In: Proceedings of the 24th International Conference on Artificial Neural Networks (ICANN 2014),
LNCS 8681, pp. 193-200, Springer Heidelberg. Hamburg, DE, September 15-19, 2014.

# Interactive Language Understanding with Multiple Timescale Recurrent Neural Networks

Stefan Heinrich and Stefan Wermter

University of Hamburg, Department of Informatics, Knowledge Technology
Vogt-Kölln-Straße 30, D - 22527 Hamburg, Germany
{heinrich,wermter}@informatik.uni-hamburg.de
http://www.informatik.uni-hamburg.de/WTM/

**Abstract.** Natural language processing in the human brain is complex and dynamic. Models for understanding, how the brain's architecture acquires language, need to take into account the temporal dynamics of verbal utterances as well as of action and visual embodied perception. We propose an architecture based on three Multiple Timescale Recurrent Neural Networks (MTRNNs) interlinked in a cell assembly that learns verbal utterances grounded in dynamic proprioceptive and visual information. Results show that the architecture is able to describe novel dynamic actions with correct novel utterances, and they also indicate that multi-modal integration allows for a disambiguation of concepts.

**Keywords:** Actions, Embodied, MTRNN, Language Acquisition

## 1 Introduction

Natural language is the cognitive capability that clearly distinguishes humans from other living beings but is by far not yet understood. Methods in neuroscience to examine processes of natural language in the brain have evolved, but are mostly restricted to non-invasive external measurements. As a direct consequence, we have limited knowledge of the characteristics of natural language in terms of connectivity, plasticity, and temporal dynamics. On the other hand, we are able to build up analogies in neuro-robotic agents that are grounded in real world scenarios to study plausible characteristics [6, 19].

Traditional generativist's theories for explaining language acquisition often fail to explain 'the how', because they leave out the evident involvement of modal embodiment and situated context, for example the close relation of language processing and motor action [3, 12, 15]. Recent constructivist's hypotheses aim to capture these characteristics, e.g. by grounding higher order symbols in action primitives and action primitives in sensory-motor experience [6, 17, 18]. Although these models help to understand how concepts can be bound to sensation, they are limited by the assumption of having a built-in lexicon of items on word level, not capturing the complex temporal dynamics in language.

In a recent study Hinoshita et al. claim that an "appropriate architecture is sufficient" for the emergence of language [11]. They state that information processing on different timescales in a *Multiple Timescale Recurrent Neural Network*

(MTRNN) can lead to the acquisition of language in a self-organised way. In our previous study we followed up on this approach and showed that a hierarchical structure with timescales and the embodiment of language in perception are important aspects of an appropriate architecture for language [9]. We demonstrated that our extended MTRNN is capable of generalisation in terms of describing novel perceived scenes with correct novel verbal utterances.

In this paper we scale up our architecture to multi-modal processing of proprioceptive and visual embodied perception, where information for all modalities is processed in similar recurrent network structures and are integrated in a cell assembly. We show that the architecture can generalise well for producing novel verbal utterances for novel dynamic visual and motor stimuli. Additionally, we analyse how embodied perception is processed and abstracted and how concepts for the respective scenes are formed in the cell assembly.

## 2 Multi-modal MTRNNs model

In our approach we employ an MTRNN to process verbal utterances over time, extended by an MTRNN to process motor proprioception, an MTRNN to process visual perception, linked by a cell assembly of fully connected neurons that process and represent the concepts of the information [5]. Refining our previous extended MTRNN model [9], we aim at capturing the hierarchical structure found for action processes in the cortex [2] and account for cortical cell assemblies that interlink concepts in language, action, and perception [15]. In Fig. 1 we provide an overview of our proposed new architecture.

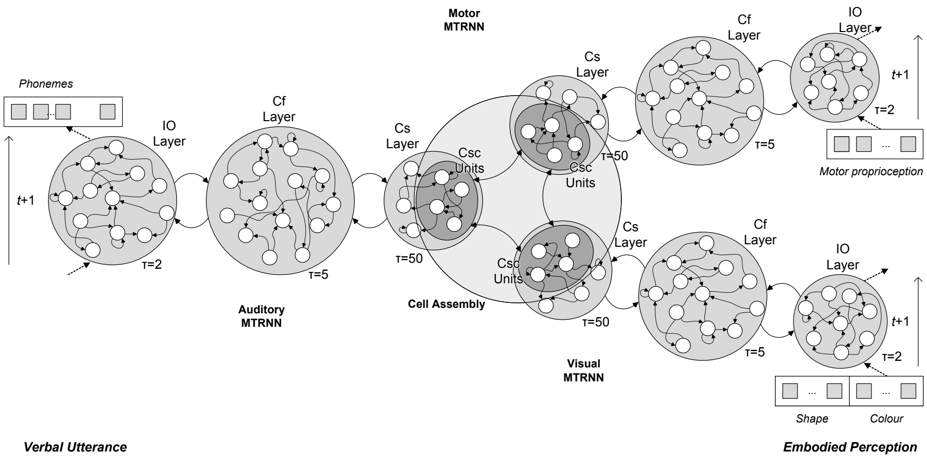


**Fig. 1.** Architecture of the multi-modal MTRNN model, consisting of MTRNNs for auditory, proprioceptive, and visual information processing as well as a cell assembly for representing and processing the concepts. A sequence of phonemes (utterance) is produced over time, based on sequences of embodied multi-modal perception.

The MTRNNs are each composed by an Input- and Output layer (IO) and two context layers called Context fast (Cf) and Context slow (Cs) [20]. Compared to other RNNs, the specific characteristic of an MTRNN is the full connectivity of neurons to all neurons of the same and of adjacent layers, and the mechanism of processing information with increasing timescales $\tau$. The cell assembly interconnects these networks and integrates information processed over time.

### Training

During training of the system, verbal utterances are presented together with sequences of the proprioceptive and visual stimuli of an action sequence. For the production of utterances, the auditory MTRNN self-organises the weights and also the internal states of some of the neurons in the Cs layer (denoted *Context Controlling units* (Csc)). In parallel, the motor MTRNN and the visual MTRNN self-organise the weights and also the internal states of the Csc units for the processing of an incoming perception. The important difference is that the auditory MTRNN self-organises towards the *initial* internal states of the Csc (start of utterance), while the motor MTRNN and the visual MTRNN self-organise towards the *final* internal states of the Csc (end of perception). Finally, the activity of the Csc units of all MTRNNs get associated in the cell assembly.

For the training of the auditory MTRNN we employ an adaptive mechanism, which is a variant of the *real-time backpropagation through time* (RTBPTT) algorithm [8]. Since we aim at an abstraction from the perception to concepts we modify the partial derivatives for the internal state $z$ at time step $t$ of neurons $i \in I_{\text{all}} = I_{\text{IO}} \cup I_{\text{Cf}} \cup I_{\text{Cs}}$ for the proprioceptive and visual MTRNN as follows:

$$\frac{\partial E}{\partial z_{t,i}} = \begin{cases} (1-\psi)\, f'(y_{t,i})\,(y_{t,i} - f(c_{T,i} + b_i)) & \text{iff } i \in I_{\text{Csc}} \wedge t = T \\ f'(y_{t,i}) \sum\limits_{k \in I_{\text{all}}} \frac{w_{k,i}}{\tau_k} \frac{\partial E}{\partial z_{t+1,k}} + \left(1 - \frac{1}{\tau_i}\right) \frac{\partial E}{\partial z_{t+1,i}} & \text{otherwise} \end{cases} , \tag{1}$$

where $f$ and $f'$ denote an arbitrary differentiable transfer function and its derivative respectively, $b$ and $w$ are the biases and weights, $y$ denotes the neurons output, and $c_{T,i}$ are internal states at the final time step $T$ of the Csc units $i \in I_{\text{Csc}} \subset I_{\text{Cs}}$. Here, we also introduce a very small *self-organisation-forcing* constant $\psi$ that allows the final internal states $c_{T,i}$ of the Csc units to adapt upon the data, although they actually serve as a target for shaping the weights of the network. Accordingly, the final internal states $c_{T,i}$ of the Csc units define the abstraction of the input data and are also updated as follows:

$$c_{T,i}^{n+1} = c_{T,i} + \psi \zeta_i \frac{\partial E}{\partial c_{T,i}} = c_{T,i} + \psi \zeta_i \frac{1}{\tau_i} \frac{\partial E}{\partial z_{T,i}} \quad \text{iff } i \in I_{\text{Csc}} \quad , \tag{2}$$

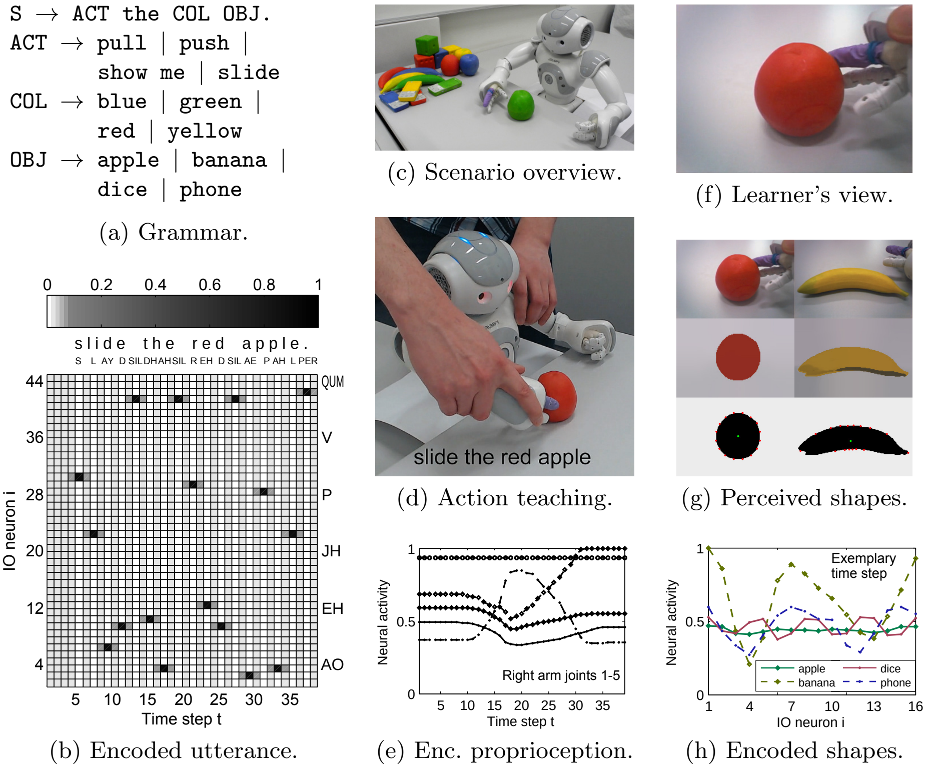
where $\zeta_i$ denotes the learning rates for the changes. Further formal descriptions of the MTRNN can be found in the work of Yamashita and Tani [20]. In our study we specify the timescale parameters as depicted in Fig. 1 based on previous studies [1, 20] and our experiences [10]. The associations of the Csc units are trained with the conventional delta rule.

### Generation

With a trained network the generation of novel verbal utterances from proprioception and visual input can be tested. The final Csc values of the respective MTRNNs are abstracted from the input sequences and associated with initial Csc values of the auditory MTRNN. These values in turn initiate the generation of a phoneme sequence.

## 3 Scenario

To understand and to approach a plausible architecture for the emergence of language, we believe it is crucial to ground the analysis in raw real world perception [7]. We therefore base the scenario in the interaction of a human teacher with a robotic learner that is supposed to acquire and ground language in embodied and situated experience. In particular a humanoid robot, NAO, should learn verbal utterances for manipulation actions of various objects to be able to describe novel actions with correct novel verbal utterances (Fig. 2c for an overview).



```
S   → ACT the COL OBJ.
ACT → pull | push |
      show me | slide
COL → blue | green |
      red | yellow
OBJ → apple | banana |
      dice | phone
```

(a) Grammar.

(b) Encoded utterance.

(c) Scenario overview.

(d) Action teaching.

(e) Enc. proprioception.

(f) Learner's view.

(g) Perceived shapes.

(h) Encoded shapes.

**Fig. 2.** Representations and scenario of language learning in human-robot interaction.

For a given scene the teacher guides the robot's arm in an interaction with a coloured object and describes the action verbally, e.g. "SLIDE THE RED APPLE". Later, the robot should be able to describe a new interaction composed of motor movements and visual experience that it may have seen before with a verbal utterance, e.g. "SHOW ME THE YELLOW APPLE".

The scenario should be controllable in terms of combinatorial complexity and mechanical feasibility for the robot, but at the same time allow for valuable analysis. For this reason we limit the corpus of verbal utterances to a small grammar as summarised in Fig 2a. We use unambiguous utterances for four different actions that can be performed differently by the NAO and four different objects with similar mass but different shapes, each in four different colours.

To obtain a biologically-inspired auditory representation, we transform every sentence from the grammar to a phonetic utterance based on the ARPAbet with additional signs for pauses and intonation of propositions, exclamations, and questions: $\Sigma = \{\text{'AA'}, ..., \text{'ZH'}\} \cap \{\text{'SIL'}, \text{'PER'}, \text{'EXM'}, \text{'QUM'}\}$, $|\Sigma| = 44$. In the next step we encode the utterance $u = (p_1, \ldots, p_l)$ into neural activation over time, where the occurrence of a phoneme $p_k$ is represented by a spike-like neural activity of a specific neuron $i$ at relative time step $r$ with some activity spreading backwards in time (rising phase) and some activity spreading forward in time (falling phase), represented as a Gaussian. A detailed description of the encoding and the used parameters can be found in [10] and [11]. The ideal neural activation for an encoded sample utterance is presented in Fig. 2b.

To gather and encode the proprioception of a corresponding action, we guide the NAO's right arm and directly measure the joint angles of five joints with a sampling rate of 20 frames per second and scale the values to $[0, 1]$, based on the minimal and maximal joint positions (see Fig. 2d). Having an encoding on the joint angle level is biologically plausible, because the (human) brain merges information from joint receptors, muscle spindles, and tendon organs into a similar proprioception representation in the S1 area [4]. Fig. 2e shows the encoded proprioception for an exemplary action.

For the visual perception we aim at capturing a representation that is biologically plausible but on a level of abstraction of shapes as found in the *posterior infero-temporal* (PIT)/V4 area [14]. Specifically, we obtain the objects shape in NAO's field of view and capture 16 points around the object from which we determine the distance to the objects centre of mass divided by the area of the object, and scale to $[0, 1]$ for all shapes (see [10] for details). The measured shape features are invariant to rotation and scaling and capture the shape persistently over time. Additionally, we obtain the shape's colour by determining the average red, green, and blue values within the object's shape. With this method we encode the perception into a sequence of equal sampling rate and length compared to the proprioception sequence. Fig. 2f–h show the NAO's view, respective object perception, and representative differences in the encoding of the objects.

Generating novel utterances from a trained system by presenting new interactions only depends on the calculation time needed for the pre-processing and encoding, and can be done in real time. No additional training is needed.

## 4 Analysis

To learn from the model's characteristics we are interested in the generalisation capabilities and the information pattern that emerges in the cell assembly. We recorded the 64 different interactions four times each with the same verbal utterance and arm starting position but with slightly varying movements and object placements to collect a data set. We divided the data 50:50 into training and test set (all variants of a specific interaction are either in the training or in the test set only) and trained ten randomly initialised systems. We repeated this process ten times with different distributions of data in training and test set to arrive at 100 runs for analysis. The MTRNNs were parametrised as follows: the auditory MTRNN consisted of $|I_{\mathrm{Cf}}| = 80$ and $|I_{\mathrm{Cs}}| = 23$ neurons; the motor and visual MTRNNs were set up each with $|I_{\mathrm{Cf}}| = 40$ and $|I_{\mathrm{Cs}}| = 23$ neurons. The number of IO neurons in all three MTRNNs were based on the representations for utterances, proprioception, and visual perception and set to 44, 5, and 19 respectively, while the number of Csc units was set to $|I_{\mathrm{Csc}}| = \lceil |I_{\mathrm{Cs}}|/2 \rceil$. We initialised all weights in the interval $[-0.025, 0.025]$ as well as the initial Csc in the intervals $[-0.01, 0.01]$ (auditory MTRNN) and $[-1.0, 1.0]$ (motor and visual MTRNNs) and set the feedback rate $\varphi = 0.1$ for the auditory MTRNN as well as the self-organisation rate $\psi = 0.001$ for the motor and visual MTRNN.

### Generalisation of novel interactions

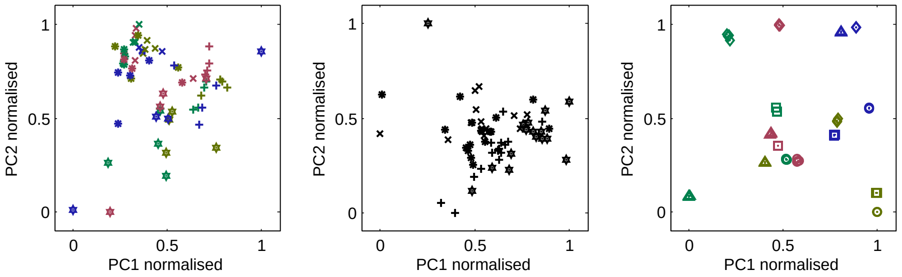
The results of the experiment show that the system is able to generalise: We obtained an $F_1$-score of 0.984 on the training and 0.476 on the test set for the best network. Although training is hard and we rarely obtained perfect but not over-fitted systems on the training data, we observed a high precision (small number of false positives) with a lower to medial recall (not exact production of desired positives) on the test data. The errors made in production were mostly minor substitution errors (single wrong phonemes) and only rarely word errors. Also, we learned that the timescale values were not crucial, but that ideal sizes of the Cf and Cs layers depend on the complexity of the problem (compare [10]).

Using a self-organisation mechanism on the final initial Csc values for the motor and visual MTRNNs caused good abstraction from the perception for the described scenario and $\psi$ value. We learned: to avoid that the cell assembly part does not converge well and produces activity for the auditory MTRNN leading to production of erroneous utterances (false positives or incomplete utterances) up to arbitrary phoneme babbling, the $\psi$ should not be very small. To avoid that the MTRNNs contribute different patterns for the same interactions to the auditory MTRNN, due to premature convergence to the final Csc values, without appropriately self-organising the weights, the $\psi$ should not be very large.

### Self-organisation in the cell assembly

Across all experiments we observed diverse patterns in the internal states of the Csc units, but always found similar patterns in the respective modality for similar utterances or perceptions. A *principal component analysis* (PCA) on the

Csc values revealed that these patterns are not very distinct for motor movements but most distinct for visual perceptions (compare Fig. 3). However, different shapes have slightly different affordances for the same 'type' of movements (e.g. different wrist angles for sliding the banana or sliding the apple), which indicates that the cell assembly integrates the perception of the different shapes and the variances of the movements into a more distinct meaning for the auditory Csc.



(a) Csc auditory. (b) Csc proprioception. (c) Csc visual.

**Fig. 3.** Internal state values of the Csc units for the initial (auditory) and final time step (motor and visual) reduced from $|I_{\mathrm{Csc}}|$ to two dimensions (PC1 and PC2) and normalised. Different shapes/movements and colours are shown with different coloured markers. For Csc auditory the distinction between shapes has been omitted for clarity.

## 5 Conclusion

To learn the brain's architectural characteristics for language acquisition, we believe, it is crucial to enable an agent to a) acquire concepts from temporally dynamic and raw sensory input and b) base information processing on hierarchical abstraction and integration of all modalities. Neuroscientists suggested that temporal dynamics can be explained by synfire chains [16], however this level of detail would hinder an analysis to learn the key characteristics on cortex level and the models would be hard to test due to the complexity. Even with our RNN architecture, which consists of similar components for different modalities, training is very complex, due to a vast parameter space and a very deep network structure. This challenge is similar to deep NNs and ongoing research [13].

Our architecture reveals the importance of the suggested characteristics and shows that verbal utterances can be learned and grounded in dynamic multimodal perception of actions, and can produce novel utterances for novel actions. The integration of multi-modal perceptions on concept level leads to a representation that disambiguates the meaning for proprioceptive and visual perceptions.

**Acknowledgments.** The authors would like to thank S. Magg and C. Weber for critical and inspiring discussions as well as G. Parisi and F. v. Stosch for support with the robot interaction data acquisition.

## References

1. Alnajjar, F., Yamashita, Y., Tani, J.: The hierarchical and functional connectivity of higher-order cognitive mechanisms: neurorobotic model to investigate the stability and flexibility of working memory. Front. Neurorobotics 7(2), 13 p. (2013)
2. Badre, D., Kayser, A.S., D'Esposito, M.: Frontal cortex and the discovery of abstract action rules. Neuron 66(2), 315–326 (2010)
3. Barsalou, L.W.: Grounded cognition. Annu. Rev. Psychol. 59, 617–645 (2008)
4. Bear, M.F., Connors, B.W., Paradiso, M.A.: Neuroscience: Exploring the Brain, 3rd edn. Lippincott Williams & Wilkins (2006)
5. Braitenberg, V.: Cell assemblies in the cerebral cortex. In: Theoretical Approaches to Complex Systems, pp. 171–188. Springer Ber. Hdb. (1978)
6. Cangelosi, A.: Grounding language in action and perception: From cognitive agents to humanoid robots. Physics of Life Reviews 7(2), 139–151 (2010)
7. Feldman, J.A.: The neural binding problem(s). Cogn. Neurodyn. 7(1), 1–11 (2013)
8. Heinrich, S., Weber, C., Wermter, S.: Adaptive learning of linguistic hierarchy in a multiple timescale recurrent neural network. In: Proc. ICANN 2012. LNCS 7552, 555–562, Springer Heidelberg. Lausanne, CH (Sep. 2012)
9. Heinrich, S., Weber, C., Wermter, S.: Embodied language understanding with a multiple timescale recurrent neural network. In: Proc. ICANN 2013. LNCS 8131, 216–223, Springer Heidelberg. Sofia, BG (Sep. 2013)
10. Heinrich, S., Magg, S., Wermter, S.: Analysing the multiple timescale recurrent neural network for embodied language understanding. In: Koprinkova-Hristova, P.D., Mladenov, V.M., Kasabov, N.K. (eds.) Artificial Neural Networks - Methods and Applications in Bio- and Neuroinformatics, SSBN 4, 149–174. Springer International Publishing (2014)
11. Hinoshita, W., Arie, H., Tani, J., Okuno, H.G., Ogata, T.: Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network. Neural Networks 24(4), 311–320 (2011)
12. Hoffmann, T., Trousdale, G. (eds.): The Oxford handbook of construction grammar. Oxford Univ. Press (2013)
13. Larochelle, H., Bengio, Y., Bengio, J., Lamblin, P.: Exploring strategies for training deep neural networks. The Journal of Machine Learning Research 10, 1–40 (2009)
14. Orban, G.A.: Higher order visual processing in macaque extrastriate cortex. Physiological Reviews 88(1), 59–89 (2008)
15. Pulvermüller, F., Moseley, R.L., Egorova, N., Shebani, Z., Boulenger, V.: Motor cognition–motor semantics: Action perception theory of cognition and communication. Neuropsychologia 55, 71–84. (2014)
16. Pulvermüller, F., Shtyrov, Y.: Spatiotemporal signatures of large-scale synfire chains for speech processing as revealed by MEG. Cereb. Cortex 19(1), 79–88 (2009)
17. Roy, D., Mukherjee, N.: Towards situated speech understanding: Visual context priming of language models. Computer Speech and Language 19, 227–248 (2005)
18. Stramandinoli, F., Marocco, D., Cangelosi, A.: The grounding of higher order concepts in action and language: A cognitive robotics model. Neural Networks 32, 165–173 (2012)
19. Wermter, S., Page, M., Knowles, M., Gallese, V., Pulvermüller, F., Taylor, J.G.: Multimodal communication in animals, humans and robots: An introduction to perspectives in brain-inspired informatics. Neural Networks 22(2), 111–115 (2009)
20. Yamashita, Y., Tani, J.: Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. PLoS Computational Biology 4(11), e1000220 (2008)