

5 Discussion

5.1 Functional hierarchy of RNN and its bifurcation

It has been reported that quite a few RNN models based on functional hierarchy, such as RNNPB, MTRNN and conceptors (Jaeger 2014), allow the bifurcation to occur in the RNN dynamics. We will give a brief discussion of how this bifurca-

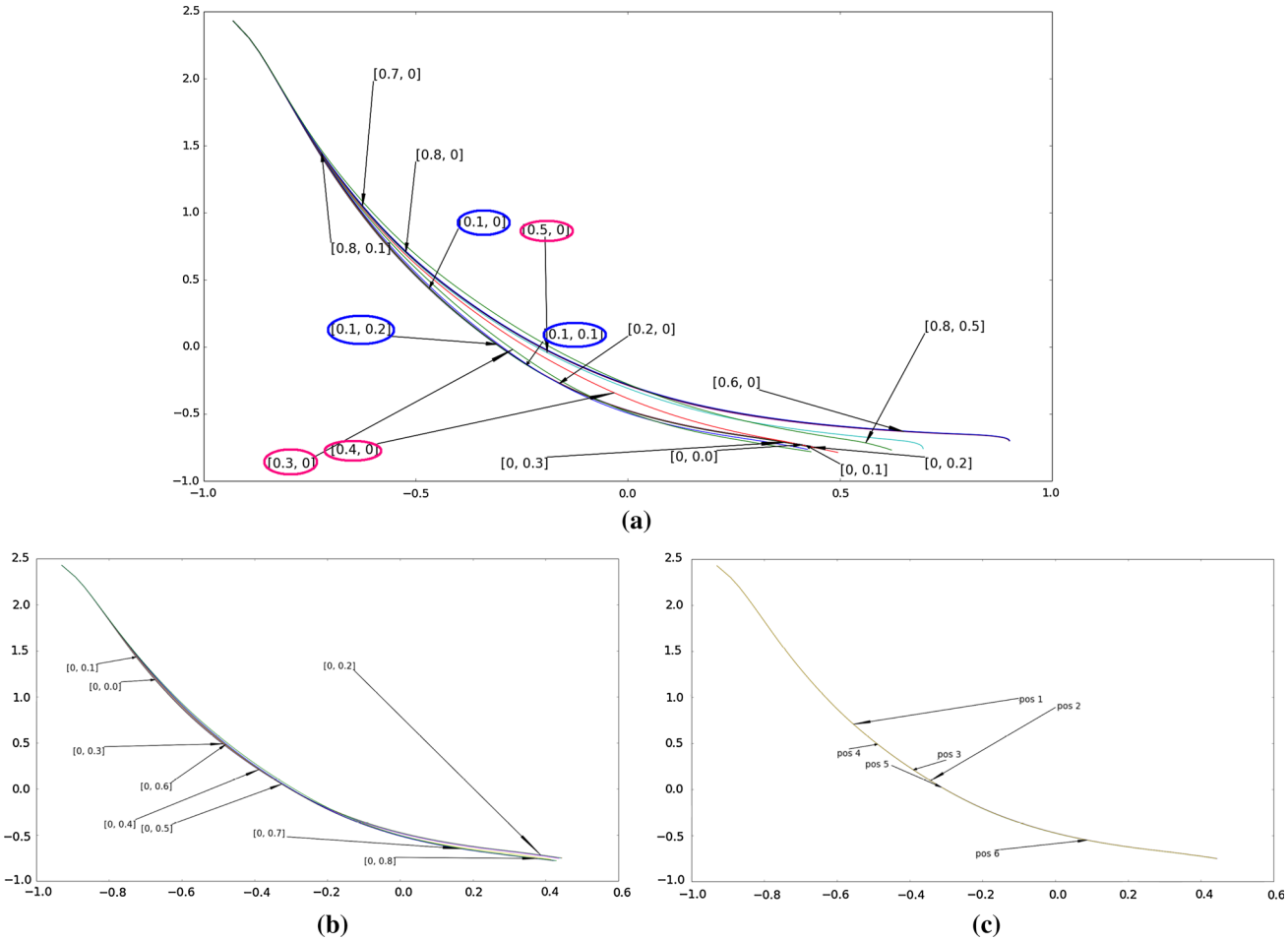


Fig. 7 Principle component analysis on the C_s neurons. With comparison, we can observe the differences in verbs result in larger divergence than nouns and locations. **a** Neural activation C_s from selected sequences. It shows that the sequences with different nouns are clustered closer than those with different verbs. Particularly we can compare (verb-noun) combinations of (0.3−0.5, 0) (red) and (0.1, 0.0−0.2) (blue). **b** C_s with different nouns. **c** C_s With different object locations (Color figure online)

tion happens. Assuming we have a simple hierarchical RNN with an additional unit (which can be regarded as a simplified version of RNNPB) as depicted in Fig. 8. The system can be described as Eq. 16.

$$\begin{cases} \dot{x}_1(t) = -x_1(t) + f(x_3(t)) \\ \dot{x}_2(t) = -x_2(t) + a \cdot f(x_1(t)) + c \cdot PB \\ \dot{x}_3(t) = -x_3(t) + b \cdot f(x_2(t)) \\ y(t) = f(x_3(t)) \end{cases} \tag{16}$$

There are three fixed points in this network. After the network has been trained, i.e. the weights a, b and c are fixed, the coordinates of fixed points only depend upon the value of PB. Furthermore, the coordinates of the fixed points $[x_1, x_2, x_3]$ are first-order functions of the value of PB units (please see appendix for the calculation in details). In other words, the coordinates of the fixed points further determine

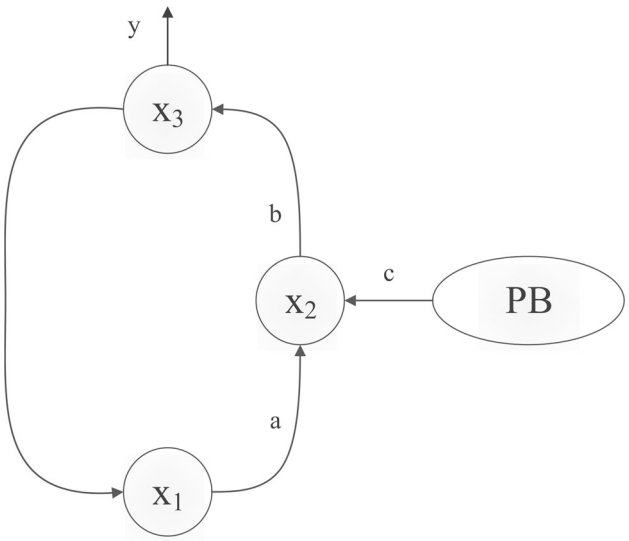


Fig. 8 A simple recurrent network with parametric bias units

the domain of different bifurcation properties. This is the reason that changing the parameter of PB units will change the qualitative structure of the non-linear dynamics of the network. From the bifurcation explanation of the simplified RNNPB model, at the next step we can also extend this to other hierarchical RNNs such as MTRNN, as they are holding a fundamentally similar theoretical foundation (Tani 2014).

5.2 Generalisation ability of MTRNN

In our experiments, the MTRNN was trained under a particular input data structure: Firstly the language commands were recorded as auditory data and transformed into a discrete symbolic representation, and secondly, the object locations and the motor behaviours were also stored as the angles of motor joints. This unique structure is a simplified representation of the common coding theory, which proposes that perceptual inputs and motor actions are sharing the same format of the representation within the cognitive processes.

The neural dynamics in our MTRNN exhibited a dynamics which are different from those reported in Hinoshita et al. (2009) and Heinrich et al. (2015). Whereas the noun (or object perceptual inputs) play a significant factor in the dynamics of context layers in these two examples, our network has minimised the effects of nouns or the object perception. This is partly because of the input data structure where the motor joints of the iCub robot have much larger dimensions than the visual perception input. Also, the spatial information for objects in our experiment setting is much easier to learn, compared to our diversified motor behaviours. The generalisation here concerns more the inference of the symbolic meaning of a language command due to the composition of neural dynamics. During the training in a hierarchical network, such as MTRNN or RNNPB, the neural connections strengthen between a particular type of sensorimotor sequence and visual perception. Particularly, in our case of 9×9 datasets, most of our network weights store the memory of motor actions.

Note that the generalisation of commands in the verb-noun combinations is not the same as we usually do in the generic recurrent neural networks (e.g. Ito and Tani 2004; Pineda 1987; Zhong et al. 2014), which expect the network to do interpolation or extrapolation with a novel input value in either temporal or spatial space. While generalizing dynamical patterns by interpolation is a non-trivial task for training motor patterns in robots, our main concern is the novel combinations in the context of lexicon acquisition. In our case, the learning of verbs and nouns results in the emergence of different dynamics that are mostly stored in different synaptic weights, and thus their combinatorial composition is realised by the non-linearity of the recurrent connections. Considering the different generalisation abilities of generic RNN, RNNPB (Kleesiek et al. 2013; Zhong et al. 2014) and MTRNN (Heinrich et al. 2015), the hierarchical RNNs appear particularly suitable for the production of flexible motor behaviour and language expression simultaneously in the real-world social robot experiments.

5.3 Hierarchical recurrent networks and further development

The hierarchical architecure was proposed the capture the *unpredict* information in the hierarchical architecture. In our application, it mainly captures the verb/motor information.

Furthermore, some machine learning methods have recently been proposed based on the two Hierarchical Recurrent Networks together (Cho et al. 2014), which achieved great performance in machine translation (Sutskever et al. 2014), image captioning (Vinyals et al. 2014), etc. The Encoder-Decoder (ED) architecture usually consists of two recurrent neural networks. One deep RNN network encodes a sequence of input vectors with arbitrary length into a fix-length vector representation in a hierarchical way, while the other deep RNN network decodes this representation into a target sequence of output vector. This specific representation between the encoder and the decoder RNNs is called "thought vectors" which is claimed to represent the meaning of the sequence in a high-dimensional space. The training of such an architecture is done by maximizing the conditional probability of the target sequence. If the input sequence is denoted as $(x_1, x_2, \cdots, x_T)$ and the corresponding output sequence is $(y_1, y_2, \cdots, y_{T'})$ (T does not necessarily equal to T'), the next symbol generation is done by maximising Eq. 17.

$$\prod_{t=1}^{T'} P(y_t | y_{t-1}, y_{t-2}, \cdots, y_1, c) = P(y_{T'}, y_{T'-1}, \cdots, y_1 | x_T, x_{T-1}, \cdots, x_1) \quad (17)$$

Generic RNNs are not able to approximate the probability of the sequence with arbitrary length because of its vanish gradient problem, but other novel RNNs, such as LSTM, BRNN (Bi-directional Recurrent Neural Networks), have been successfully employed to construct the ED architecture to "understand" (encode) and to "generate" (decode) the temporal sequences. Furthermore, due to the recent popularity of parallel computation by GPU, it has become possible to train and use such architectures to solve problems such as machine translation and image captioning.

As the MTRNN can also avoid the vanish gradient problem, and larger MTRNN can be implemented via GPU, it

is also possible to embed the MTRNN into the ED architecture. In fact, the context slow level C_s already exhibits a similar feature of "thought vectors", using a stable neural vector to represent the basic profiles of motor actions and object instances (in our robotic experiment). They also have similar information bi-directional flows which allow the networks to recognise and to generate the time sequences. Despite their similarities, compared with LSTM, the MTRNN have other distinct features: First, from the above experiments and from other MTRNN experiments (Heinrich et al. 2015; Hinoshita et al. 2009), it has been shown that the fast context layers and slow context layers exhibit various dynamics to explicitly represent the relationship between the verbs and nouns. The deep LSTM, on the contrary, has not been reported to have similar dynamics. Second, differently from the static vector representation from LSTM, the context layers allow a "slow" change through time which is more realistic for an interaction environment, where it can be used to dynamically exhibit the meaning of sentences and sensorimotor information.

Admittedly, the training of deep RNNs, e.g. LSTMs and MTRNNs, costs a large amount of computational effort. But the recent development of GPU computing provides an opportunity to construct and test such a big scale neural network with a reasonable time and budget. The combination of MTRNN, the concept of "thought vectors" and its embodiment in robotic systems, will allow us to further explore issues such as:

1. The comparison of the performances of MTRNN, LSTM, and BRNN within the ED architecture and examine their performances in the robotic platforms.
2. The robot motor action, as a natural temporal sequence, can be further incorporated as the training of RNNs of ED architecture with connections to other modalities.

6 Conclusion

This paper presents a neurorobotic study on noun and verb generation and generalisation, utilising with the MTRNN networks, with a large data-set, consisting of vocal language commands, visual object, and motor action data. Although the generalisation abilities of hierarchical RNNs (RNNPB, MTRNN) have been reported in previous research, this is the first study to demonstrate its generalisation capability using such a large data-set, which enables the robot to learn to handle real-world objects and actions. These experiments showed that the generalisation ability of the network is possible even with a large number of test-sets (9 motor actions and 9 objects placing placed in 6 different locations). This is particularly important because the recurrent connections between the verbs and nouns are associated with different modalities of the training-data, which is strengthened during embodiment training by the sensorimotor interaction. Detailed analyses on the robot's neural controller showed that the dynamics on different layers are self-organized in the MTRNN. These self-organised dynamics further constitute a functional hierarchical representation on different layers, which associate different lexical structures with different modalities of the sensorimotor inputs. The MTRNN showed how the embodied information about the verbs dominates a large portion of the network dynamics, since the proprioception information plays a significant role in the training sequences. As such, the hierarchical RNNs, such as MTRNN, are shown to be particularly beneficial in building a neurorobotics cognitive architecture about language learning for robotic systems, where the recurrent connections are able to self-organise and build associations between embodied information in different modalities and the lexical structure information.

Acknowledgements This research has been supported by the EU project POETICON++ under Grant Agreement 288382, the UK EPSRC project BABEL and Waseda SGU Program and the New Energy and Industrial Technology Development Organization (NEDO) of Japan. We are grateful to Dr. Christopher Ford for his helpful review.

Appendix

To calculate the coordinates of the fixed points, we should let $f'(x) = 0$, which means that we need to solve the following equations

$$\begin{cases} -x_1(t) + f(x_3(t)) = 0 \\ -x_2(t) + a \cdot f(x_1(t)) + c \cdot PB = 0 \\ -x_3(t) + b \cdot f(x_2(t)) = 0 \end{cases} \tag{18}$$

The first solution for the first coordinate $[x_1^1, x_2^1, x_3^1]$ is:

$$\begin{cases} x_1^1 = \sqrt{\dfrac{36N^2 + (6M - 6N^2 + Na)^2}{36N^2b^2 + 36N^2 + (6M - 6N^2 + Na)^2}} \\ x_2^1 = \dfrac{a}{6} + \dfrac{M}{N} - N \\ x_3^1 = 6\sqrt{\dfrac{N^2b^2}{36N^2 + (6M + N(-6N + a))^2}} \end{cases} \tag{19}$$

Similarly, the coordinate of the second fixed point $[x_1^2, x_2^2, x_3^2]$ is calculated by:

$$\begin{cases} x_1^2 = \sqrt{\frac{[2250000(173N-100)^2+(-30000M+224727N^2+43250Na+259650N+25000a+75000)^2]}{[225000b^2(173N-100)^2+225000(173N-100)^2+(-30000M+224727N^2+43250Na+259650N+25000a+75000)^2]}} \\ x_2^2 = \frac{a}{6}+\frac{2M}{-1+\sqrt{3}}N-(-\frac{1}{2}+\frac{\sqrt{3}}{2}\cdot N) \\ x_3^2 = 6\sqrt{\frac{b^2\cdot(1.73N-1.0)^2}{-12M+(1.73N-1.0)\cdot(5.196N+a+3.0)^2+36(1.73N-1.0)^2}} \end{cases} \tag{20}$$

And the coordinate of the third fixed point $[x_1^3, x_2^3, x_3^3]$ is given by

$$\begin{cases} x_1^2 = \sqrt{\frac{[2250000(173N+100)^2+(-30000M+224727N^2+43250Na+259650N+25000a+75000)^2]}{[225000b^2(173N+100)^2+225000(173N+100)^2+(-30000M+224727N^2+43250Na+259650N+25000a+75000)^2]}} \\ x_2^2 = \frac{a}{6}+\frac{2M}{-1-\sqrt{3}}N-(-\frac{1}{2}-\frac{\sqrt{3}}{2}\cdot N) \\ x_3^2 = 6\sqrt{\frac{b^2\cdot(1.73N+1.0)^2}{-12M+(1.73N+1.0)\cdot(5.196N+a+3.0)^2+36(1.73N+1.0)^2}} \end{cases} \tag{21}$$

For the above solutions, we define the parameter M as:

$$M = -\frac{a^2}{36} - \frac{b^2}{6} - \frac{1}{3} \tag{22}$$

and N as

$$N = \left[-\frac{a^3}{216} + \frac{ab^2}{4} + \frac{-\frac{ab^2}{2}-1}{12} + \frac{a}{4} + \frac{c\cdot PB}{4} + \sqrt{M^3 + \frac{\left(-\frac{a^3}{108}+\frac{ab^2}{2}+\frac{-\frac{ab^2}{2}-1}{6}+\frac{a}{2}+\frac{c\cdot PB}{2}\right)^2}{4}} \right]^{1/3} \tag{23}$$

Although the equations seem to be complicated, remember that variables a, b and c (weights) are constant after training, which means that M is a constant as well. Thus PB value is a first-order variable in the function of N. Similarly, from observation from Eqs.19–21 we can see that the solutions are first-order function of variable N, which means that the coordinates of this non-linear system are a first-order function of PB.

References

Andreas, J., & Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1173–1182).

Asada, M. (2009). Cognitive developmental robotics: A survey. *IEEE Transactions on Autonomous Mental Development*, *1*(1), 12–34.

Attamimi, M., et al. (2016). Learningword meanings and grammar for verbalization of daily life activities using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models. *Advanced Robotics*, *30*(11–12), 806–824.

Awano, H., et al. (2011). Use of a sparse structure to improve learning performance of recurrent neural networks. In *Neural information processing* (pp. 323–331). Berlin: Springer.

Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain. *Developmental Psychobiology*, *40*(3), 293–310.

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for motionese: Modifications in mothers infantdirected action. *Developmental Science*, *5*(1), 72–83.

Brand, R. J., et al. (2007). Fine-grained analysis of motionese: Eye gaze, object exchanges, and action units in infantversus adult-directed action. *Infancy*, *11*(2), 203–214.

Breazeal, C. L. (2004). *Designing sociable robots*. Cambridge: MIT press.

Cangelosi, A. (2010a). Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, *7*(2), 139–151.

Cangelosi, A. (2010b). Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, *2*(3), 167–195.

Cangelosi, A., & Parisi, D. (2001). How nouns and verbs differentially affect the behavior of artificial organisms. In *Proceedings of the 23rd annual conference of the cognitive science society* (pp. 170–175). London: LEA.

Cangelosi, A., & Parisi, D. (2004). The processing of verbs and nouns in neural networks: Insights from synthetic brain imaging. *Brain and Language*, *89*(2), 401–408.

Cangelosi, A., & Schlesinger, M. (2015). *Developmental robotics: From babies to robots*. Cambridge: MIT Press.

Chang, N., Feldman, J., & Narayanan, S. (2005). Structured connectionist models of language, cognition and action. In *Progress in neural processing* (Vol. 16, p. 57).

Cho, K., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679–704.

Dominey, P. F. (2013). Recurrent temporal networks and language acquisitionfrom corticostriatal neurophysiology to reservoir computing. *Frontiers in Psychology*, *4*, 500.

Dominey, P. F., Mallet, A., & Yoshida, E. (2009). Real-time spoken-language programming for cooperative interaction with a humanoid apprentice. *International Journal of Humanoid Robotics*, *6*(02), 147–171.

Eberhard, K. M. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Gentner, D. (1982). Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In *Center for the study of reading technical report*, no. 257.

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285.

Golosio, B. (2015). A cognitive neural architecture able to learn and communicate through natural language. *PloS ONE*, *10*(11), e0140866.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, *39*(5), 350–365.

Hall, D. G., & Waxman, S. R. (1993). Assumptions about word meaning: Individuation and basic-level kinds. *Child Development*, *64*(5), 1550–1570.

Heinrich, S., Magg, S., & Wermter, S. (2015). Analysing the multiple timescale recurrent neural network for embodied language understanding. In *Artificial neural networks* (pp. 149–174). Berlin: Springer.

Heinrich, S., & Wermter, S. (2018). Interactive natural language acquisition in a multi-modal recurrent neural architecture. *Connection Science*, *30*(1), 99–133.

Hinaut, X., & Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PloS ONE*, *8*(2), e52946.

Hinoshita, W., et al. (2009). Emergence of evolutionary interaction with voice and motion between two robots using RNN. In *IEEE/RSJ international conference on intelligent robots and systems, 2009. IROS 2009* (pp. 4186–4192). IEEE.

Hinoshita, W., et al. (2011). Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network. *Neural Networks*, *24*(4), 311–320.

Holzer, S. (1994). From constructivism to active learning. *The Innovator*, *2*, 4–5.

Howard, I. S., & Messum, P. (2011). Modeling the development of pronunciation in infant speech acquisition. *Motor Control*, *15*(1), 85–117.

Ito, M., & Tani, J. (2004). Generalization in learning multiple temporal patterns using rnnpb. In *Neural information processing* (pp. 592–598). Springer.

Iwahashi, N. (2008). Interactive learning of spoken words and their meanings through an audio-visual interface. *IEICE Transactions on Information and Systems*, *91*(2), 312–321.

Jaeger, H. (2014). Controlling recurrent neural networks by conceptors. arXiv preprint arXiv:1403.3369.

Kaschak, M. P., et al. (2005). Perception of motion affects language processing. *Cognition*, *94*(3), B79–B89.

Kersten, A. W. (1998). An examination of the distinction between nouns and verbs: Associations with two different kinds of motion. *Memory & Cognition*, *26*(6), 1214–1232.

Kleesiek, J., et al. (2013). Action-driven perception for a humanoid. In *Agents and artificial intelligence* (pp. 83–99). Berlin: Springer.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing*, *21*(1), 1–6.

Konidaris, G., Kaelbling, L. P., & Lozano-Perez, T. (2015). Symbol acquisition for probabilistic high-level planning. In *Proceedings of the 24th international conference on artificial intelligence* (pp. 3619–3627). AAAI Press.

Landy, D., Allen, C., & Zednik, C. (2014). A perceptual account of symbolic reasoning. *Frontiers in Psychology*, *5*, 275.

Longobardi, E., et al. (2015). Noun and verb production in maternal and child language: Continuity, stability, and prediction across the second year of life. In *Language learning and development* (pp. 1–16).

Macaluso, E., & Driver, J. (2005). Multisensory spatial interactions: A window onto functional integration in the human brain. *Trends in Neurosciences*, *28*(5), 264–271.

Maguire, M., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). A unified theory of word learning: Putting verb acquisition in context. In *Action meets word: How children learn verbs,* (p. 364).

Mangin, O., & Oudeyer, P. -Y. (2012). Learning to recognize parallel combinations of human motion primitives with linguistic descriptions using non-negative matrix factorization. In *2012 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. (pp. 3268–3275). IEEE.

Matuszek, C., et al. (2013). Learning to parse natural language commands to a robot control system. In *Experimental robotics* (pp. 403–415). Berlin: Springer.

Metta, G., et al. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems.* (pp. 50–56). New York: ACM.

Mirolli, M., & Parisi, D. (2009). Language as a cognitive tool. *Minds and Machines*, *19*(4), 517–528.

Mirolli, M., & Parisi, D. (2011). Towards a Vygotskyan cognitive robotics: The role of language as a cognitive tool. *NewIdeas in Psychology*, *29*(3), 298–311.

Misra, D. K., et al. (2014). Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions. In *Proceedings of robotics: science and systems (RSS)*, Berkeley, USA.

Noë, A. (2001). Experience and the active mind. *Synthese*, *129*(1), 41–60.

Novianto, R. (2014). Flexible attention-based cognitive architecture for robots. In *International conference on social robotics* (pp. 279–289). Berlin: Springer.

Ogata, T., & Okuno, H. G. (2013). Integration of behaviors and languages with a hierarchal structure selforganized in a neurodynamical model. In *2013 IEEE workshop on robotic intelligence in informationally structured space, RiiSS 2013–2013 IEEE symposium series on computational intelligence*. SSCI.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying different-modality properties for concepts produces switching costs. *Psychological Science*, *14*(2), 119–124.

Peniak, M., et al. (2011). Aquila:Anopen-source GPU-accelerated toolkit for cognitive and neuro-robotics research. In *The 2011 international joint conference on neural networks (IJCNN)* (pp. 1753–1760). New York: IEEE.

Pineda, F. J. (1987). Generalization of back-propagation to recurrent neural networks. *Physical Review Letters*, *59*(19), 2229.

Pulvermüller, F. (2002). *The neuroscience of language: On brain circuits of words and serial order*. Cambridge: Cambridge University Press.

Reale, R. A., & Imig, T. J. (1980). Tonotopic organization in auditory cortex of the cat. *Journal of Comparative Neurology*, *192*(2), 265–291.

Rohlfing, K. J. (2016). An alternative to mapping a word onto a concept in language acquisition: Pragmatic frames. *Frontiers in Psychology*, *7*, 470.
Saygin, A. P. (2010). Modulation of BOLD response in motion-sensitive lateral temporal cortex by real and fictive motion sentences. *Journal of Cognitive Neuroscience*, *22*(11), 2480–2490.
Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, *15*, 31–90.
Sperry, R. W. (1952). Neurology and the mind-brain problem. *American Scientist*, *40*(2), 291–312.
Steels, L., & Hild, M. (2012). *Language grounding in robots*. Berlin: Springer Science & Business Media.
Stramandinoli, F., Marocco, D., & Cangelosi, A. (2012). The grounding of higher order concepts in action and language: A cognitive robotics model. *Neural Networks*, *32*, 165–173.
Sugita, Y., & Tani, J. (2005). Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adaptive Behavior*, *13*(1), 33–52.
Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).
Tani, J. (2014). Self-organization and compositionality in cognitive brains: A neurorobotics study. *Proceedings of the IEEE*, *102*(4), 586–605.
Tani, J., Ito, M., & Sugita, Y. (2004). Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks*, *17*(8), 1273–1289.
Tellex, S., et al. (2011). Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*.
Tenenbaum, J. B. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.
Tomasello, M., & Farrar, M. J. (1986). Object permanence and relational words: A lexical training study. *Journal of Child Language*, *13*(03), 495–505.
Van Essen, D. C. (1985). Functional organization of primate visual cortex. *Cerebral Cortex*, *3*, 259–329.
Vinyals, O., et al. (2014). Show and tell: A neural image caption generator. arXiv preprint arXiv:1411.4555.
Warlaumont, A. S., et al. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, *38*, 64–75.
Weng, J., et al. (2001). Autonomous mental development by robots and animals. *Science*, *291*(5504), 599–600.
Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880.
Yamashita, Y., & Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Computational Biology*, *4*(11), e1000220.
Yürüten, O., Şahin, E., & Kalkan, S. (2013). The learning of adjectives and nouns from affordance and appearance features. *Adaptive Behavior*, *21*(6), 437–451.
Zhong, J. (2015) . Artificial neural models for feedback pathways for sensorimotor integration.
Zhong, J., & Canamero, L. (2014). From continuous affective space to continuous expression space: Non-verbal behaviour recognition and generation. In *2014 Joint IEEE international conferences on development and learning and epigenetic robotics (ICDLEpirob)*. IEEE. (pp. 75–80).
Zhong, J., Cangelosi, A., & Wermter, S. (2014). Toward a self-organizing pre-symbolic neural model representing sensorimotor primitives. *Frontiers in Behavioral Neuroscience*, *8*, 22.
Zhong, J., Weber, C., & Wermter, S. (2011). Robot trajectory prediction and recognition based on a computational mirror neurons model. In *Artificial neural networks and machine learning-ICANN 2011* (pp. 333–340). Berlin: Springer.
Zhong, J., Weber, C., & Wermter, S. (2012a). Learning features and predictive transformation encoding based on a horizontal product model. In *Artificial neural networks and machine learning-ICANN 2012* (pp. 539–546). Berlin: Springer.
Zhong, J., Weber, C., & Wermter, S. (2012b). A predictive network architecture for a robust and smooth robot docking behavior. *Paladyn, Journal of Behavioral Robotics*, *3*(4), 172–180.

Junpei Zhong is a researcher at the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo. He is also a visiting researcher at Plymouth University, UK. He obtained a B.Eng from South China University of Technology, a M.Phil degree from the Hong Kong Polytechnic University and a doctoral degree (Dr.rer.nat) from the University of Hamburg. Between year 2010 and 2015, he has been engaged in several EU FP7 projects on cognitive robotics (RobotDoC, ALIZ-E and POETICON++) in Germany and the UK. His research is committed to build machine learning models for cognitive robots and service robots.

Martin Peniak is a development engineer at Cortexica Vision Systems, London, UK, after he graduated from Plymouth University, UK. His research interests include parallel programming, cognitive robotics, astrophotography.

Jun Tani received a B.S. in Mechanical Engineering from Waseda University, a dual M.S. in Electrical Engineering and Mechanical Engineering from the University of Michigan, and a Dr. Eng. from Sophia University in 1995. He started his research career in Sony Laboratory in 1990. He had been a Team Leader of the Lab. for Behavior and Dynamic Cognition, Brain Science Institute, RIKEN in Tokyo for 12 years until May 2012, and a professor at Korean Advanced Institute of Science and Technology (KAIST) from 2012 to 2017. He also held the position

of Visiting Associate Professor at the Univ. of Tokyo between 1997 and 2002. Now he is a full professor at Okinawa Institute of Science and Technology (OIST), Japan. He serves as editorial board members for IEEE Trans. on AMD, Connection Science, Frontiers in Neurorobotics and Adaptive Behavior. His research interests include neurorobotics, neural network models, embodied cognition, phenomenology and complex systems.

Tetsuya Ogata received the BS, MS and DE degrees in Mechanical Engineering, in 1993, 1995 and respectively, from Waseda University. From 1999 to 2001, he was a Research Associate in Waseda University. From 2001 to 2003, he was a Research Scientist in the Brain Science Institute, RIKEN. From 2003 to 2012, he was an Associate Professor in the Graduate School of Informatics, Kyoto University. Since 2012, he has been a Professor of the Faculty of Science and Engineering, Waseda University. From 2009 to 2015, he was a JST (Japan Science and Technology Agency) PRESTO Researcher (5 years). From 2015, he has been a Visiting Researcher of Artificial Intelligence Research Center, AIST. His research interests include human-robot interaction, dynamics of human-robot mutual adaptation and inter-sensory translation in robot systems with neuro-dynamical models.

Angelo Cangelosi is Professor of Artificial Intelligence and Cognition and the Director of the Centre for Robotics and Neural Systems at Plymouth University (UK). Cangelosi's main research expertise is on language grounding and embodiment in humanoid robots, developmental robotics, human-robot interaction, and on the application of neuromorphic systems for robot learning. Cangelosi has produced more than 250 scientific publications. Overall, he has secured over £15m of research grants as coordinator/PI (e.g. Coordinator of H2020 ITN APRIL). He has chaired numerous workshops and conferences including the IEEE ICDL-EpiRob 2011 and 2013 Conferences (Frankfurt 2011, Osaka 2013). In 2012-13 he was Chair of the IEEE Technical Committee on Autonomous Mental Development. Cangelosi is Editor (with K. Dautenhahn) of the journal Interaction Studies, and in 2015 was Editor-in-Chief of IEEE Transactions on Autonomous Development. His latest book Developmental Robotics: From Babies to Robots (MIT Press; co-authored with Matt Schlesinger) was published in January 2015.