Interactive Natural Language Acquisition in a Multi-modal Recurrent Neural Architecture

Stefan Heinrich* and Stefan Wermter

Knowledge Technology Institute, Department of Informatics, Universität Hamburg, Vogt-Koelln Straße 30, Hamburg, Germany

(*submitted June 2016, revision December 2016, accepted February 2017*)

The human brain is one of the most complex dynamic systems that enables us to communicate in natural language. We have a good understanding of some principles underlying natural languages and language processing, some knowledge about socio-cultural conditions framing acquisition, and some insights about *where* activity is occurring in the brain. However, we were not yet able to understand the behavioural and mechanistic characteristics for natural language and *how* mechanisms in the brain allow to acquire and process language.

In an effort to bridge the gap between insights from behavioural psychology and neuroscience, the goal of this paper is to contribute a computational understanding of the appropriate characteristics that favour language acquisition, in a brain-inspired neural architecture. Accordingly, we provide concepts and refinements in cognitive modelling regarding principles and mechanisms in the brain – such as the hierarchical abstraction of context – in a plausible recurrent architecture. On this basis, we propose neurocognitively plausible model for embodied language acquisition from real world interaction of a humanoid robot with its environment. The model is capable of learning language production grounded in both, temporal dynamic somatosensation and vision. In particular, the architecture consists of a continuous time recurrent neural network, where parts have different leakage characteristics and thus operate on multiple timescales for every modality and the association of the higher level nodes of all modalities into cell assemblies. Thus, this model features hierarchical concept abstraction in sensation as well as concept decomposition in production, multi-modal integration, and self-organisation of latent representations.

Keywords: Language Acquisition; Recurrent Neural Networks; Embodied Cognition; Multi-modal Integration; Developmental Robotics

1. Introduction

The human brain is seen as one of the most complex and sophisticated dynamic systems. Humans can build precise instruments and write essays about higher purpose of life because they have reached a state of specialisation and knowledge by externalising information and by interaction with each other. We not only utter short sounds to indicate an intention, but also describe complex procedural activity, share abstract declarative knowledge, and may even completely think in language (Bergen, 2012; Christiansen and Chater, 2016; Feldman, 2006; Håkansson and Westander, 2013). For us, it is extremely easy as well as important to share information about matter, space, and time in complex interactions through natural language. Often it is claimed that language is the cognitive capability that differentiates humans most from other beings in the animal kingdom.

*Corresponding author. Email: heinrich@informatik.uni-hamburg.de

However, humans' natural language processing perhaps is the least well understood cognitive capability. The main reason for this may be the complexity of human language and our inability to observe and study this capability in less complex related species. Another reason is that the neural wiring in the human brain perhaps is not the only component, which is necessary for language to develop. It seems that socio-cultural principles are as well important, and only the inclusion of all factors may allow us to understand language processing. Nevertheless, it is our brain that enables humans to acquire perception capabilities, motor skills, language, and social cognition. The capability for *language acquisition* thus may result from the concurrence of general mechanisms on information processing in the brain's architecture. In particular, in recent studies in neuroscience it was found that the brain indeed includes both hemispheres and all modalities in language processing, and the embodied development of representations might be key in language acquisition (Barsalou, 2008; Glenberg and Gallese, 2012; Hickok and Poeppel, 2007; Huth et al., 2016; Pulvermüller and Fadiga, 2010). Furthermore, hierarchical dependencies in connectivity were identified, including different but specific delays in information processing. In linguistic accounts and behavioural studies a number of important principles, such as compositional *and* holistic properties in entities, body-rationality, and social interaction, have been found that might ease – or actually enable – the acquisition of a language competence (Karmiloff and Karmiloff-Smith, 2002; Smith and Kirby, 2012; Smith and Gasser, 2005). In light of the mechanistic conditions of the brain as well as enabling factors of how we learn language *and* other higher cognitive functions, the key objective is to understand the *characteristics* of a brain-inspired *appropriate* neural architecture that *facilitates* language acquisition.

In this paper, we propose[1] a novel embodied multi-modal model for language acquisition to study these important characteristics. As a significant innovation, this model grounds spoken language into the temporal dynamic processing of somatosensory and visual perception and explores a mechanism that abstracts latent representations from these dynamics in a self-organising fashion. Our contribution to knowledge is adding to the understanding of whether connectivity and plasticity attributes of the human brain allow for emergence and development of languages. Results from analytical as well as empirical studies with computer simulations and an interactive humanoid robot will reveal the importance of this self-organisation as well of specific timing in processing speech and multi-modal sensory information.

1.1. *Previous Work on Modelling Language Acquisition and Grounding*

In the past, researchers have suggested valuable models to explain grounding of language in embodied perception and action, based on neuroscientific data and hypotheses (compare (Cangelosi and Schlesinger, 2015; Coradeschi et al., 2013; Tani, 2014) for an overview). This includes early work on symbol grounding e.g. (Cangelosi, 2010; Cangelosi and Riga, 2006), studies on language evolution and symbol emergence, e.g. (Schulz et al., 2011; Steels et al., 2012), and research sentence comprehension and role filler assignment e.g. (Dominey et al., 2009; Dominey and Ramus, 2000). However, due to the tremendous complexity, models are rare which consider the dynamics in full scale and avoid assumptions on predefined word representations (short cutting language processing) or on static or categorically predefined observations (short cutting dynamics in grounding). From studies that approach this complexity, we can adopt important insights.

[1]Parts of this work have been presented at ICANN 2014 (Heinrich and Wermter, 2014).

1.1.1. Integrating Dynamic Vision

Models for grounding in *dynamic vision* are supposed to represent language in the alteration of e.g. perceived objects. Objects can, for example, be altered in terms of changing morphology or motion by self-induced manipulation. Due to complexity, models were often based on a certain decoupling and simplification of the visual stream to achieve a feasible level of coherence in visually perceived features. For example, Yu (2005) developed a model that coupled lexical acquisition with object categorisation. Here, the learning processes of visual categorisation and lexical acquisition were modelled in a close loop. This led to the emergence of the most important associations, but also to the development of links between words and categories and thus to linking similar fillers for a role. The visual features reflect little morphology over time since perception in the visual stream stemmed from unchanging preprocessed shapes in front of a plain background. With changing morphology, Monner and Reggia (2012) modelled grounding of language in visual object properties. This model is designed for a micro-language that stems from a small context-sensitive grammar and includes two input streams for scene and auditory information and an input-output stream for related prompts and responses. In between the input and input-output layer, several layers of LSTM blocks are employed to find statistical regularities in the data. This includes the overall meaning of a particular scene in terms of finding a latent symbol system that is inherent in the used grammar and dictionary. Yet, fed in object properties are – in principle – present as given prompts for the desired output responses. This way the emerging symbols in internal memory layers can be determined or shaped by the prompt and response data and are perhaps less latent. Thus it remains unclear how we can relate the emergence of pre-defined or latent symbols to the problem of grounding natural language in dynamic sensory information to eventually understand how noisy perceived information contributes.

Overall, the studies show that dynamic vision can be integrated as embodied sensation if the dynamics of perception can be reasonably abstracted. For a novel model, however, it is crucial to control complexity in perception to attempt explaining the emerging internal representation.

1.1.2. Dynamic Multi-modal Integration

Integrating multiple modalities into language acquisition is particularly difficult because the linked processes in the brain are extraordinary complex – and in fact – in large parts not yet understood. For this reason, to the best of the authors' knowledge, there is no model available that describes language processing integrated into multi-modal perception with full spatial and temporal resolution for the cortex without making difficult assumptions or explicit limitations. However, frameworks were studied that included temporally dynamic perception that forms the basis for grounding. Marocco et al. (2010) defined a controller for a simulated *Cognitive Universal Body* (iCub) robot based on *Recurrent Neural Networks* (RNNs). The iCub's neural architecture was trained to receive linguistic input (bit-strings representing pseudo-words) before the robot started to push an object (ball, cube, or cylinder) and observe the reaction in a sensorimotor way. Experiments showed that the robot was not only able to distinguish between objects via correct "linguistic" tags, but could reproduce a linguistic tag via observing the dynamics without receiving linguistic input and a correct object description. Thus, the authors claimed that the meaning of labels is not associated with a static representation of the object, but with its dynamical properties. Farkaš et al. (2012) modelled the grounding of words in both, object-directed actions and visual object sensations. In the model, motor sequences were learned by a continuous actor-critic learning that integrated the joint po-

sitions with linguistic input (processed in an *Echo State Network* (ESN)) and a visually perceived position of an object (learned a priori in a *Feed Forward Network* (FFN)). A specific strength of the approach is that the model, embedded into a simulated iCub, can adapt well to different motor constellations and can generalise to new permutations of actions and objects. However, the action, shape and colour descriptions (in binary form) are already present in the input of motor and vision networks. Thus, this information is inherently included in the filtered representations that are fed into the model part for a linguistic description. In addition, the linguistic network was designed as a fixed-point classifier that outputs two active neurons per input: one 'word' for an object and one for an action. Accordingly, the output is assuming a word representation and omits the sequential order. In a framework for multi-modal integration, Noda et al. (2014) suggested integrating visual and sensorimotor features in a deep auto-encoder. The employed time delay neural network can capture features on varying timespan by time-shifts and hence can abstract from higher level features to some degree. In their study, both modalities of features stem from the perception of interactions with some toys and form reasonable complex representations in the sequence of 30 frames. Although language grounding was not pursued, the shared multi-modal representation in the central layer of the network formed an abstraction of perceived scenes with a certain internal structuring and provided certain noise-robustness.

Nevertheless, all in all, we can obtain the insight that forming representation for language can perhaps get facilitated by the shared multi-modal representations and combinations of mechanisms of the brain that filter features on multiple levels.

1.2. *Paper Organisation*

This paper is structured as follows: With the related work in mind from the introduction, in Sec. 2 we will introduce important principles and mechanisms that have been found underlying language acquisition. In Sec. 3 we will develop a novel computational model by integrating these principles and mechanisms into a general recurrent architecture with aims at neurocognitive plausibility with respect to representation and temporal dynamic processing. We include a complete formalisation to ease re-implementation and will introduce a novel mechanism for unsupervised learning based on gradient descent. Then, in Sec. 4, we follow up with our evaluation and the analysis. We will specify the scenario for a language learning robot as well as a complete description of the utilised neurocognitively-inspired representations for verbal utterances and embodied multi-modal perception. In addition, we report experiments for generalisation and self-organisation. Finally, in Sec. 5 we will discuss our findings, conclusions, and future prospects.

2. Fundamental Aspects of Language Acquisition

Research on language acquisition is approached in different disciplines by complementary methods and research questions. Research in linguistics investigates different aspects of language in general and complexity of formal languages in particular. Ongoing debates about nature versus nurture and symbol grounding led to valuable knowledge of *principles* of learning and *mechanisms* of information fusion in the brain that facilitate language competence (compare Broz et al. (2014) and Cangelosi and Schlesinger (2015) for a roadmap). Recent research suggests that the well-known principles of statistical frequency and of compositionality in language acquisition are particularly important for forming representation by means of structuring multi-sensory data.

Researchers in different fields related to behavioural psychology study top-down both, the development of language competence in growing humans and the reciprocal effects of interaction with their environment, and have identified important socio-cultural principles. In computational neuroscience, many researchers look bottom-up into the *where* and *when* of language processing and refined the map of activity across the brain for language comprehension and production. New imaging methods allow for much more detailed studies on both, temporal and spatial level, and led to a major paradigm shift in our understanding of language acquisition and underlying mechanisms.

2.1. *Principles Found in Developmental Psychology*

For language acquisition, the first year after birth of a human infant is most crucial. In contrast to other mammals, the infant is not born mobile and matured, but develops capabilities and competencies postnatal (Karmiloff and Karmiloff-Smith, 2002). Development of linguistic competence occurs in parallel – and highly interwoven – with cognitive development of other capabilities such as multi-modal perception, attention, motion control, and reasoning, while the brain matures and wires various regions (Feldman, 2006; Karmiloff and Karmiloff-Smith, 2002). In this process of individual learning the infant undergoes several phases of linguistic comprehension and production competence, ranging from simple phonetic discrimination up to complex narrative skills (Grimm, 2012; Karmiloff and Karmiloff-Smith, 2002).

During this development the infant's cognitive system makes use the following crucial principles among others (Cangelosi and Schlesinger, 2015):

- **Preposition for reference.** The temporally coherent perception of a physical entity in the environment and a describing stream of spoken natural language leads to the association of both (Smith and Yu, 2008).
- **Body-rationality.** Representations, which an infant might form, develop through sensorimotor-level environmental interactions accompanied by goal-directed actions (Piaget, 1954). In addition, the embodiment is suggested as a necessary precondition for building up higher thoughts (Smith and Gasser, 2005).
- **Social cognition.** The development of language is seen only possible by interaction of a child with a caregiver that provides digestible amounts of spoken language (Tomasello, 2003). In particular, mothers provide an age-dependent simplification of grammar and focus on more common words first (Hayes and Ahrens, 1988).

Overall this means that postnatal development of the processes of thought together with an appropriate interaction of a teacher enables acquisition of language.

2.2. *Mechanistic Characteristics Found in Neuroscience*

Based on new imaging methods, several hypotheses have been introduced stating that many cortical areas are involved in language processing. In particular, it was claimed that several pathways between superior temporal gyrus and inferior frontal gyrus are involved in both language production and comprehension (Friederici, 2012; Hagoort and Levelt, 2009; Hickok and Poeppel, 2007; Huth et al., 2016). These pathways are suggested to include dorsal streams for sensorimotor integration and ventral streams for processing syntax and semantics. An important mechanism found is the activation of conceptual networks that are distributed over sensory areas during processing of words related to body parts (somatosensory areas) or object shapes (visual areas) (Pulvermüller, 2003;

Pulvermüller and Fadiga, 2010). Other seemingly important mechanisms found are:

- **Cell assemblies.** In higher stages of the spatial or temporal hierarchy, neurons are organised in cell assemblies (Damasio, 1989; Palm, 1990). Those might be distributed over different cortical areas or even across hemispheres and the activation of large and highly distributed cell assemblies can form higher-level concepts. Other cell assemblies exist that represent specific semantics like morphemes and lemmas in language processing or are mediators between different levels (Levelt, 2001). The aforementioned conceptual networks can be seen as cell assemblies on word (morpheme) level.
- **Phonological and lexical priming.** The structure of brain connectivity and timing leads to priming, for example in cohort activation of most relevant sounds or lemmas (Levelt et al., 1991; Marslen-Wilson and Zwitserlood, 1989).
- **Spatial and temporal hierarchical abstraction.** Strongly varying *timescales* take place in the brain. For example in the frontal lobe on caudal-rostral axis, processing of information occurs on much greater timescales from the pre-motor area up to mid-dorsolateral pre-frontal cortex, suggesting that these timings might be relevant for the processing of a plan for motor movement, over sequentialisation, and execution of motor primitives (Badre and D'Esposito, 2009; Badre et al., 2010). Similar temporal hierarchies have been found in lower auditory processing (Brosch and Schreiner, 1997; Ulanovsky et al., 2004) and higher vision (Schmolesky et al., 1998; Smith and Kohn, 2008).

Overall this indicates the tight involvement of *general* processes in the brain for reducing and representing complexity in language processing.

3. Neurocognitively Plausible Multi-modal Grounding Model

Based on aforementioned principles and mechanistic characteristics we can build up a model, which is a neurocognitively plausible *constraint* of a general nonlinear neural architecture. As a starting point we adopt the *Continuous Time Recurrent Neural Network* (CTRNN) as a valid abstraction for cortex-level processing (Dayan and Abbott, 2005):

$$\tau \frac{\mathrm{d}y_i}{\mathrm{d}t} = -y_i + f\left(\sum_{j \in I_{\mathsf{In}}} w_{ij} x_j + b_i + \sum_{k \in I_{\mathsf{Rec}}} w_{ik} y_k\right) \quad , \tag{1}$$

where the activity y of a neuron i is derived over time t as an accumulation of previous activity and a function over presynaptic input x (can be sensory input I_{In}, recurrent input I_{Rec}, or both), plastic connections w and a bias b. The derivation is governed by a time constant τ that describes how fast the firing rate approaches the steady state value. Although we can deduce the CTRNN from the *Leaky Integrate-and-Fire* (LIF) model and thus from a simplification of the Hodgkin-Huxley model from 1952, the network architecture was suggested independently by Hopfield and Tank (1986) as a nonlinear graded-response neural network and by Doya and Yoshizawa (1989) as an adaptive neural oscillator. The CTRNN is thus the most general computational network model as it allows us to define arbitrary input, output, or recurrence characteristics within one (horizontal) layer. Because of the recurrent connections, the network is arbitrarily deep[2] and nonlinear, based on continuous information that are processed over time.

[2]Compared to non-recurrent feed-forward networks, depth is depending on the arbitrary length of a sequence.

3.1. *Multiple Timescale Recurrent Neural Network*

To explore the mechanism of *timescales* as a constraint of the CTRNN, Tani et al. replicated the learning of mammal body motions in an experimental setup along the developmental robotic approach (Nishimoto and Tani, 2009; Tani et al., 2008; Yamashita and Tani, 2008). These *Multiple Timescale Recurrent Neural Networks* (MTRNNs) were specified by three layers (called *Input-Output* (IO) layer, *Context-fast* (Cf) layer, and *Context-slow* (Cs) layer) with variable timescales and have been trained with a gradient descent method for sequences. The analysis revealed that for a trained network, which could reproduce sequences best (merely indicated by converging to the smallest training error)[3], the patterns in different layers self-organised towards a decomposition of the body movements. The researchers were able to interpret from the neural activity that the Cf layer is always coding for the same short primitive, while the Cs layer patterns are unique per sequence and consist of slow changing values, which function as triggering points for primitives.

3.1.1. MTRNN with Context Bias

In those original experiments, the researchers were able to train an MTRNN for reasonably diverse and long sequences by initialising the network's neural activity at first time step with specific values of the experimenter's choice (Nishimoto and Tani, 2004; Yamashita and Tani, 2008). These *initial states* were kept for the training of each specific sequence and represented the (nonlinear) association of a constant (starting) value and its dynamic pattern. In later experiments, Nishide et al. adapted and integrated the idea of *Parametric Bias* (PB) units into the MTRNN (Awano et al., 2010; Nishide et al., 2009). Therein, bias units are part of the Cs layer and parametrise the motion sequence with a certain property (e.g. which tool is used in a certain action), while another initial neural activity is not specified. However, for these bias or *Context-controlling* (Csc) units only an initialisation before training is necessary, while the values of these units can self-organise during training. Similar to the *Recurrent Neural Network with Parametric Bias* (RNNPB), these initial states can be seen as a general context of a sequence. By modulating these internal states, differing other sequences can be generated. Overall, for the conducted experiments on motor primitives, the slow context codes for the general *concept* of a certain body motion.

By combining the characteristics of the various experiments on CTRNNs with multiple timescales and context bias properties (similar to parametric bias but also changing over time), we arrive at a general description of the MTRNN as illustrated in Fig. 1. For certain contexts, provided as initial states to some of the neurons with the highest timescale $I_{\mathsf{Csc}} \subset I_{\mathsf{Cs}}$ (slowest neurons), the network is processing certain sequences over time. The constraints on connectivity and relative timescale setting are inspired by the brain (Badre and D'Esposito, 2009) and have been challenged in developmental robotics studies to confirm a hierarchical compositionality e.g. in body motion. For further models, we can process dynamic sequences in terms of discretised time steps (e.g. for linguistic processing of smallest graphemic or phonetic units, or visual and sensorimotor processing with a certain sampling rate), but can regard any task as continuous by means of the absolute variability of the timescales.

[3]The best network during the experiments was shaped by timescale values of 1.0 for IO, 5.0 for Cf, and 70.0 for Cs layers (Yamashita and Tani, 2008).

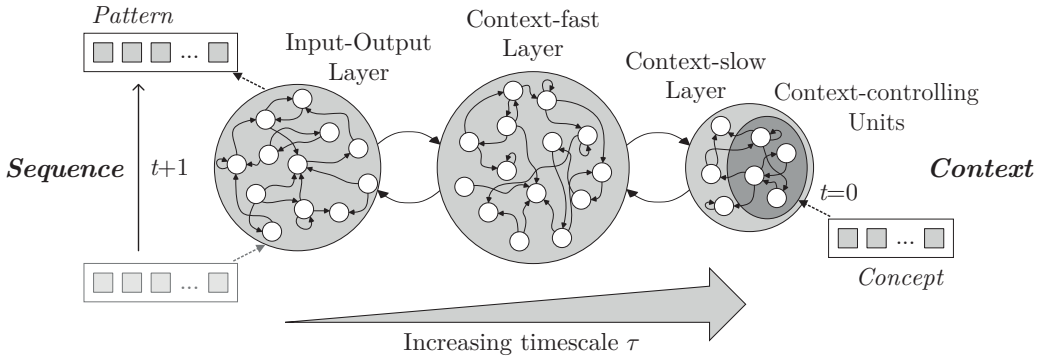


Figure 1. The overall *Multiple Timescale Recurrent Neural Network* (MTRNN) architecture with exemplary three horizontally parallel layers: *Input-Output* (IO), *Context-fast* (Cf), and *Context-slow* (Cs), with increasing timescale τ, where the Cs layer includes some *Context-controlling* (Csc) units. While the IO layer processes dynamic patterns over time, the Csc units at first time step ($t = 0$) contain the *context* of the sequence, where a certain concept can trigger the generation of the sequence.

3.1.2. Information Processing in the MTRNN

By defining the time constant as a neuron- or unit-dependent variable τ_i and solving the equation with respect to a time step t, we can also describe this special CTRNN in detail[4]: In the MTRNN information is processed continuously with a unit-specific firing rate as a sequence of T discrete time steps. Such a sequence $s \in S$ is represented as a flow of activations of neurons in the IO layer ($i \in I_{\mathsf{IO}}$). The input activation x of a neuron $i \in I_{\mathsf{All}} = I_{\mathsf{IO}} \cup I_{\mathsf{Cf}} \cup I_{\mathsf{Cs}}$ at time step t is calculated as:

$$x_{t,i} = \begin{cases} y_{t-1,i} & \text{iff } t \geq 1 \wedge i \notin I_{\mathsf{IO}} \\ x^*_{t,i} & \text{iff } t \geq 1 \wedge i \in I_{\mathsf{IO,input}} \\ y^*_{t-1,i} & \text{iff } t \geq 1 \wedge i \in I_{\mathsf{IO,output}} \end{cases} , \tag{2}$$

where we can either project desired (sensory) input x^* to the IO layer ($I_{\mathsf{IO,input}}$) or read out the desired output y^* of the IO layer ($I_{\mathsf{IO,output}}$), depending on how the architecture is employed in a task. The input activation for neurons $i \notin I_{\mathsf{IO,input}}$ is initialised with 0 at the beginning of the sequence. The internal state z of a neuron i at time step t is determined by:

$$z_{t,i} = \begin{cases} c_{0,i} & \text{iff } t = 0 \wedge i \in I_{\mathsf{Csc}} \\ \left(1 - \frac{1}{\tau_i}\right) z_{t-1,i} + \frac{1}{\tau_i} \left(\sum_{j \in I_{\mathsf{All}}} w_{ij} x_{t,j} + b_i \right) & \text{otherwise} \end{cases} , \tag{3}$$

where $c_{0,i}$ is the initial internal state of the Csc units $i \in I_{\mathsf{Csc}} \subset I_{\mathsf{Cs}}$ (at time step 0), $w_{i,j}$ are the weights from jth to ith neuron, and b_i is the bias of neuron i. The output (activation value) y of a neuron i at time step t is defined by an arbitrary differentiable activation function:

$$y_{t,i} = f\left(z_{t,i}\right) \quad , \tag{4}$$

depending on the representation for neurons in IO and on the desired shape of the activation for postsynaptic neurons, e.g. decisive normalisation (softmax) or sigmoidal.

[4]Notation style follows the original description of the MTRNN in (Yamashita and Tani, 2008).

3.1.3. Learning in the MTRNN

During learning the MTRNN can be trained with sequences, and self-organises the weights and also the internal state values of the Csc units. The overall method can be a variant of *Backpropagation Through Time* (BPTT), sped up with appropriate measures based on the task characteristics.

For instance, if the MTRNN produces continuous activity (IO) we can modify the input activation with a prorated *Teacher Forcing* (TF) signal $\alpha \in]0,1[$ of the desired output y^* together with the generated output y of the last time step:

$$x_{t,i} = \begin{cases} (\alpha) y^*_{t-1,i} + (1-\alpha) y_{t-1,i} & \text{iff } t \geq 1 \wedge i \in I_{\mathsf{IO}} \\ y_{t-1,i} & \text{iff } t \geq 1 \wedge i \notin I_{\mathsf{IO}} \end{cases} \quad . \tag{5}$$

In the forward pass, an appropriate error function E is accumulating the error between activation values (y) and desired activation values (y^*) of IO neurons at every time step based on the utilised activation function. In the second step, the partial derivatives of calculated activation (y) and desired activation (y^*) are derived in a backward pass. In the case of e.g. a decisive normalisation function (softmax) in IO and a sigmoidal function f_{sig} in all other layers, we can specify the error on the internal states of all neurons as follows:

$$\frac{\partial E}{\partial z_{t,i}} = \begin{cases} y_{t,i} - y^*_{t,i} + \left(1 - \frac{1}{\tau_i}\right) \frac{\partial E}{\partial z_{t+1,i}} & \text{iff } i \in I_{\mathsf{IO}} \\ \sum\limits_{k \in I_{\mathsf{All}}} \frac{w_{ki}}{\tau_k} \frac{\partial E}{\partial z_{t+1,k}} f'_{\mathsf{sig}}(z_{t,i}) + \left(1 - \frac{1}{\tau_i}\right) \frac{\partial E}{\partial z_{t+1,i}} & \text{otherwise} \end{cases} \quad , \tag{6}$$

where the gradients are 0 for the time step $T+1$. For the error function E of the decisive normalisation the *Kullback-Leibler Divergence* (KLD) is used, where the cross-entropy is generalised to $|I_{\mathsf{IO}}|$ classes (Kullback and Leibler, 1951). Importantly, the error propagated back from future time steps is particularly dependent on the (different) timescales.

Finally, in every epoch n the weights w but also biases b are updated:

$$w_{ij}^{(n)} = w_{ij}^{(n-1)} - \eta_{ij} \frac{\partial E}{\partial w_{ij}} = w_{ij} - \eta_{ij} \sum_t \frac{1}{\tau_i} \frac{\partial E}{\partial z_{t,i}} x_{t,j} \quad , \tag{7}$$

$$b_i^{(n)} = b_i^{(n-1)} - \beta_i \frac{\partial E}{\partial b_i} = b_i - \beta_i \sum_t \frac{1}{\tau_i} \frac{\partial E}{\partial z_{t,i}} \quad , \tag{8}$$

where the partial derivatives for w and b are respectively the accumulated sums of weight and bias changes over the whole sequence, and η and β denote the learning rates for weight and bias changes. To facilitate the application of different methods for speeding up training, we can use individual learning rates for all weights and biases to allow for individual modifications of the weight and bias updates respectively.

The initial internal states $c_{0,i}$ of the Csc units define the behaviour of the network and are also updated as follows:

$$c_{0,i}^{(n)} = c_{0,i}^{(n-1)} - \zeta_i \frac{\partial E}{\partial c_{0,i}} = c_{0,i} - \zeta_i \frac{1}{\tau_i} \frac{\partial E}{\partial z_{0,i}} \quad \text{iff } i \in I_{\mathsf{Csc}} \quad , \tag{9}$$

where ζ_i denotes the learning rates for the initial internal state changes.

3.1.4. Adaptive Learning Rates

For speeding up training we employ an adaptation of the *Resilient Propagation* (RPROP) algorithm that makes use of different individual learning rates η and β and adapt the learning rates ζ for the update of the initial internal states $c_{0,i}$ as well (Heinrich et al., 2012; Riedmiller and Braun, 1993). In particular, the learning rates ζ are adapted proportionally to the average of all learning rates η over all weights that are connected with unit i and neurons of the same (Cs) and adjacent (Cf) layer:

$$\zeta_i \propto \frac{1}{|I_{\mathsf{Cf}}| + |I_{\mathsf{Cs}}|} \sum_{j \in (I_{\mathsf{Cf}} \cup I_{\mathsf{Cs}})} \eta_{ij} \quad . \tag{10}$$

Since the update of the $c_{0,i}$ depends on the same partial derivatives (time step $t = 0$) as the weights, we do not need additional parameters in this adaptive mechanism.

3.2. *Novel Unsupervised MTRNN with Context Abstraction*

In the MTRNN with context bias we found that the timescale characteristic is crucial for a hierarchical compositionality of temporal dynamic output sequences. The increasingly slower information processing in the context led to generation of a sequence from an abstract concept. In order to design an architecture that can abstract a context from temporal dynamic input sequences, we can reverse the notion of the context bias and thus reverse the processing from the context to the IO layer. The structure of such a novel MTRNN with context abstraction is visualised in Fig. 2. For certain sequential input, provided as a dynamic pattern to the fastest neurons (lowest timescale) I_{IO}, the network is accumulating a common *concept* in the slowest neurons (highest timescale) $I_{\mathsf{Csc}} \in I_{\mathsf{Cs}}$. Since the timescale characteristics yield a slow adaptation of these so-called Csc units, information in the units will accumulate abstract pattern from the input sequence (filtered by neurons in a potential intermediate layer). The accumulation of information is characterised by a logarithmic skew to the near past and a reach-out to the long past depending on timescale values τ_{Cs} (and τ_{Cf}).

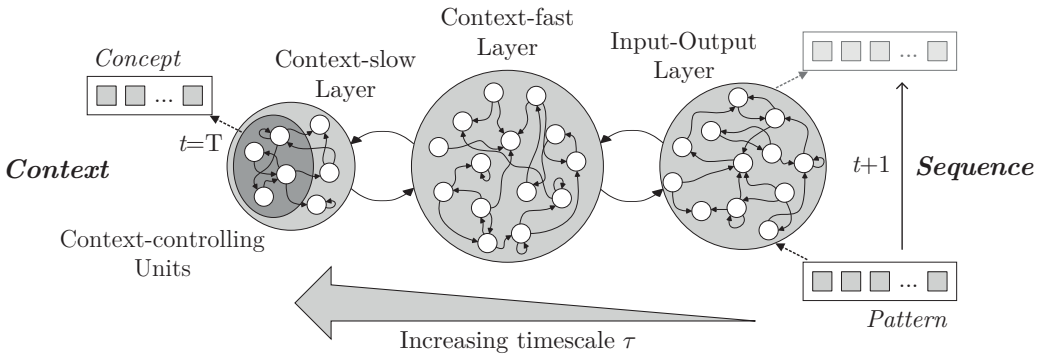


Figure 2. The *Multiple Timescale Recurrent Neural Network* (MTRNN) with context abstraction architecture providing exemplary three horizontally parallel layers: *Context-slow* (Cs), *Context-fast* (Cf), and *Input-Output* (IO), with increasing timescale τ, where the Cs layer includes some *Context-controlling* (Csc) units. While the IO layer processes dynamic patterns over time, the Csc units abstract the context of the sequence at *last* time step ($t = T$). The crucial difference to the MTRNN with context bias is an inversion of the direction of procession and an accumulation of abstract context instead of production from a given abstract context.

3.2.1. From Supervised Learning to Self-organisation

The MTRNN with context abstraction can be trained in supervised fashion to capture a certain concept from the temporal dynamic pattern. This is directly analogue to fixed-point classification with *Elman Recurrent Neural Networks* (ERNNs) Elman (1989) or CTRNNs: We can determine the error between a desired temporal static concept pattern and the activity in the Csc units at final time step ($t = T$). With a gradient descent method we can propagate the error backwards through time over the whole temporal dynamic pattern from which the concept was abstracted. However, for an architecture that is supposed to model the processing of a certain cognitive function in the brain, we are also interested in removing the necessity of providing a *desired* target concept a priori. Instead, the representation of the concept should *self-organise* based on regularities latent in the stimuli.

For the MTRNN with parametric bias, this was realised in terms of modifying the Csc units' activity in the initial time step ($t = 0$) backwards by the partial derivatives for weights connecting from those units. Thus the internal states of the initial Csc units self-organised in Csc space towards values that were suited best for generating the sequences of the data set (Hinoshita et al., 2011). To foster a similar self-organisation of the Csc units at final time step of the MTRNN with context abstraction, a semi-supervised mechanism is developed that allows us to modify the desired concept pattern based on the derived error.

Since we aim at an abstraction from perception input to the overall concept, the *Least Mean Square* (LMS) error function is modified for the internal state z at time step t of neurons $i \in I_{\mathsf{All}} = I_{\mathsf{IO}} \cup I_{\mathsf{Cf}} \cup I_{\mathsf{Cs}}$, introducing a *self-organisation forcing constant* ψ as follows:

$$\frac{\partial E}{\partial z_{t,i}} = \begin{cases} (1-\psi)\left(y_{t,i} - f(c_{T,i} + b_i)\right) f'_{\mathsf{sig}}\left(z_{t,i}\right) & \text{iff } i \in I_{\mathsf{Csc}} \wedge t = T \\ \sum\limits_{k \in I_{\mathsf{All}}} \dfrac{w_{ki}}{\tau_k} \dfrac{\partial E}{\partial z_{t+1,k}} f'(z_{t,i}) + \left(1 - \dfrac{1}{\tau_i}\right) \dfrac{\partial E}{\partial z_{t+1,i}} & \text{otherwise} \end{cases} , \tag{11}$$

where $c_{T,i}$ are internal states at the *final* time step T (indicating the last time step of a sequence) of the Csc units $i \in I_{\mathsf{Csc}} \subset I_{\mathsf{Cs}}$.

The particularly small self-organisation forcing constant 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:

$$c_{T,i}^{(n)} = c_{T,i}^{(n-1)} - \psi\zeta_i \frac{\partial E}{\partial c_{T,i}} = c_{T,i}^{(n-1)} - \psi\zeta_i \frac{1}{\tau_i} \frac{\partial E}{\partial z_{T,i}} \quad \text{iff } i \in I_{\mathsf{Csc}} \quad , \tag{12}$$

where ζ_i denotes the learning rates for the final internal state changes. Thereby the learning error E is used in one part (ψ) to modify the final internal states and in another part ($1 - \psi$) to modify the weights.

Thus, similarly to the parametric bias units, the final internal states $c_{T,i}$ of the Csc units self-organise during training in conjunction with the weights (and biases) towards the highest entropy. We can observe that the self-organisation forcing constant and the learning rate are dependent, since changing ζ would also shift the self-organisation – for arbitrary but fixed ψ. However, this is an useful mechanism to self-organise towards concepts that are most appropriate with respect to the structure of the data.

3.2.2. Preliminarily Evaluating the Abstracted Context

To test in a preliminary experiment how the abstracted concepts form for different sequences using this unsupervised learning mechanism, the architecture was trained for abstracting two contrary cosine waves into context patterns. In particular, for a sequence two cosines waves were presented to two input neurons and discretised to 33 time step. By differently phase-shifting the cosines, four different sequences were prepared. The key aspect of this task is to learn abstract the different phase-shifts in the otherwise identical sequences. In particular because of the ambiguous nature of saddle points, the network cannot simply learn to predict the next time step, but must capture the whole sequence. Processing such a sequence by the MTRNN with context abstraction is supposed to result in a specific pattern of the final Csc units' activity as the abstracted concept.

For determining how those patterns self-organise, the architecture was trained with predefined unchanging patterns (chosen randomly: $\forall i \in I_{\mathsf{Csc}}, c_{T,i} \in \mathbb{R}_{[-1.0,1.0]}$) as well as with randomly initialised patterns that adapt during training by means of the varied self-organisation forcing parameter ψ. To measure the result of the self-organisation, two distance measures d_{avg}, and d_{rel} are used:

$$d(c_k, c_l) = \sqrt{\sum_{i \in I_{\mathsf{Csc}}} (c_{k,i} - c_{l,i})^2} \quad , \tag{13}$$

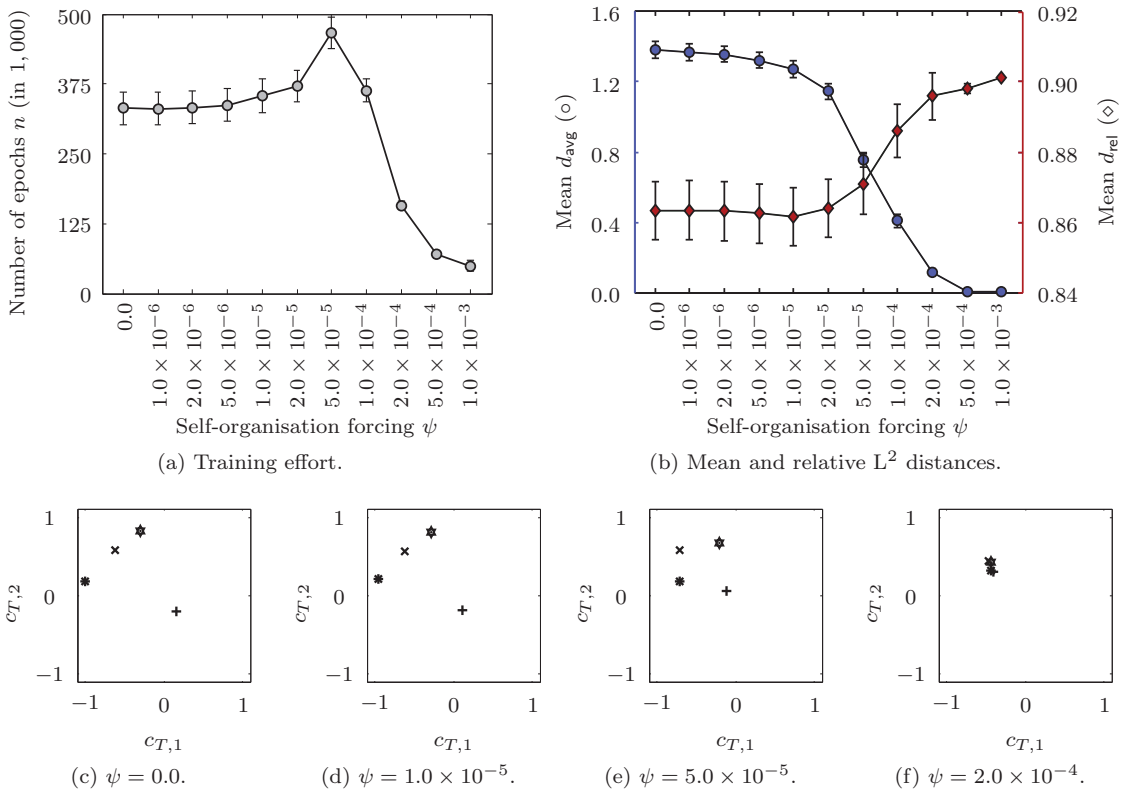
$$d_{\mathsf{avg}} = \frac{1}{(|S|-1) \cdot (|S|/2)} \sum_{k=1}^{|S|-1} \sum_{l=k+1}^{|S|} d(c_k, c_l) \quad , \tag{14}$$

$$d_{\mathsf{rel}} = \prod_{k=1}^{|S|-1} \prod_{l=k+1}^{|S|} \left(\frac{d(c_k, c_l)}{d_{\mathsf{avg}}} \right)^{\frac{1}{(|S|-1)\cdot(|S|/2)}} \quad , \tag{15}$$

where $|S|$ describes the number of sequences and $c_k = c_{k,T,i}$ denotes the final Csc units of sequence k. With d_{avg}, which uses the standard *Lebesgue* L^2 or *Euclidean* distance, we can estimate the average distance of all patterns, while with d_{rel} we can describe the relative difference of distances. For example, in case the distances between all patterns are exactly the same, this measure would yield the best possible result[5] of $d_{\mathsf{rel}} = 1.0$. Comparing both measures for varied settings of ψ provides an insight on how well the internal representation is distributed after self-organisation.

The results for the experiment are presented in Fig. 3. From the plots, we can obtain that patterns of the abstracted context show a fair distribution for no self-organisation (the random initialisation) up to especially small values of about $\psi = 1.0 \times 10^{-5}$, a good distribution for values around $\psi = 5.0 \times 10^{-5}$ and a degrading distribution for larger ψ. The scatter plots for arbitrary but representative runs in Fig. 3c–f visualise the resulting patterns for no ($\psi = 0.0$), too small ($\psi = 1.0 \times 10^{-5}$), good ($\psi = 5.0 \times 10^{-5}$), and too large self-organisation forcing ($\psi = 2.0 \times 10^{-4}$). From inspecting the Csc units, we can learn that a "good" value for ψ leads to a marginal self-organisation towards an ideal distribution of the concepts over the Csc space during the training of the weights. Furthermore, a larger ψ is driving a stronger adaptation of the Csc patterns than of the weights, thus leading to a convergence to similar patterns for all sequences.

[5]Given the dimensionality of the Csc units is ideal with respect to the number of sequences. However, in cases of a dimensionality that is lower, this value is smaller 1.0. For example, it is not possible to arrange four points in a 2D-plane with equal distance > 0. In this case, when representing four sequences with two Csc units, we can derive a theoretical optimal $d_{\mathsf{rel}} = 0.9863$. This example can be visualised as having four points in a the 2D-plane arranged optimally on the edges of a square.



(a) Training effort.

(b) Mean and relative L^2 distances.

(c) $\psi = 0.0$.

(d) $\psi = 1.0 \times 10^{-5}$.

(e) $\psi = 5.0 \times 10^{-5}$.

(f) $\psi = 2.0 \times 10^{-4}$.

Figure 3. Effect of the self-organisation forcing mechanism on the development of distinct concept patterns for different sequences of contrary cosine waves: Training effort (a) and mean d_{avg} and d_{rel} with standard error bars over varied ψ (b), each over 100 runs; representative developed Csc patterns (c–f) for different sequences for selected parameter settings of no, small, "good", and large self-organisation forcing respectively.

The task in this preliminary experiment is quite simple, thus a random initialisation within a feasible range of values ($[-1.0, 1.0]$) of the Csc units often already provides a fair representation of the context and allows for convergence to very small error values. However, for larger numbers of sequences, which potentially share some primitives, the random distribution of respective concept abstraction values is unlikely to provide a good distribution, thus self-organisation forcing mechanism can drive the learning.

3.3. *Previous MTRNN models for Language Processing*

In previous studies, the MTRNN with context bias was tested for modelling language processing due to the mechanism of spatial and temporal hierarchical compositionality. In particular, Hinoshita et al. (2011) utilised the architecture in a model to learn language from continuous input of sentences composed of words and graphemes that stem from a small grammar. For the model, no implicit information is provided on word segmentation and on roles or categories for words. Instead, the input is modelled as streams of spike-like activities on graphemic level. During training, the architecture self-organises to the decomposition of the sentences hierarchically, based on the explicit structure of the inputs and the specific characteristics of some layers. The authors found that the characteristics of information processing on different timescales indeed leads to a hierarchical decomposition of the sentences in a way that certain character orders form words and certain word orders form the sentences. Although the model was reproducing learned symbolic sentences quite well in their experiments, generalisation was not possible to

test, because the generation of sentences was initiated by the internal state of the Csc units, which had to be trained individually for every sentence in the model.

Heinrich et al. (2015) extended this model to process the language embodied in a way that visual input will trigger the model to produce a meaningful verbal utterance that appropriately represents the input. The architecture, called EMBMTRNN model, consists of similar MTRNN layers for the language network, where a verbal utterance is processed as a sequence on phoneme level based on initial activity on an overall concept level. The overall concept is associated with raw feature input over merged shape and colour information of a visually perceived object. Thereby the model incorporates the following hypotheses: a) speech is processed on a multiple-time resolution, and b) semantic circuits are involved in the processing of language. Experiments revealed that the model can generalise to new situations, e.g. describe an object with a novel combination of shape and colour with the correct corresponding utterance due to the appropriate hierarchical component structure. Yet, in this model the multi-modal complexity of real-world scenarios has not yet been tackled exhaustively. The temporal dynamic nature of visual observations or sensations from another modality was not included and especially not processed on multiple-time resolution.

3.4. *Novel Recurrent Neural Model with Embodied Multi-modal Integration*

Previous models of language processing (compare in Sec. 3.3) provided insight into the architectural characteristics of language production, grounded in *some* perception. In recent neuroscientific studies, we learned about the importance of *conceptual networks* that are activated in processing speech and that most of the involved processes operate in producing speech as well (compare Borghi et al. (2010); Glenberg and Gallese (2012); Indefrey and Levelt (2004); Levelt (2001); Pulvermüller et al. (2014)). Central findings include that the sensorimotor system is involved in these conceptual networks in general and in action and language comprehension in particular.

For the action comprehension phenomenon, these networks supposedly seem to involve multiple senses. As an example, for actions perceived from visual stimuli, Singer and Sheinberg (2010) found that there is a tight connection between perceiving the form and the motion of an action. A sequence of body poses is perceived as an action if the frames are integrated within 120 milliseconds. Additionally, they found that the visual sequence is represented best as an action if both cues are present, but that in such a case the representation is mostly based on form information. Since *body-rational* motion information is hierarchically processed in proprioception as well, an integration of visual form and somatosensory motion seems more important. These multi-modal contributions – visual and somatosensory – are suggested to be strictly hierarchically organised (compare Friston (2005); Sporns et al. (2004)).

The structure of integration in a conceptual network seems to derive from spatial conditions of the areas on the cortex that have been identified for higher abstraction from the sensory stimuli. These areas, for example the *Superior Temporal Gyrus* (STG), but also the *Inferior Frontal Gyrus* (IFG), are connected more densely, compared to the sensory regions, but they also show a high interconnectivity with other areas of higher abstraction. From the studies on *Cell Assemblies* (CAs) we deduced that such a particularly dense connectivity, on the one hand, can form general concepts (for example about a certain situated action) and, on the other hand, may invoke activation first (Pulvermüller et al., 2014).

3.4.1. Model Requirements

From these recent findings, hypotheses, and the previous related work, we can adopt that the computational neural model for natural language production should be embedded in an architecture that integrates multiple modalities of contributing perceptual (sensory) information. The perceptual input should also be processed horizontally from sensation encoding over primitive identification (if compositional) up to the conceptual level. Highly interconnected neurons between higher conceptual areas should form CAs and thus share the representations for the made experiences. Importantly the representations should form based on the structure in the perceptual input without a specific target.

In line with the developmental robotics approach (Cangelosi and Schlesinger, 2015), the multi-modal perception should be based on real world data. Both, the perceptual sensation as well as the auditory[6] production, should be represented neurocognitively plausible. By employing this approach, an embodied and situated agent should be created that acquires a language by interaction with its environment as well as a verbally describing teacher. In this case, the interaction is experienced in terms of the temporal dynamic manipulation of different shaped and coloured objects.

With these requirements, the model implements the principles and mechanistic characteristics described in Sec. 2.1. Properties of the model supposedly are generalisation despite dynamic embodied perception and disambiguation of inherently focused but limited uni-modal sensation by multi-modal integration. All in all, goals of this model are a) to refine connectivity characteristics that foster language acquisition and b) to investigate merged conceptual representation.

3.4.2. Multi-modal MTRNNs Model

In order to meet the requirements of such a multi-modal model, the following hypotheses are added to the previous EMBMTRNN model (compare Sec. 3.3) into a novel model named MULTIMTRNNS: a) somatosensation and visual sensation are processed hierarchically by means of a multiple-time resolution, and b) higher levels of abstractions are encoded in CAs that are distributed over the sensory and motor areas. As a specific refinement of the previous model, the neural circuits for processing the perceptions are modelled each as an MTRNN with context abstraction. The first one, called $\text{MTRNN}_{\mathtt{s}}$, processes somatosensation, specifically proprioceptive perception, while the second one, named $\text{MTRNN}_{\mathtt{v}}$, processes visual perception. The Csc units of all MTRNNs (within the layers with the highest timescale Cs) are linked as fully connected associator neurons that constitute the CAs for representing the concepts of the information. Based on the abstract concepts the MTRNN with context bias, here called $\text{MTRNN}_{\mathtt{a}}$, processes the verbal utterance, again as a sequence on phoneme level. All recurrent neural structures are specifications of a CTRNN to maintain neurocognitive plausibility.

The notation of the IO, Cf, and Cs layers in the novel perception components of the MULTIMTRNNS model, stand for input, fusion, and context of both modalities, somatosensory and vision, respectively. An overview of the architecture is presented in Fig. 4. An arising hypothesis for the computational model is that during learning a composition of a general feature emerges, which is invariant to the length of the respective sensory input. A second hypothesis is that features are ambiguous if uni-modal sensations are ambiguous for a number of overall *different* observations, but that the association can provide a distinct representation for the production of a verbal utterance.

[6]For consistency with some related literature the speech production, which adopts the involvement of the auditory system, is called auditory production.

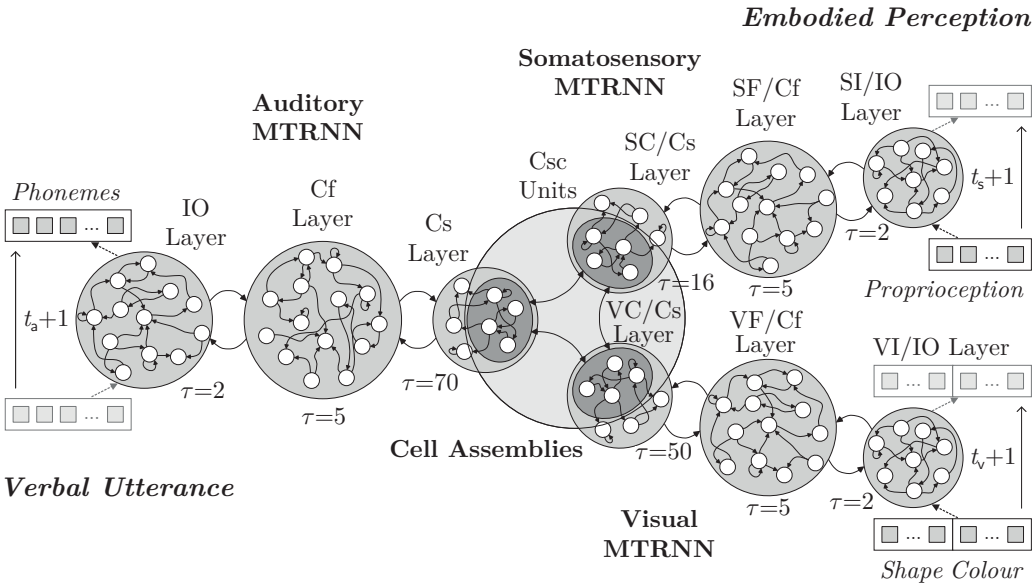


Figure 4. Architecture of the multi-modal MTRNN model, consisting of an MTRNN with context bias for auditory, two MTRNNs with context abstraction for somatosensory as well as visual information processing, and *Cell Assemblies* (CAs) for representing and processing the concepts. A sequence of phonemes (utterance) is produced over time, based on sequences of embodied multi-modal perception.

3.4.3. Information Processing, Training, and Production

For every scene, verbal utterances are presented together with sequences of proprioceptive and visual stimuli of an action sequence. During training of the system, the somatosensory MTRNN_s and the visual MTRNN_v self-organise weights and also the internal states of the Csc units in parallel, for processing of an incoming perception. For the production of utterances, the auditory MTRNN_a self-organises weights and also the internal states of Csc units. The important difference is that the MTRNN_s and the MTRNN_v self-organise towards the *final* internal states of the Csc (end of perception), while the MTRNN_a self-organises towards the *initial* internal states of the Csc (start of utterance). Finally, the activity of the Csc units of all MTRNNs gets associated in the CAs. The output layers of the MTRNN_a are specified by the decisive normalisation, while all other neurons are set up with a sigmoidal function (using a logistic function with $\kappa_\mathsf{h} = 0.35795$ for range, and $\kappa_\mathsf{w} = 0.92$ for slope as suggested in (Heinrich and Wermter, 2014). This particularly includes the neurons in the IO layers of the MTRNN_s and MTRNN_a as well.

For training of the auditory MTRNN_a the procedure and mechanisms are kept identical to the training in all previous models: the adaptive BPTT variant is utilised by specifying the KLD and the LMS as the respective error functions. The training of the MTRNN_s and MTRNN_v is conducted similarly, but it includes for both the suggested self-organisation forcing mechanism as described in Eq. 11 (Sec. 3.2.1). For these MTRNN with context abstraction, again the error is measured on randomly initialised (desired) activities of the Csc units at the final time step and is used for self-organising both, the weights and the desired internal Csc states. For the CAs, associations between the Csc units of the MTRNN_s, MTRNN_v, and MTRNN_a are trained with the LMS rule on the activity of

the Csc units:

$$\frac{\partial E}{\partial z_i} = (y_i - f_{\mathsf{sig}}(c_{\mathsf{a},0,i}))f'_{\mathsf{sig}}(z_i) \quad , \tag{16}$$

$$z_i = \sum_{j \in I_{\mathsf{s},\mathsf{Csc}}} w_{ij} f_{\mathsf{sig}}(c_{\mathsf{s},T,j}) + \sum_{k \in I_{\mathsf{v},\mathsf{Csc}}} w_{ik} f_{\mathsf{sig}}(c_{\mathsf{v},T,k}) + b_i \quad \forall i \in I_{\mathsf{a},\mathsf{Csc}} \quad , \tag{17}$$

where $c_{\mathsf{a},0,i}$, $c_{\mathsf{s},T,j}$ and $c_{\mathsf{v},T,k}$ denote the internal states of the Csc units for the MTRNN$_\mathsf{a}$, MTRNN$_\mathsf{s}$ and MTRNN$_\mathsf{v}$ respectively.

With a trained network the generation of novel verbal utterances from proprioception and visual input can be tested. The final Csc values of the MTRNN$_\mathsf{s}$ and MTRNN$_\mathsf{v}$ are abstracted from the input sequences respectively and associated with initial Csc values of the auditory MTRNN$_\mathsf{a}$. These values, in turn, initiate the generation of a phoneme sequence. 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 and Results

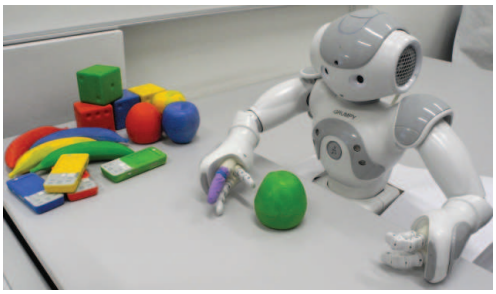
In order to analyse the proposed model's characteristics, we are first of all interested in identifying a parameter setting for the best (relative) generalisation capabilities. Particularly, this enables to analyse the information patterns that emerges for different parts of the architecture. Inspired by infant learning such an analysis will be embedded in a real world scenario, where a robot learns language from interaction with a teacher and its environment (Cangelosi and Schlesinger, 2015). As a prelude for such an analysis the self-organisation forcing mechanisms need to be inspected further for the impact on the developed internal representation of the abstracted proprioception.

4.1. *Multi-modal Language Acquisition Scenario*

Premised on the principle of social cognition (compare Sec. 2.1), the scenario is also based in the interaction of a human teacher with a robotic learner to acquire and ground language in embodied and situated experience. For testing the refined model, a *NAO humanoid robot* (NAO) is supposed to learn to describe the manipulation of objects with various characteristics to be able to describe novel manipulation actions with correct novel verbal utterances. Manipulations are to be done by the NAO's effectors and thus to be observed by its motor feedback (proprioception) and visual perception (see Fig. 5a for an overview). In this study, for the developmental robotics approach, it is particularly important to include the influence of natural variances in interaction, which originate in varying affordances of different objects, but also in unforeseen natural noise.

For a given scene in this scenario, the human teacher guides the robot's arm in an interaction with a coloured object and verbally describes the manipulation action, e.g. '`slide the red apple`'. Later, the robot should be able to describe a new interaction composed of motor movements (proprioception) 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 analysing how the permutation is handled. For this reason, the corpus is limited to a set of verbal utterances, which are generated from the small grammar as summarised in Fig. 5c. For every single object of



(a) Scenario overview.

pull	move the arm behind the object and drag it towards the torso
push	move the arm in front of the object and push it away from the torso
show me	point with the hand to the object
slide	move the arm to the right of the object and slide it horizontally to the left

(b) Instructions for manipulation action teaching.

Example:
'slide the red apple.'

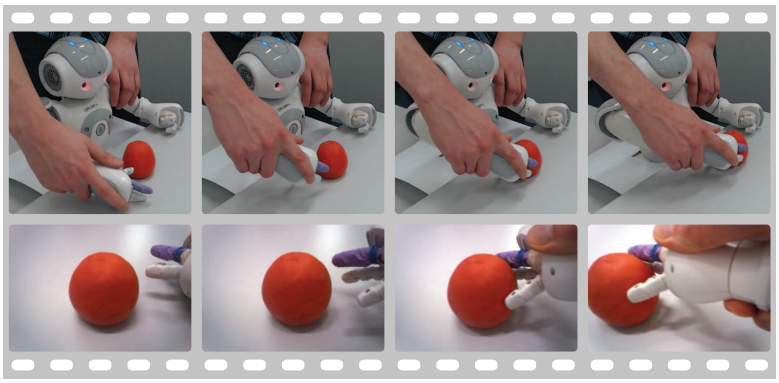
```
S → ACT the COL OBJ.

ACT → pull | push
      | show me | slide

COL → blue | green
      | red | yellow

OBJ → apple | banana
      | dice | phone
```

(c) Grammar.



(d) Action teaching over time (bottom: learner's view).

Figure 5. Scenario and manipulation action recording for multi-modal language learning scenario.

the same four distinct shapes (apple, banana, phone, or dice) and four colours (blue, green, red, or yellow), four different manipulations are feasible with the arm of the NAO: pull, push, show me, and slide. The grammar is overall unambiguous, meaning that a specific scene can only be described by one specific utterance. Nevertheless, all objects have a similar mass and similar surface conditions (friction). This way the proprioceptive sensation alone is mostly ambiguous for a certain manipulation action on objects with differing colours, but also with different shapes.

In order to collect data for this study, the 64 different possible interactions were recorded four times, each with the same verbal utterance and arm-starting position but with slightly varying movements and object placements. This was done by asking different subjects (colleagues from the computer science department) to perform the teaching of such interactions in order to minimise the experimenter's bias (instructions listed in Fig. 5b).

Neurocognitively Plausible Encoding

To encode an utterance into a sequence $s = (p_1, \ldots, p_T)$ of neural activation over time, a phoneme-based adaptation of the encoding scheme suggested by Hinoshita et al. (2011) is used: All verbal utterances for the descriptions are taken from the symbolic grammar, but are transformed into phonetic utterances based on phonemes from the ARPAbet[7] and four additional signs to express pauses and intonations in propositions, exclamations, and questions: $B = \{\text{'AA'}, ..., \text{'ZH'}\} \cup \{\text{'SIL'}, \text{'PER'}, \text{'EXM'}, \text{'QUM'}\}$,

[7]ARPAbet is a general American English phone set, transcribed in ASCII symbols that was developed in the 1976 Speech Understanding Project by the Advanced Research Projects Agency. Transformation was done using the *Carnegie Mellon University* (CMU) pronouncing dictionary.

with size $|B| = 44$. The occurrence of a phoneme p_k is represented by a spike-like neural activity of a specific neuron at relative time step t_{rel}. In addition, some activity is spread backwards in time (rising phase) and forwards in time (falling phase), represented as a Gaußian function g over the interval $[-\omega/2, \ldots, -1, 0, +1, \ldots, \omega/2]$. All activities of spike-like peaks are normalised by a decisive normalisation function for every absolute time step t over the set of input neurons. On the absolute course of time t the peaks mimic priming effects in articulatory phonetic processing. For example, the previous occurrence of the phoneme '`P`' could be related to the occurrence of the phoneme '`AH`' leading to an excitation of the respective neuron for '`AH`', when the neuron for '`P`' was activated. A sketch of the utterance encoding is shown in Fig. 6. The Gaußian g for p_k is defined by:

$$g(p_k, t_{\mathsf{rel}}, i) = \begin{cases} \exp\left(\dfrac{-t_{\mathsf{rel}}^2}{2\sigma^2}\right) & \text{iff } p_k = B_i \\ 0 & \text{otherwise} \end{cases} \quad , \tag{18}$$

where $t_{\mathsf{rel}} = 0$ is the mean and the variance σ represents the filter sharpness factor. A peak occurs for the neuron $i \in I_{\mathsf{IO}}$ with $|I_{\mathsf{IO}}| = |B|$, if the phoneme p_k is equal to the ith phoneme in the phoneme alphabet B. From the spike-like activities the internal state z of a neuron i at time step t is determined by:

$$z_{t,i} = \begin{cases} \lambda \cdot \max\left(g(p_{k=1\ldots|s|}, t_{\mathsf{rel}} = -\omega/2 \ldots \omega/2, i)\right) & \text{iff } t = \gamma + k\nu + t_{\mathsf{rel}} \\ 0 & \text{otherwise} \end{cases} \quad , \tag{19}$$

$$\lambda = \ln\left(\frac{0.9}{1.0 - 0.9}\left(|I_{\mathsf{IO}}| - 1\right)\right) \quad , \tag{20}$$

where ω is the filter width, γ is a head margin to put some noise to the start of the sequence, ν is the interval between two phonemes, and λ is a scaling factor for the neuron's activity y^* of maximal values for possibly overlapping spikes. The scaling factor depends on the number of IO neurons and scales the activity to $y^* \in]0, 0.9]$ for the specified decisive normalisation (softmax) function:

$$y_{t,i}^* = f_{\mathsf{softmax}}(z_{t,i}) = \frac{\exp\left(z_{t,i}\right)}{\sum\limits_{j \in I_{\mathsf{IO}}} \exp\left(z_{t,j}\right)} \quad . \tag{21}$$

For the scenario, the constants are set to $\gamma = 4$, $\omega = 4$, $\sigma^2 = 0.3$, and $\nu = 2$. The ideal neural activation for an encoded sample utterance is visualised in Fig. 8a.

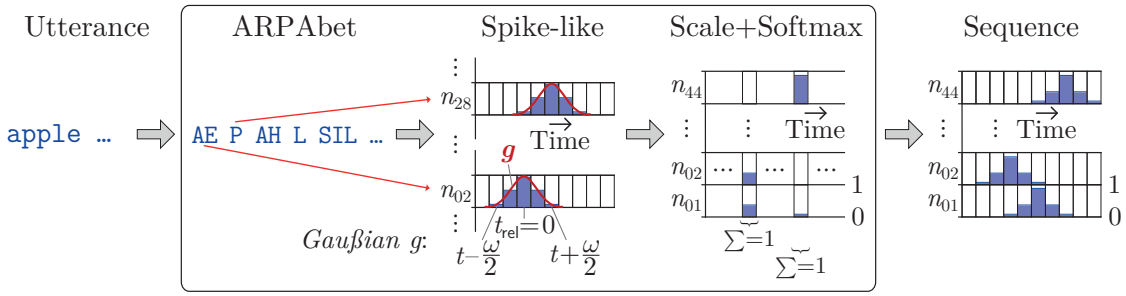


Figure 6. Schematic process of utterance encoding. The input is a symbolic sentence, while the output is the neural activity over $|I_{\mathsf{IO}}|$ neurons times T_{a} time steps.

The utterance encoding is neurocognitively plausible because it reflects both, the neural priming mechanism as well as the fluent activation on a spatially distinct phonetic map (Marslen-Wilson and Zwitserlood, 1989; Rauschecker and Tian, 2000). Although research on neural spatial organisation of phoneme coding is in its infancy, there is evidence for an early organisation of the *Primary Auditory Cortex* (A1) and the *Superior Temporal Sulcus* (STS) forming a map for speech related and speech unrelated sounds (Chang et al., 2010; Liebenthal et al., 2005). The input representation is also in line with an ideal input normalisation to the mean of the activation function, as suggested in (LeCun et al., 1998).

To gather and encode the proprioception of a corresponding manipulation action, the right arm of the NAO is guided by the human teacher. From this steered arm movement, joint angles of the five joints are directly measured with a sampling rate of 20 *Frames Per Second* (FPS). The resulting values are scaled to $[0, 1]$, based on the minimal and maximal joint positions (see Fig. 8b for an example of the proprioceptive features F_{pro}). In a data recording conducted via this scheme, the human teachers are instructed about the four different movements as listed in Fig. 5b. Having an encoding on the joint angle level is neurocognitively plausible because the (human) brain merges information from joint receptors, muscle spindles, and tendon organs into a similar proprioception representation in the S1 area (Gazzaniga et al., 2013). Fig. 8c shows the encoded proprioception for the exemplary manipulation action.

For the visual perception, we aim at capturing a representation that is neurocognitively plausible but on a level of abstraction of shapes and colours and make use of conventional visual perception methods as shown in Fig. 7. At first, the mean shift algorithm is employed for segmentation on an image taken by the robotic learner (Comaniciu and Meer, 2002). The algorithm finds good segmentation parameters by determining modes that describe best the clusters in a transformed 3-D feature space by estimating best matching probability density functions. Secondly, the *Canny edge detection* as well as the OpenCV contour finder are applied for object discrimination (Canny, 1986; Suzuki and Abe, 1985). The first algorithm basically applies a number of filters to find strong edges and their direction, while the second determines a complete contour by finding the best match of contour components. Thirdly, the centre of mass and 16 distances to salient points around the contour are calculated. Here, salient means, for example, the largest or shortest distance between the centre of mass and the contour within intervals of 22.5°. Finally, the distances are scaled by the square root of the object's area and ordered clockwise, starting with the largest. The resulting encoding of 16 values in $[0, 1]$ represents the characteristic shape, which is invariant to scaling and rotation. Encoding of the perceived colour is realised by averaging the three R, G, and B values of the area

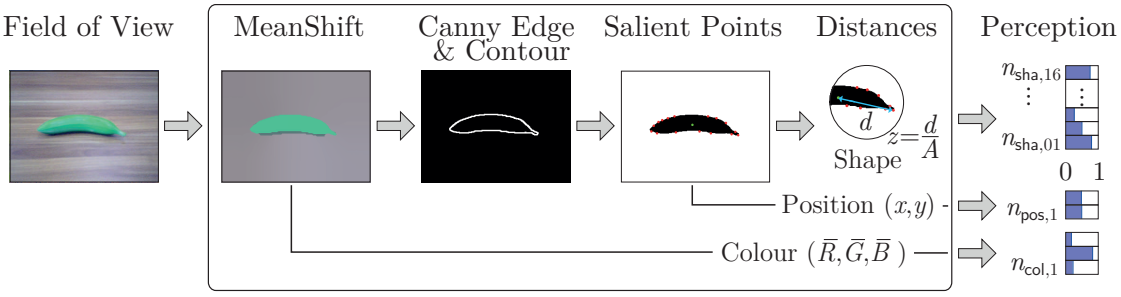
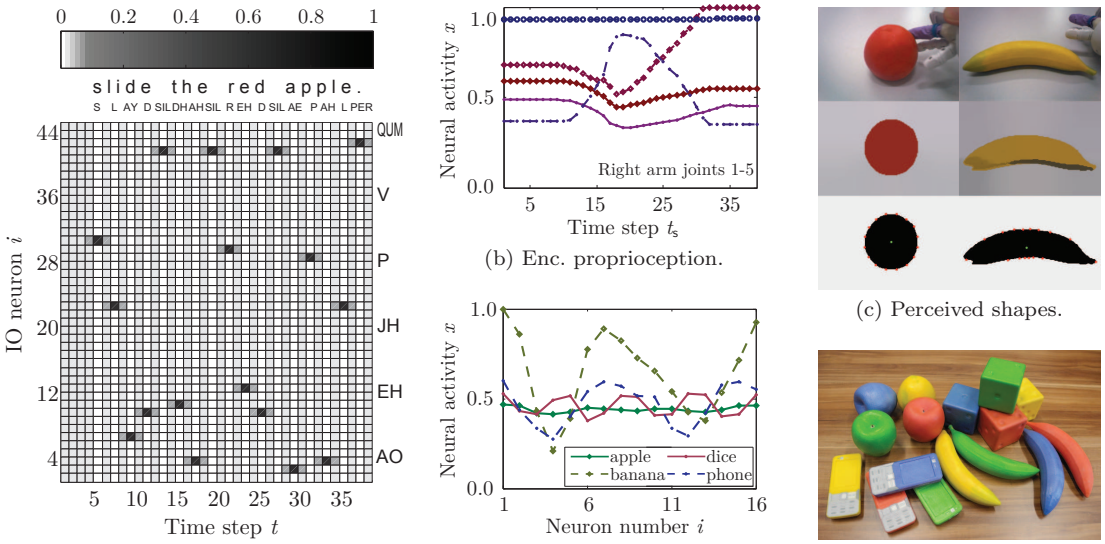


Figure 7. Schematic process of visual perception and encoding. The input is a single frame taken by the NAO camera, while the output is the neural activity over N neurons, with N being the sum over shape + colour + position features.



(a) Encoded utterance.

(b) Enc. proprioception.

(c) Perceived shapes.

(d) Enc. visual shapes.

(e) Some objects of interest.

Figure 8. Representations in the multi-modal language acquisition scenario.

within the shape. Other colour spaces e.g. based on only *hue* and *saturation* could be used as well, but they are in this step mainly a technical choice. Additionally, the perceived relative position of the object is encoded by measuring the two values of the centroid coordinate in the field of view to allow for tests on interrelations between multiple objects later.

The resulting encoding is plausible because in the brain is representing visual information in the process of recognising objects similarly by primarily integrating shape and colour features received from the *Visual Cortex Four* (V4) area (Krüger et al., 2013; Orban, 2008). The shape representation codes the discrimination of objects by combining a number of contour fragments described as the curvature-angular position relative to the objects' centre of mass (Pasupathy and Connor, 1999; Yau et al., 2012). The colour representation codes hue and saturation information of the object invariant to luminance changes (Gegenfurtner, 2003; Tanigawa et al., 2010). For an overview, Fig. 8c provides two prototypical example results of the perception process, Fig. 8d provides a sketch of the visual shape perception encoding, and Fig. 8e shows some of the used objects. The objects have been designed via 3D-print to possess similar masses despite different shapes and comparable colour characteristics across the shapes to provide for robustly and controllably perceivable characteristics.

Capturing motion features also in visual perception is deliberately avoided for several reasons. First of all, from a conceptual perspective, it is desired to keep the visual sensation ambiguous on its own as well as to study the multi-model integration on a conceptual level. Secondly, an agent could experience the movement of an entity in the field of view simply by tracking the said entity with its head or the eyes. This would shift the perception to the somatosensory level and would introduce a redundancy with respect to the arm sensation, which could be difficult to preclude in an analysis.

4.2. *Experimental Setup and Evaluation Measures*

For evaluation, the data was divided 50:50 into training and test sets (all variants of a specific interaction are either in the training or in the test set only) and used to train

ten randomly initialised systems. In addition, this whole process was repeated 10 times as well (10-fold cross-validation) to obtain 100 runs for analysis.

The MTRNNs were parametrised as follows (all parameters are given in Tab. 1). The auditory MTRNN$_\mathsf{a}$ and the visual MTRNN$_\mathsf{v}$ were specified in size based on the previous studies for the EMBMTRNN model (Heinrich et al., 2015; Heinrich and Wermter, 2014). The somatosensory MTRNN$_\mathsf{s}$ was shaped similarly with $|I_{\mathsf{s,Cf}}| = 40$ and $|I_{\mathsf{s,Csc}}| = 23$, based on the experience acquired as well as on other work (Yamashita and Tani, 2008). 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. Also based on previous experience and independent of the data set the number of Csc units were set to $|I_{\mathsf{Csc}}| = \lceil |I_{\mathsf{Cs}}|/2 \rceil$. All weights were initialised similarly within the interval $[-0.025, 0.025]$, while the initial Csc units (auditory MTRNN$_\mathsf{a}$) were randomly taken from interval $[-0.01, 0.01]$ and the final Csc units (somatosensory MTRNN$_\mathsf{s}$ and visual MTRNN$_\mathsf{v}$) from interval $[-1.0, 1.0]$. The learning mechanisms and parameters were identically chosen as in previous studies (Heinrich et al., 2015). Likewise, the timescales for the MTRNN$_\mathsf{a}$ and the MTRNN$_\mathsf{v}$ were based on the resulting values for the related models ($\tau_{\mathsf{a,IO}} = 2$, $\tau_{\mathsf{a,Cf}} = 5$, and $\tau_{\mathsf{a,Cs}} = 70$) (Heinrich et al., 2015; Hinoshita et al., 2011). A good starting point for the timescale setting of the MTRNN$_\mathsf{s}$ were the parameters suggested in original studies ($\tau_{\mathsf{s,IO}} = 2$, $\tau_{\mathsf{s,Cf}} = 5$, and $\tau_{\mathsf{s,Cs}} = 50$) to provide a progressive abstraction (Nishimoto and Tani, 2009; Yamashita and Tani, 2008). For this scenario, the timescales for the somatosensory modality seem not particularly crucial, since the manipulation actions are not strongly dependent on shared motion primitives. A preliminary parameter search (not shown) confirmed these suggestions and revealed good settings for the vision modality in similar ranges ($\tau_{\mathsf{v,IO}} = 2$, $\tau_{\mathsf{v,Cf}} = 5$, and $\tau_{\mathsf{v,Cs}} = 16$).

For the self-organisation forcing parameter of the visual MTRNN$_\mathsf{v}$, a parameter exploration was conducted similarly and is excluded here for brevity. This search revealed that the self-organisation is more crucial for this data set, but that a setting of $\psi_\mathsf{v} = 5.0 \times 10^{-5}$ again is good[8].

[8]detailed results are omitted, but detailed results for the somatosensory MTRNN$_\mathsf{s}$ will be presented within this section

Table 1. Standard meta and training parameter settings for evaluation.

Parameter	Description	Value	Parameter	Description	Value
$\lvert I_{\mathsf{a,IO}}\rvert$	Auditory IO neurons	44	$\tau_{\mathsf{a,IO}}$	Timescale auditory IO	2
$\lvert I_{\mathsf{a,Cf}}\rvert$	Auditory Cf neurons	80	$\tau_{\mathsf{a,Cf}}$	Timescale auditory Cf	5
$\lvert I_{\mathsf{a,Cs}}\rvert$	Auditory Cs neurons	23	$\tau_{\mathsf{a,Cs}}$	Timescale auditory Cs	70
$\lvert I_{\mathsf{s,IO}}\rvert$	Somatos. IO neurons	5	$\tau_{\mathsf{s,IO}}$	Timescale somatos. IO	2
$\lvert I_{\mathsf{s,Cf}}\rvert$	Somatos. Cf neurons	40	$\tau_{\mathsf{s,Cf}}$	Timescale somatos. Cf	5
$\lvert I_{\mathsf{s,Cs}}\rvert$	Somatos. Cs neurons	23	$\tau_{\mathsf{s,Cs}}$	Timescale somatos. Cs	50
$\lvert I_{\mathsf{v,IO}}\rvert$	Visual IO neurons	19	$\tau_{\mathsf{v,IO}}$	Timescale visual IO	2
$\lvert I_{\mathsf{v,Cf}}\rvert$	Visual Cf neurons	40	$\tau_{\mathsf{v,Cf}}$	Timescale visual Cf	5
$\lvert I_{\mathsf{v,Cs}}\rvert$	Visual Cs neurons	23	$\tau_{\mathsf{v,Cs}}$	Timescale visual Cs	16
$\lvert I_{\mathsf{a,Csc}}\rvert$	Auditory Csc units	12	α	Teacher forcing	0.1
$\lvert I_{\mathsf{s,Csc}}\rvert$	Somatos. Csc units	12	η_{max}	Maximal learning rate	1.0
$\lvert I_{\mathsf{v,Csc}}\rvert$	Visual Csc units	12	η_{min}	Minimal learning rate	1.0×10^{-6}
$\mathbf{C}^0_{\mathsf{a},0}$	Initial Csc values range	± 0.01	ξ_+	Increasing factor	1.01
$\mathbf{C}^0_{\mathsf{s},T}$	Init. final Csc r. somatos.	± 1.00	ξ_-	Decreasing factor	0.96
$\mathbf{C}^0_{\mathsf{v},T}$	Init. final Csc r. visual	± 1.00	η^0, β^0, ζ^0	Initial learning rates	0.05
$\mathbf{W}^0$	Initial weights range	± 0.025	ψ_v	Self-organisation f. visual	5.0×10^{-5}

4.3. *Generalisation of Novel Interactions*

Based on good parameters for dimensions, timescales, and learning, a variation of the self-organisation forcing parameter ψ_{s} of the somatosensory MTRNN$_{\mathsf{s}}$ was conducted to test the overall performance of the model. The results of the experiment show that the system is able to generalise well: a high F_1-score and a low edit distance (insertion=1, deletion=1, substitution=2) of 0.984 and 0.00364 on the training as well as 0.638 and 0.154 on the test set was determined for the best network. On average over all runs an F_1-score and an edit distance of 0.952 and 0.0185 for the training as well as 0.281 and 0.417 for the test have been measured ($q_{\mathrm{F_1\text{-}score,mixed}} = 0.617$, $q_{\mathsf{edit\text{-}dist,mixed}} = 0.219$). Note, due to the rigid training scheme there is a high chance that the system had to describe scenes, for which not all aspect (shape, colour, or manipulation action) have been learned before (intended to keep the scenario challenging). For a parameter variation of the self-organisation forcing ψ_{s} over $\{1, 2, 5\}\cdot 10^{-k}$, $k \in \{4, 3, 2\}$, all results are provided in Fig. 9a and c. Notably, the best results originated from setting $\psi_{\mathsf{s}} = 5.0 \times 10^{-4}$.

Training is challenging and rarely perfect yet not over-fitted systems were obtained on the training data. Nevertheless, a high precision (small number of false positives) with a lower up to medium recall (not the exact production of desired positives) was observed on the test data. The errors made in production were mostly minor substitution errors (single wrong phonemes) and only rarely word errors.

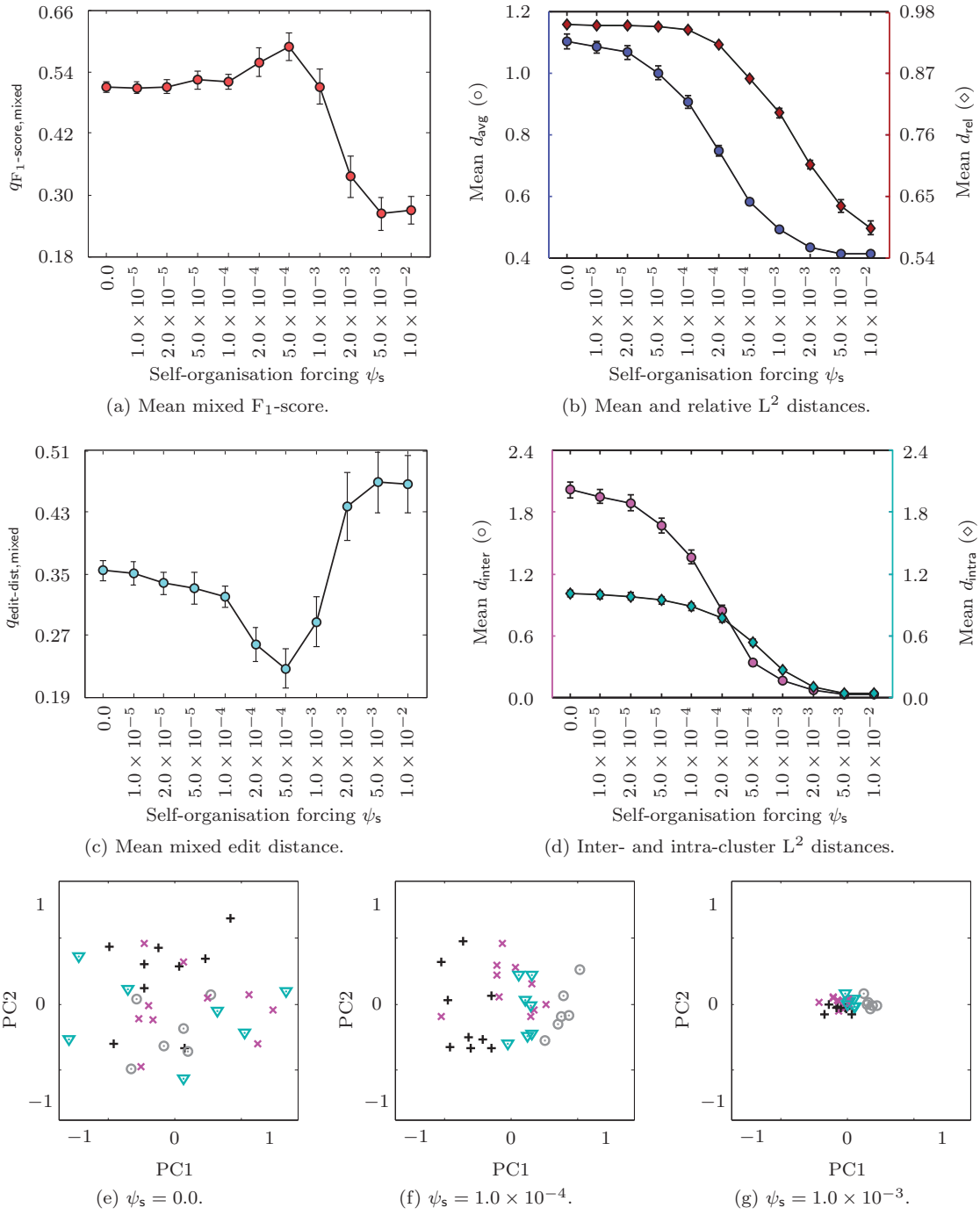
Using a self-organisation mechanism on the final initial Csc values for the somatosensory and visual MTRNNs caused good abstraction from the perception for the described scenario and the chosen ψ_{s} and ψ_{v} values. In this scenario, in fact, the mechanism is very crucial. For both sensory modalities the performance was significantly worse (threshold for $\rho_{\mathsf{t\text{-}test}} < 0.001$) when using static random values for the final internal states of the Csc units in abstracting the sensation $\psi = 0.0$. In particular for proprioception the rate of successfully described novel scenes nearly doubled when using self-organisation forcing with $\psi_{\mathsf{s}} = 5.0 \times 10^{-4}$ compared to random patterns. Based on the experience acquired in the preliminary test (compare Sec. 3.2.2), the obvious hypothesis is that the MTRNN$_{\mathsf{s}}$ self-organised a better distribution of the Csc patterns in the Csc space. However, measuring the Csc space by using the L^2 distance metrics revealed that the patterns are not spreading out, but rather shrink towards small context values, regardless ψ_{s} is set too large (see Fig. 9b): For smaller ψ_{s} the shrinking develops similar but less strong.

To find an alternative hypothesis, the patterns were inspected again in detail. They showed some regularity for scenes including the same manipulation action. Thus, a good performance might correlate with a self-organisation towards similar patterns for similar manipulations. To quantify this effect, two additional measures are used to describe the difference between patterns for scenes with the same or with different manipulations $M = \{\texttt{pull}, \texttt{push}, \texttt{show me}, \texttt{slide}\}$:

$$d_{\mathsf{inter}} = \frac{1}{|M|} \sum_{m_k \in M} d_{\mathsf{avg}}(C_{m_k}) \tag{22}$$

$$d_{\mathsf{intra}} = \frac{1}{(|M|-1)\cdot(|M|/2)} \sum_{k=1}^{|M|-1} \sum_{l=k+1}^{|M|} d\left(\mathrm{centroid}(C_{m_k}), \mathrm{centroid}(C_{m_l})\right) \ , \tag{23}$$

where the inter-cluster distance d_{inter} is the average of all unweighted pair distances of patterns over the scenes that include the same manipulation (e.g. pull, push, show me, and slide) – subsequently averaged over all manipulations. The intra-cluster distance d_{intra} provides the mean of all distances of centroids for the clusters C that contain patterns



(a) Mean mixed F_1-score.

(b) Mean and relative L^2 distances.

(c) Mean mixed edit distance.

(d) Inter- and intra-cluster L^2 distances.

(e) $\psi_s = 0.0$.

(f) $\psi_s = 1.0 \times 10^{-4}$.

(g) $\psi_s = 1.0 \times 10^{-3}$.

Figure 9. Effect of the self-organisation forcing mechanism on the development concept patterns in the MultiMTRNNs model: Mean mixed F_1-score (a) and mixed edit distance (b) – "mixed" measures indicate a combination of training and test results with equal weight, mean of average and relative pattern distances (c), and intra- and intra-cluster distances (d) with interval of the standard error, each over 100 runs and over varied ψ_v respectively; representative developed Csc patterns (e–g) reduced from $|I_{Csc}|$ to two dimensions (PC1 and PC2) for selected parameter settings of no, "good", and large self-organisation forcing respectively. Different words for shapes and colours are shown with different coloured markers (black depicts 'position' utterance).

of the same manipulation. The measurements of the inter- and intra-cluster distances over the varied $\psi_{\mathtt{s}}$ are presented in Fig. 9c. The plots are compared on the same absolute scale and show that the inter-distance is decreasing rapidly with increased $\psi_{\mathtt{s}}$, but the intra-distance decreases much slower. At some point, in fact (e.g. for $\psi_{\mathtt{s}} = 5.0 \times 10^{-4}$), the inter-distance is smaller than the intra-distance. This means that the patterns are indeed clustered best for certain $\psi_{\mathtt{s}}$ values before the shrinkage for the Csc patterns is too strong and the distances vanish. In Fig. 9e–g we can visually confirm this measured clustering on a representative example ("good" in 9f).

4.4. *Self-organisation in the Cell Assemblies*

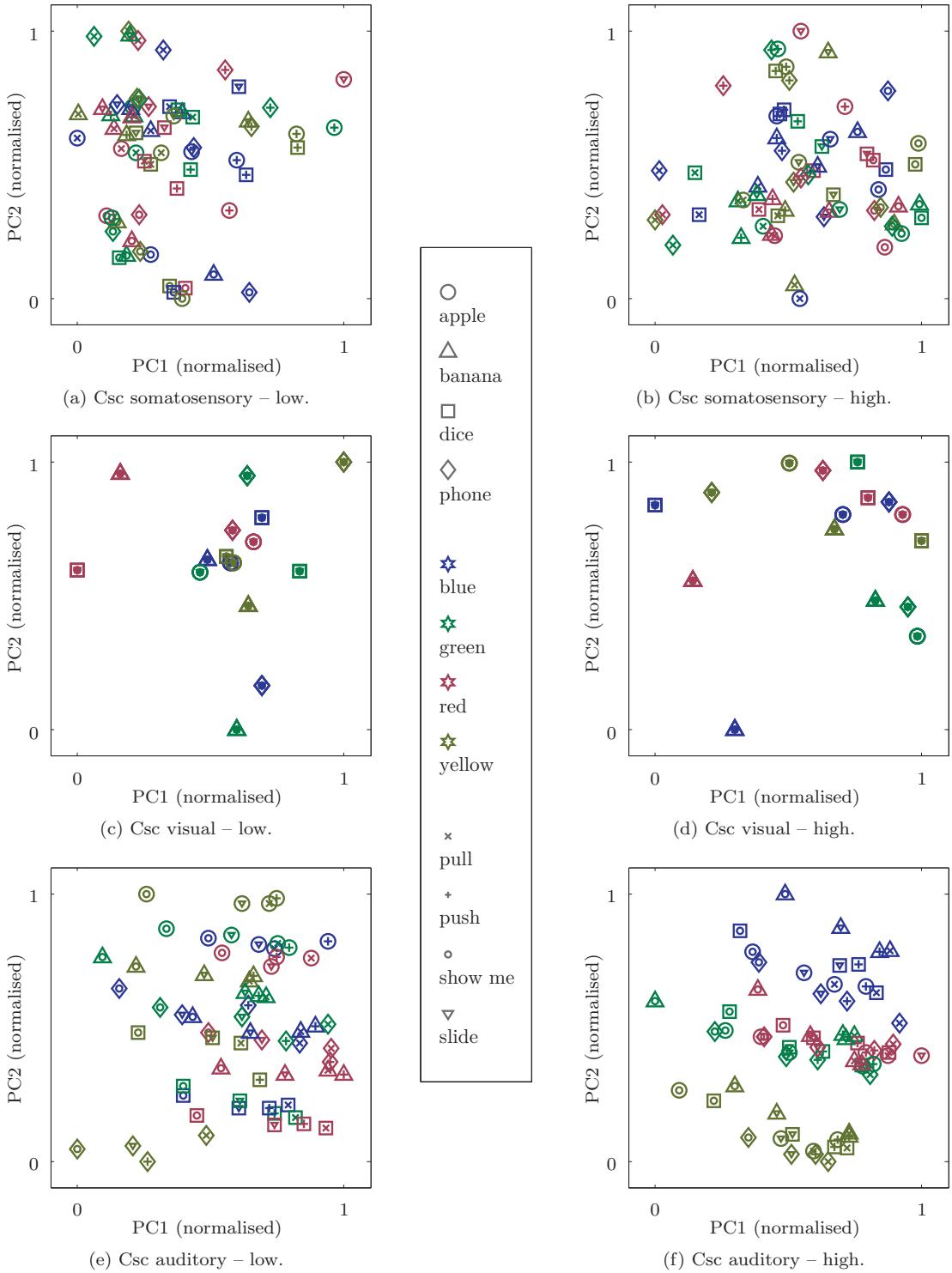
Throughout all tests of the model, diverse patterns of the internal states of the Csc units developed across the modalities. Nonetheless, frequently similar patterns emerged in the respective modality for similar utterances or perceptions. This is particularly the case for the Csc units of the sensory modalities ($\text{MTRNN}_{\mathtt{s}}$ and $\text{MTRNN}_{\mathtt{v}}$), as shown in the last experiment (where a clustering towards patterns for similar perceptions emerged), but also for Csc units of the auditory production subsequently to the activation within the CAs. During training, the Csc units in the auditory $\text{MTRNN}_{\mathtt{a}}$ also self-organised for the presented sequences (utterances). However, within the formation of the CAs by means of the associations, patterns emerged that are able to cover the whole space of scenes in training and test data.

To inspect how these patterns self-organise, we can look into the generated Csc patterns after the whole model is activated by the perception on somatosensory and visual modalities from the training *and* the test data. An example for such Csc activations is presented in Fig. 10 for well-converged architectures with a low[9] generalisation rate (a, c, and e) and a high generalisation rate (b, d, and f). The visualisation is provided by reducing the activity of the Csc units to two dimensions using again *Principle Component Analysis* (PCA) and normalising the values[10]. The results confirm that the patterns form dense and sparse clusters for the visual Csc (the patterns, in fact, overlap each other for different manipulations on the *same* coloured and shaped object). For the somatosensory Csc, the clusters are again reasonable distinct for the same manipulations, although there is a notable mixing between some manipulations on certain objects. For the auditory Csc in case of high generalisation, the patterns are also distinctly clustered. In the example, presented in Fig. 10f, we can discover clustering by colour (prominently on PC2), by manipulation (notably on PC1) and by shape (in between and on lower components). The low generalisation example of Fig. 10e shows the clusters less clear with more patterns scattered across the PC1 and PC2.

Inspecting the sensory data revealed that visual shape and colour sequences are strikingly similar for different manipulation on the same objects, while the proprioception sequences show some differences for some objects. For example, the slide manipulation on banana-shaped objects was notably different than on the other objects. Apart from that, the proprioception sensation is mostly ambiguous with respect to the specific scene (which object of which shape was manipulated) – which was intended in the scenario design. Thus it seems that in the CAs there is a tendency of restructuring the characteristics (shape, colour, or proprioception), which were overlapping for the single modalities, into a representation where all characteristics are distributed.

[9] Test set F_1-score: low generalisation rate 0.117, high generalisation rate 0.638.

[10] The first two components explain the variance in the patterns as follows: low/proprioceptive: 90.75%, low/visual: 52.42%, low/auditory: 83.34%, high/proprioceptive: 97.59%, high/visual: 43.52%, high/auditory: 65.66%.



(a) Csc somatosensory – low.

(b) Csc somatosensory – high.

(c) Csc visual – low.

(d) Csc visual – high.

(e) Csc auditory – low.

(f) Csc auditory – high.

Figure 10. Activity in the Csc units after the model has been activated by proprioception and visual perception for the final internal states (somatosensory and visual) and the initial internal states (auditory), reduced from $|I_{\mathsf{Csc}}|$ to two dimensions (PC1 and PC2) and normalised, each. Visualisation a, c, e are shown for an representative example for low and b, d, f for high generalisation.

4.5. *Summary*

In sum, embedding MTRNNs with context abstraction and an MTRNN with context bias into one coherent architecture allows for a composition of temporal dynamic multi-modal perception into overall concepts and for the decomposition into meaningful sequential actuation, e.g. in terms of a verbal description. From the results, we can deduce that the self-organisation forcing indeed is facilitating the clustering of concepts for similar perceptions by self-organising the space of the internal states of the Csc units upon the structure of the data. Self-organising the patterns in the CAs towards well-distributed clusters highly correlated with the ability to generalise well.

In the novel model, good clustering self-organised for the abstracted context patterns of visual perception and also for somatosensation. For vision, this clustering occurs in particularly dense clusters that are sparsely distributed over the Csc space. For models that generalise well, we found that in the CAs associations emerged that projected the Csc space of the multi-modal sensation (shape, colour, proprioception) into a well-distributed Csc space of auditory production. This distribution self-organised again towards sparsely-distributed dense clusters. Models that are able to successfully describe all training data, but cannot generalise, showed a less well-distributed auditory Csc space.

For the generalisation this means that a well-distributed (sparse) but well-structured (conceptual clusters) auditory Csc space facilitates the grounding of language acquisition into the temporal dynamic features. Such a Csc space allows *modulating* which motor sequence needs to be selected to describe the perception. A good overall abstraction of the respective perceptual features into the CAs thus fosters a correct (good) decomposition into a chain of words and then into phonemes. As a consequence, the CAs fuse but more importantly disambiguate single modalities, which are ambiguous on their own, into an overall coherent representation. Since in the model this happens temporally concurrent, it seems sufficient that different aspects of an observation, just need to co-occur to form a rich but latent overall representation for all modalities.

5. Discussion

For the brain, it has been shown that spatial characteristics of neural connectivity and temporal characteristics of neural activation lead to a hierarchical processing of sensation and actuation (compare Sec. 2). In previous studies researchers have adopted these natural conditions of the cortex in order to model similar hierarchical processing in motor movements and speech production aspects (compare Sec. 3.1). In particular, these conditions were utilised to constrain CTRNNs with timescales and to integrate a context bias. Such a so-called *MTRNN with context bias* model can decompose an initial context into a sequence of primitives. In this paper, this concept is developed further and reversed to allow for composing a sequence of primitives into an abstracted context. A mechanism is proposed to force an entropy-based self-organisation of such a context, which supposedly serves as a key component of an overall model for grounding language in embodied multi-modal sensation.

Self-organising Compositional Representations

The self-organisation forcing mechanism provides the development of a latent representation for the respective abstracted context of a sequential perception, without the need of an a priori definition. In the model, the self-organisation forcing parameter is quite sensitive as too small values hinder a self-organisation, while too large values lead to a

fast premature convergence of the architecture. A cause for the latter case is that both, the forward activity from small weights as well as a too strong adaptation towards this activity, lead to small errors. Thus, the internal states of the final Csc values are self-organised to match the activity from the network, before the network is self-organised to cover the regularities of the data. This issue could be further approached by using a regularisation for the self-organisation or by using weight initialisations based on the eigenvalue of the weight matrix. For the first option, it would be important to consider methods that are independent of the direction of the gradient. For example, a simple normalisation of the internal states of the final Csc units would only skew the distribution and hence could lead to a convergence towards similar Csc patterns. For the second option, a divergence could occur because the randomly initialised Csc pattern could by chance be all similarly small or similarly large. Utilising weight initialisation and normalisation techniques, used in learning deep FFNs (LeCun et al., 2015), might be interesting but can lead to additional instability during RNN training.

For our model, however, this means that for forming a compositional representation it perhaps is sufficient that the data contains regularities as well as irregularities. It seems that a compositional representation is formed solely by minimising differing activity for similar temporal dynamic patterns (in production and sensation), thus by the entropy of different versus similar patterns. For the concepts of the whole temporal dynamic sequences, this entropy-based descent, which is inherent in the self-organisation forcing mechanism, leads to a restructuring of the concept space to represent similar sequences with similar temporally static concept patterns. All in all, the regularities in the data, which that are also rich in our natural environment (Smith and Gasser, 2005), also seem sufficient for an architecture with different timescales.

Multi-modal Context in Language Acquisition

In the novel model, the density of the formed clusters of certain observations was observed to be closely related to the similarity of the abstracted sequences. This observation is logical since the data for the somatosensory and the visual modalities was not compositional and thus the patterns in the Csc formed as a compression of the temporal dynamic observations. As a consequence, the clustering of sequences is limited by the variability of the sequences, since there is no mapping required to a category within the single modalities. By associating the (clustered) multi-modal sensory representations with the auditory production representations, the cell assemblies form as a direct link of the active patterns. The resulting mappings show a close relation to the action-perception circuits measured in the brain (Pulvermüller et al., 2014): the Csc space is re-organised to form specific conceptual webs for co-occurring multi-modal patterns. Since this effect was not built in but emerged from the entropy-based learning, it seems that the conceptual webs are the obvious consequence of the self-organisation.

Regarding our model, this means that the contexts for the single modalities indeed restructure towards a clustering of similar up to identical patterns for similar perceptions. In this way, the model self-organises towards capturing the features that are different in the otherwise ambiguous sequences. By associating the abstracted temporally static context representations of multiple perception modalities with the speech production modality, concept-level cell assemblies emerge that provide a well distributed unambiguous context space. Thereby the context space is modulated to produce novel but correct speech productions. With regard to the brain this relates to the finding of synchronous firing between individual neurons, which react to the same stimulus but scaled-up to cortex level (Alho et al., 2014; Engel and Singer, 2001).

Again, both, the uni-modal representations and the associations, self-organise themselves, driven by the regularities in the data. However, the structuring in the single modalities seems less complex and is easier to re-organise. Hence, the hierarchical abstractions seem to operate like a filter on *some* features from the rich perception. Summarising, this means, that the multi-modal context is an abstraction for important aspects of the perception on various pathways, to cope with the inherently varying temporal resolutions and information densities of the different modalities.

Conclusion and Future Work

Overall, in this paper, we present a neurocognitively plausible model for embodied multimodal language grounding and demonstrate it in a natural interaction of a robotic agent with its environment. The model is an extension of a previous model on embodied grounding, which showed that the spatial and temporal abstraction is an important characteristic for language in the brain (Heinrich et al., 2015), and includes the processing of temporal dynamic somatosensory and visual perception. The characteristics of a neural architecture that facilitating language acquisition that we obtained from the novel model are: a) shared representations of abstracted multi-modal sensory stimuli and motor actions can integrate novel experience and modulate novel production and b) self-organisation might occur naturally because of the structure in the sensorimotor data and both, the spatial and temporal nesting that has evolved in the human brain.

Future research must address the demanding training, a scaling up to larger, and more natural language corpora to cover wider ranges of sensorimotor contingencies. Perhaps we can ease the training by fuzzy characteristics of the neurons, e.g. a stochastic variance in the neurons' firing rate, or the recruitment of new connections without changing the architecture's dimension (LeCun et al., 2015; Murata et al., 2013). In conceptually similar tasks in application, namely sequence to sequence mapping such as video annotation, the machine learning community made tremendous progress recently, e.g. (Donahue et al., 2015; Sutskever et al., 2014). Although many utilised architectures employ computational mechanisms that are not neurocognitively plausible, some aspects like pooling, drop-out, and normalisation can be utilised to short-cut parts of the training that are conceptually not crucial for the model, such as preprocessing for visual feature extraction. For scaling-up, the complexity might get reduced by employing the principle of scaffolding in learning a language corpus (Håkansson and Westander, 2013; Rohlfing et al., 2006; Wrede et al., 2010): words and holo-phrases first, and then more complex utterances without altering the weights from a "word"-layer (e.g. Cf) to the phonetic output. With respect to modelling further phenomena in the brain, it has been suggested that the same conceptual networks may be involved in speech processing, motor action as well as somatosensation (Garagnani and Pulvermüller, 2016; Glenberg and Gallese, 2012). Further refinements can embed hierarchical abstraction and decomposition in utterance comprehension and motor action as well, and test how such a model can replicate an action for verbal descriptions, which were passively learned before or in co-occurrence with the production of an utterance.

With the outcome from our novel model and further refinements, we can design novel neuroscientific experiments on discovering multi-modal integration as well as hierarchical dependencies particularly in language processing and perhaps construct future robotic companions that participate in fascinating discourses.

Acknowledgements

We would like to thank Sascha Griffiths, Sven Magg, Wolfgang Menzel, and Cornelius Weber for fruitful discussions on model characteristics and experimental design as well as for valuable comments on earlier versions of this manuscript. Also, we want to thank Erik Strahl and Carolin Mönter for important support with the robotic hardware and experimental data collections. Finally, we gratefully acknowledge partial support from the German Research Foundation (DFG) under project Crossmodal Learning, TRR-169.

References

Alho, J., Lin, F.-H., Sato, M., Tiitinen, H., Sams, M., and Jääskeläinen, I. P. (2014). Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. *Frontiers in Psychology*, 5(394):1–10.

Awano, H., Ogata, T., Nishide, S., Takahashi, T., Komatani, K., and Okuno, H. G. (2010). Human-robot cooperation in arrangement of objects using confidence measure of neuro-dynamical system. In *Proc. 2010 IEEE Int. Conf. on Systems Man and Cybernetics (SMC)*, p. 2533–2538. IEEE.

Badre, D. and D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9):659–669.

Badre, D., Kayser, A. S., and D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, 66(2):315–326.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59:617–645.

Bergen, B. K. (2012). *Louder Than Words: The New Science of How the Mind Makes Meaning*. Basic Books, New York, US.

Borghi, A. M., Gianelli, C., and Scorolli, C. (2010). Sentence comprehension: effectors and goals, self and others. An overview of experiments and implications for robotics. *Frontiers in Neurorobotics*, 4(3):8 p.

Brosch, M. and Schreiner, C. E. (1997). Time course of forward masking tuning curves in cat primary auditory cortex. *Journal of Neurophysiology*, 77(2):923–943.

Broz, F., et al. (2014). The ITALK project: A developmental robotics approach to the study of individual, social, and linguistic learning. *Topics in Cognitive Science*, 6(3):534–544.

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

Cangelosi, A. and Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, 30(4):673–689.

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

Canny, J. (1986). A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 8(6):679–698.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, 13(11):1428–1432.

Christiansen, M. H. and Chater, N. (2016). *Creating Language – Integrating Evolution, Acquisition, and Processing*. The MIT Press, Cambridge, US.

Comaniciu, D. and Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619.

Coradeschi, S., Loutfi, A., and Wrede, B. (2013). A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intelligenz*, 27(2):129–136.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, 33(1):25–62.

Dayan, P. and Abbott, L. F. (2005). *Theoretical neuroscience*. The MIT Press, Cambridge, US.

Dominey, P. F., Inui, T., and Hoen, M. (2009). Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing. *Brain and Language*, 109(2):80–92.

Dominey, P. F. and Ramus, F. (2000). Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, 15(1):87–127.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proc. 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, p. 2625–2634.

Doya, K. and Yoshizawa, S. (1989). Adaptive neural oscillator using continuous-time back-propagation learning. *Neural Networks*, 2(5):375–385.

Elman, J. L. (1989). Structured representations and connectionist models. In *Proc. 11th Annual Conf. of the Cognitive Science Society (CogSci 1989)*, p. 17–23. Lawrence Erlbaum Assoc.

Engel, A. K. and Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, 5(1):16–25.

Farkaš, I., Malík, T., and Rebrová, K. (2012). Grounding the meanings in sensorimotor behavior using reinforcement learning. *Frontiers in Neurorobotics*, 6(1):13 p.

Feldman, J. A. (2006). *From Molecule to Metaphor: A Neural Theory of Language*. The MIT Press, Cambridge, US.

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, 16(5):262–268.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

Garagnani, M. and Pulvermüller, F. (2016). Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *European Journal of Neuroscience*, 43(6):721–737.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2013). *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton & Company, New York, US, 3 edition.

Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, 4(7):563–572.

Glenberg, A. M. and Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex*, 48(7):905–922.

Grimm, H. (2012). *Störungen der Sprachentwicklung*. Hogrefe, Göttingen, DE, 3 edition.

Hagoort, P. and Levelt, W. J. M. (2009). The speaking brain. *Science*, 326(5951):372–373.

Håkansson, G. and Westander, J. (2013). *Communication in humans and other animals*. Number 4 in Advances in Interaction Studies. John Benjamins Pub., Amsterdam, UK.

Hayes, D. P. and Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of 'motherese'? *Journal of child language*, 15(2):395–410.

Heinrich, S., Magg, S., and Wermter, S. (2015). Analysing the multiple timescale recurrent neural network for embodied language understanding. In Koprinkova-Hristova, et al., eds., *Artificial Neural Networks – Methods and Applications in Bio-/Neuroinformatics*, vol. 4 of *SSBN*, ch. 8, p. 149–174. Springer Int. Pub. Switzerland.

Heinrich, S., Weber, C., and Wermter, S. (2012). Adaptive learning of linguistic hierarchy in a multiple timescale recurrent neural network. In *Proc. 22nd Int. Conf. on Artificial Neural Networks (ICANN 2012)*, vol. 7552 of *LNCS*, p. 555–562. Springer Ber. Hdg.

Heinrich, S. and Wermter, S. (2014). Interactive language understanding with multiple timescale recurrent neural networks. In *Proc. 24th Int. Conf. on Artificial Neural Networks (ICANN 2014)*, vol. 8681 of *LNCS*, p. 193–200. Springer Int. Pub. Switzerland.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402.

Hinoshita, W., Arie, H., Tani, J., Okuno, H. G., and Ogata, T. (2011). Emergence of hierarchical structure mirroring linguistic composition in a recurrent neural network. *Neural Networks*, 24(4):311–320.

Hopfield, J. J. and Tank, D. W. (1986). Computing with neural circuits: A model. *Science*, 233(4764):625–633.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Indefrey, P. and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1–2):101–144.

Karmiloff, K. and Karmiloff-Smith, A. (2002). *Pathways to language: From fetus to adolescent*. Harvard Univ. Press, Cambridge, US.

Krüger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodríguez-Sánchez, A. J., and Wiskott, L. (2013). Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *Pattern Analysis and Machine Intelligence, IEEE Trans.*, 35(8):1847–1871.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998). Efficient backprop. In Orr, G. et al., eds., *Neural Networks: Tricks of the Trade*, vol. 1524 of *LNCS*, p. 9–50. Springer Ber. Hdg.

Levelt, W. J., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., and Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, 98(1):122.

Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13464–13471.

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., and Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex*, 15(10):1621–1631.

Marocco, D., Cangelosi, A., Fischer, K., and Belpaeme, T. (2010). Grounding action words in the sensorimotor interaction with the world: Experiments with a simulated iCub humanoid robot. *Frontiers in Neurorobotics*, 4(7):15 p.

Marslen-Wilson, W. and Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human perception and performance*, 15(3):576.

Monner, D. and Reggia, J. A. (2012). Emergent latent symbol systems in recurrent neural networks. *Connection Science*, 24(4):193–225.

Murata, S., Namikawa, J., Arie, H., Sugano, S., and Tani, J. (2013). Learning to reproduce fluctuating time series by inferring their time-dependent stochastic properties: Application in robot learning via tutoring. *Autonomous Mental Development, IEEE Trans.*, 5(4):298–310.

Nishide, S., Nakagawa, T., Ogata, T., Tani, J., Takahashi, T., and Okuno, H. G. (2009). Modeling tool-body assimilation using second-order recurrent neural network. In *Proc. 2009 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2009)*, p. 5376–5381. IEEE.

Nishimoto, R. and Tani, J. (2004). Learning to generate combinatorial action sequences utilizing the initial sensitivity of deterministic dynamical systems. *Neural Networks*, 17(7):925–933.

Nishimoto, R. and Tani, J. (2009). Development of hierarchical structures for actions and motor imagery: a constructivist view from synthetic neuro-robotics study. *Psychological Research*, 73(4):545–558.

Noda, K., Arie, H., Suga, Y., and Ogata, T. (2014). Multimodal integration learning of robot behavior using deep neural networks. *Robotics and Autonomous Systems*, 62(6):721–736.

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews*, 88(1):59–89.

Palm, G. (1990). Cell assemblies as a guideline for brain research. *Concepts in Neuroscience*, 1(1):133–147.

Pasupathy, A. and Connor, C. E. (1999). Responses to contour features in macaque area v4. *Journal of Neurophysiology*, 82(5):2490–2502.

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books, New York, US.

Pulvermüller, F. (2003). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge Univ. Press, Cambridge, UK.

Pulvermüller, F. and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, 11(5):351–360.

Pulvermüller, F., Garagnani, M., and Wennekers, T. (2014). Thinking in circuits: toward neurobiological explanation in cognitive neuroscience. *Biological Cybernetics*, 108(5):573–593.

Rauschecker, J. P. and Tian, B. (2000). Mechanisms and streams for processing of 'what' and 'where' in auditory cortex. *Proceedings of the Nat. Acad. of Sciences*, 97(22):11800–11806.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: the rprop algorithm. In *Proc. IEEE Int. Conf. on Neural Networks (ICNN93)*, vol. 1, p. 586–591. IEEE.

Rohlfing, K. J., Fritsch, J., Wrede, B., and Jungmann, T. (2006). How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199.

Schmolesky, M. T., Wang, Y., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., and Leventhal, A. G. (1998). Signal timing across the macaque system. *Journal of Neurophysiology*, 79(6):3272–3278.

Schulz, R., Glover, A., Milford, M. J., Wyeth, G., and Wiles, J. (2011). Lingodroids: Studies in spatial cognition and language. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA2011)*, p. 178–183. IEEE.

Singer, J. M. and Sheinberg, D. L. (2010). Temporal cortex neurons encode articulated actions as slow sequences of integrated poses. *The Journal of Neuroscience*, 30(8):3133–3145.

Smith, K. and Kirby, S. (2012). Compositionality and linguistic evolution. In Werning, M., et al., eds., *The Oxford handbook of compositionality*, ch.r 25. Oxford Univ. Press.

Smith, L. B. and Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1–2):13–29.

Smith, L. B. and Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Smith, M. A. and Kohn, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. *The Journal of Neuroscience*, 28(48):12591–12603.

Sporns, O., Chialvo, D. R., Kaiser, M., and Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8(9):418–425.

Steels, L., Spranger, M., van Trijp, R., Höfer, S., and Hild, M. (2012). Emergent action language on real robots. In Steels, L. and Hild, M., eds., *Language Grounding in Robots*, ch. 13, p. 255–276. Springer Science+Business Media LLC New York.

Sutskever, I., Vinyals, O., and Le, Q. V. V. (2014). Sequence to sequence learning with neural networks. In *Proc. 28th Annual Conf. on Neural Information Processing Systems (NIPS2014)*, vol. 27 of *Adv. in NIPS*, p. 3104–3112. Curran Assoc., Inc.

Suzuki, S. and Abe, K. (1985). Topological structural analysis of digitized binary images by border following. *Graphical Models and Image Processing*, 30(1):32–46.

Tani, J. (2014). Self-organization and compositionality in cognitive brains: A neurorobotics study. *Proceedings of the IEEE*, 102(4):586–605.

Tani, J., Nishimoto, R., Namikawa, J., and Ito, M. (2008). Codevelopmental learning between human and humanoid robot using a dynamic neural network model. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans.*, 38(1):43–59.

Tanigawa, H., Lu, H. D., and Roe, A. W. (2010). Functional organization for color and orientation in macaque v4. *Nature Neuroscience*, 13(12):1542–1548.

Tomasello, M. (2003). *Constructing a Language*. Harvard Univ. Press, Cambridge, US.

Ulanovsky, N., Las, L., Farkas, D., and Nelken, I. (2004). Multiple time scales of adaptation in auditory cortex neurons. *The Journal of Neuroscience*, 24(46):10440–10453.

Wrede, B., Kopp, S., Rohlfing, K., Lohse, M., and Muhl, C. (2010). Appropriate feedback in asymmetric interactions. *Journal of Pragmatics*, 42(9):2369–2384.

Yamashita, Y. and Tani, J. (2008). Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLoS Comput Biol*, 4(11):e1000220.

Yau, J. M., Pasupathy, A., Brincat, S. L., and Connor, C. E. (2012). Curvature processing dynamics in macaque area V4. *Cerebral Cortex*, 23(1):198–209.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: a computational study. *Connection Science*, 17(3):381–397.