Research Highlights:

- A computational Developmental Deep Model of Action and Naming (DDMAN), similarly to children, produced scale errors
- Scale errors were frequent at the beginning of training, and decreased linearly; after training DDMAN sporadically produced scale errors
- Objects representations were coarsely organized by shape, causing the initial oversight of object size during action selection; gradually the model learned to attain to size
- Our simulations demonstrate that scale errors are a natural consequence of learning to associate objects with actions

Title: **Children's scale errors are a natural consequence of learning to associate objects with actions: a computational model**

Beata J. Grzyb[1,2], Yukie Nagai[3], Minoru Asada[2], Allegra Cattani[1], Caroline Floccia[1] and Angelo Cangelosi[1,4]

1 University of Plymouth, Drake Circus, Plymouth, PL4 8AA, United Kingdom, **2** Osaka University, 1-1 Yamadaoka, Suita, Osaka, 656-0871, Japan, **3** National Institute of Information and Communications Technology, 1-4 Yamadaoka, Suita, 565-0871, Osaka, Japan, **4** School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom

Correspondence

Beata J. Grzyb, Division of Psychology and Language Sciences, University College London, 26 Bedford Way, Bloomsbury, London WC1H 0AP, United Kingdom.

Email: b.grzyb@ucl.ac.uk

Abstract

Young children sometimes attempt an action on an object, which is inappropriate because of the object size -- they make scale errors. Existing theories suggest that scale errors may result from immaturities in children's action planning system, which might be overpowered by increased complexity of object representations or developing teleofunctional bias. We used computational modelling to emulate children's learning to associate objects with actions and to select appropriate actions, given object shape and size. A computational Developmental Deep Model of Action and Naming (DDMAN) was built on the dual-route theory of action selection, in which actions on objects are selected via a direct (non-semantic or visual) route, or an indirect (semantic) route. As in case of children, DDMAN produced scale errors: the number of errors was high at the beginning of training and decreased linearly but did

not disappear completely. Inspection of emerging object-action associations revealed that these were coarsely organized by shape, hence leading DDMAN to initially select actions based on shape rather than size. With experience, DDMAN gradually learned to use size in addition to shape when selecting actions. Overall, our simulations demonstrate that children's scale errors are a natural consequence of learning to associate objects with actions.

Key words: scale errors, computational model, action selection, perception-action coupling

Acknowledgements

This work was funded by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme (ORATOR FP7-PEOPLE-2012-IEF).

Children sometimes make serious attempts to perform an action on an object, which is impossible because of the object size. These action errors can be referred to as body scale errors, when children act on an object which has an inappropriate size to accommodate for their body (e.g. Brownell, Zerwas, & Ramani, 2007; DeLoache, LoBue, Vanderborght, & Chiong, 2013; DeLoache, Uttal, & Rosengren, 2004). Examples include children attempting to sit in a tiny chair, put doll shoes on their own feet or get inside small cars. Other types of scale errors are referred to as object/tool scale errors, when children use an object to act on another object that is too small or too big to afford its function (Casler, Eshleman, Greene, & Terziyan, 2011; Ware, Uttal, Wetter, & DeLoache, 2006). Examples include children attempting to put a doll on a tiny bed or use a scooping net too big to fit into an aquarium.

Scale errors are common during early childhood; almost all (if not all) typically developing children commit scale errors**.** In laboratory situation, body scale errors were reported in nearly half of 18 to 30 month old children (DeLoache et al., 2004), and tool scale errors were observed in almost all (with one exception) of 18 to 42 month old children (Casler et al., 2011). These action errors are quite frequent also in everyday life. Almost all parents (with one exception) reported that their child (13 – 21 month old at the onset of the study) made at least one scale error over a period of 6 months (Rosengren, Gutierrez, Anderson, & Schein, 2009). The prevalence of scale errors decreases linearly with age (Brownell, Zerwas, & Ramani, 2007; Grzyb, Cangelosi, Cattani, & Floccia, 2018; Rosengren, Carmichael, Schein, Anderson, & Gutierrez, 2009; but see DeLoache et al., 2004; Ware, Uttal, & DeLoache, 2010). Scale errors, however, do not disappear completely after childhood. Casler, Hoffman and Eshleman (2014) showed that even adults are prone to make scale errors. When presented with two alternative tools, one that is typically used “for the job” but has an

inappropriate size to be effective (either too big or too small), and another one that is less typical but perfectly suitable (in terms of size), adults tend to choose the typical tool despite its size.

The existing theories agree that scale errors originate from immaturities in the action planning/selection system (Brownell, Zerwas, & Ramani, 2007; Casler et al., 2011; DeLoache et al., 2004; Glover, 2004). Indeed, scale errors do not seem to originate from inadequate action execution, since children's behaviour while attempting to act on the inappropriately sized objects demonstrates that the incorrectly selected action is appropriately scaled to the object size during action execution. For instance, when children initiate interaction with the miniature car, they first approach it and bend over or kneel down to get closer to it, use a precise grip to open the small door, and aim their foot for the tiny opening. The theories, therefore, focus on different aspects of cognitive processing that may influence children's abilities to form appropriate action plans.

DeLoache et al. (2004) suggested that scale errors reflect an immaturity in inhibitory control and in the integration of information processed by the two distinct visual systems (Goodale & Milner, 1995). Upon seeing a miniature object, such as a tiny chair, the action planning system (or the ventral stream) forms an action plan based on the features of the object (e.g. shape, colour), leading to the selection of an action plan appropriate for interacting with a full-sized chair (e.g. sitting). The action control system (or the dorsal system) should then inhibit the selected action plan based on incompatible size information, but fails to do so because of the child's weak inhibitory control, resulting in a scale error.

What would cause such failures in the integration of visual information for action? Mandler and DeLoache (2012) pointed to developmental changes in the complexity of

object representations as a potential culprit of children's scale errors. In this view, when the children's representation of objects gains in complexity, and the motor routines become more integrated, the perceptual system becomes unstable. The sight of a familiar object causes strong activations that override the perception of the object size, resulting in the selection of inappropriate actions for the size of the object.

Casler and colleagues (2011) suggested that children's perception might be overtaken by a rapidly developing bias to view objects teleofunctionally, that is, as entities existing for a particular purpose (e.g., a chair is for sitting, a paintbrush for painting). Children (and also adults, see Casler et al., 2014) have difficulties dissociating an object from its function, even though the particular instance has the wrong size to be effective. Under certain – unfavourable – circumstances, such as when no viable options or no visual cues for comparison exist, the actions can be guided by an object's categorical function. In other words, a sight of an object, miniature or full-sized, invokes the typical function associated with that object; children may attempt to use it for its typical purpose despite its size.

Alternatively, children might generate inappropriate actions because of the immature, developing knowledge about their own body size in relation to objects (Brownell et al., 2007). In this view, children's scale errors arise due to failures in the action planning processes that incorporate/access information about their own body size and relate this information to the retrieved information about the object size. For instance, to avoid trying to slide down a too small slide, children must process their own body's size and compare/relate it to the size of a slide.

In sum, the existing explanations focus on the immaturity of the action planning system, and within this, vary as to which cognitive processes are responsible for

errors. In the current study, we used computational modelling as a mean to explore the possible reasons of failures in the action selection processes in children.

Our computational model builds on the dual-route theory of action selection, first proposed by Riddoch, Humphreys and Price (1989), and later on, instantiated in a "quasi-modular" connectionist model by (Yoon, Dietmar, Heinke, & Humphreys, 2002). The dual-route theory proposes that actions can be selected via a direct (non-semantic or visual) route based on the visual properties of objects (e.g. affordances), or via indirect (semantic) route, based on extra perceptual knowledge such as object name or canonical object functions. This theory builds upon an assumption that both dorsal and ventral areas are recruited/engaged in action selection process: the direct route corresponds to the dorsal stream that focuses on object use, and the indirect route corresponds to the ventral visual stream that focuses on object knowledge. Once the action has been selected (via direct or indirect routes), visual information may be used directly to adjust the specific motor response (Milner & Goodale, 1995).

The dual-route theory was previously instantiated in a "quasi-modular" connectionist model, in which the values of connections between different modules were fixed based on neuroscientific findings. The model successfully accounted for error patterns in action selection in normal subjects (when they responded under a fast deadline) and patterns of impairment found in brain-lesioned patients, demonstrating that the dual-route theory is a promising framework to study action errors. Compelled by its accountability for action errors in adults, we implemented the dual-route theory in a Developmental Deep Model of Action and Name (DDMAN) to emulate the learning/developmental processes of action selection in young children. The values of connections between different modules in our model were not fixed as in Yoon et al.'s model (2002), but set initially to random small values and adjusted during the training

process to reflect the input patterns. As input, we used various objects from four categories (i.e., chair, car, motorbike, and umbrella) and in two sizes (full-sized and miniature). Over training, DDMAN learns to associate objects with their actions in the visual route, and with their actions and names in the lexico-semantic route. For instance, the lexico-semantic route of DDMAN learns to associate a full-sized chair with an action "sitting" and a name "chair", and to associate a miniature chair with the same name, but a different action "grasping". After each training epoch, we presented the model with test objects and recorded actions that were selected by the model as appropriate for these objects.

The primary goal of our simulations was to examine whether action selection processes that use emerging object-action associations to act upon objects would produce scale errors. Examination of action error patterns and the emerging representations (i.e., object-action and object-action-name associations) over training epochs will allow us to observe whether scale errors occur, and if so, under which circumstances.

The second goal of our simulations was to determine the shape of the function of scale errors with age (i.e., training epochs in our computational study). Some studies show an inverted-U shape development with the peak of scale errors occurring at 20 – 24 months and their disappearance by 30 months (DeLoache et al., 2004; Ware et al., 2010), while others showed a linear decrease from infancy to 40 months (Brownell et al., 2007; Grzyb et al., 2018; Rosengren et al., 2009). Whereas a decreasing linear function points to a maturation process linked to any number of developing concurrent abilities (object representation, inhibitory control, body knowledge, etc.), an inverted U-shaped function would suggest that scale errors are a stage experienced by all typically developing children. In our computational study, we will observe

patterns of scale errors to determine whether these errors take an inverted-U shape or linear function.

Finally, we will examine whether object naming increases the number of scale errors produced by DDMAN. In a tool-based scale error scenario, object naming affects children's scale errors: when the objects are not named, then scale errors decrease over the two test trials; however, when the objects are named, scale errors remain high in the second test trial (Hunley & Hahn, 2016). In addition, Oláh, Elekes, Pető, Peres and Király (2016) showed that children were more likely to make scale errors when tools were named by a non-word by a native speaker as opposed to a foreign speaker, suggesting that meaningful linguistic context is needed for this facilitation to occur. Finally, Grzyb et al. (2018) showed that body scale errors are more likely to be observed in early talkers than late talkers. Object name learning changes some aspects of object representation (Balaban & Waxman, 1997; Welder & Graham, 2001; Xu, 2002; Xu, Cote, & Baker, 2005), which may in turn modulate the incidence of scale errors. If object naming influences scale errors, we may observe a larger number of scale errors produced in the indirect semantic route then in the direct visual route of DDMAN.

In summary, we aimed at expanding our understanding of the mechanism underlying children's scale errors by emulating children's learning processes to associate objects with actions and to select actions that are appropriate for the object category and size. We will examine the patterns of action errors, scale errors produced and emerging object representations, to determine (1) the possible origins of scale errors; (2) the shape of the function of scale errors with age (i.e., training epochs); (3) the influence of object naming on the prevalence of scale errors.

Method

We used computational modelling to explore how objects are associated with actions, and how the action selection system learns to select an appropriate action for a given object. The dual-route theory was instantiated in a deep belief network, named the Developmental Deep Model of Action and Naming (DDMAN). Training and test set included pixel images depicting silhouettes of objects along with binary vectors encoding their size. Objects from four categories (i.e., chair, car, motorbike, and umbrella) and in two different sizes (full-sized and miniature) were used; these were coupled during training with action and name labels. Importantly, no patterns representing scale errors were included, as objects were always presented with appropriate action labels. For instance, a full-sized chair was always presented with the action "sitting", while a miniature chair was presented with the action "grasping". Hence, a miniature chair was never presented with the action corresponding to a full-sized chair. After each training epoch, the ability of DDMAN to select appropriate actions for test objects was examined, and all action errors and scale errors were recorded.

Developmental Deep Model of Action and Naming (DDMAN)

We chose to implement the dual-route theory in deep neural networks for the following reasons. First, deep neural networks are particularly suitable to simulate human brain structure (LeCun, Bengio, & Hinton, 2015) and provide novel computational ways to explore how knowledge is acquired and represented (e.g., Zorzi, Testolin, & Stoianov, 2013). Second, the generative properties of a deep neural network make it extremely suitable for building multimodal associations (e.g., Srivastava & Salakhutdinov, 2014; Ngiam et al., 2011). For instance, deep neural networks have been used to learn a shared representation between audio and visual

data (Ngiam et al., 2015), and to replicate the McGurk effect, that is, an audio-visual perception phenomenon where a visual /ga/ with an audio /ba/ is perceived as /da/ by most participants (McGurk & McDonald, 1976).

The main objective of our simulations was to examine whether scale errors are produced while the network learns to associate objects with actions and names, and to examine how developing object representations may influence the action selection system. Hence, we simplified the neural architecture and the type of input data. Similarly to what was done by Hinton, Osidero and Teh (2006), we used pixel images to code for objects and explicit class labels for object actions and names.

Regarding the choice of visual input format, we simplified the perceptual input to avoid the problem of size-invariant object recognition, which is known to be a hard problem in machine learning and is typically approached with convolutional neural networks (e.g. van Noord, & Postma, 2017; Xu, Xiao, Zhang, Yang, & Zhang, 2014). Hence, the input to our networks were pixel images depicting object silhouettes and binary vectors that explicitly encode object size (i.e., width, length and height).

In our simulations, the action labels correspond to a general action category. However, other works have shown that these labels could be replaced by multilayer pathway that uses, for instance, spectrograms from speakers saying isolated words (e.g. Ngiam et al., 2011). Therefore, our computational model could potentially be extended in the future to incorporate also action control (i.e., parametrization) and object names (i.e., spectrograms for words).

Finally, following the existing theories of scale errors (e.g., DeLoache et al., 2004; Casler et al., 2011), we focused entirely on the action planning / action selection system as a candidate responsible for scale errors in children, and not on the actual

action control system which seems to be able to accurately adjust once a initiated faulty action plan has been selected.

As mentioned earlier, the general framework of DDMAN adapts the "boxes and arrows" dual-route theory proposed by Riddoch and colleagues (1989), and later on, instantiated in a connectionist model (Name and Action Model, NAM) by Yoon, Heinke and Humphreys (2002). In the dual-route theory of action selection (as well as in NAM), the information flows from the visual module via direct route to the action system, or via indirect route through the semantic system to the action system. Based on the growing literature from neuroscience and psychology on the reciprocal relation between perception and action (e.g., Rizzolatti & Craighero, 2004; Rizzolatti & Fadiga, 1998; Thill et al., 2013), we added in the direct visual route of DDMAN an associative system that links the visual and action systems. Such a bidirectional link allows DDMAN to select actions when visual object is presented as input, but also to generate object visual features for a given action. For instance, given a visual input of a chair, an action "sitting" can be selected; for an action "driving" the visual features of a typical car can be generated.

We implemented DDMAN in a class of deep neural networks called deep belief networks (e.g. Hinton, Osindero, & Teh, 2006). The visual module and associative systems in DDMAN are organised as layers of stochastic artificial neural networks, called Restricted Boltzmann Machine (or RBM) (Freund & Haussler, 1992; Hinton, 2002; Smolensky, 1986). Each RBM consists of a number of "neuron-like" elements (i.e. units) organised in two layers: a visible and a hidden layer. The units within the same layer are not connected, but each unit from a visible layer is connected to each neuron in the hidden layer. The connections between units are assigned numbers called "weights". A greater magnitude of the weight between a visible unit and a

hidden unit means that the unit has greater influence over the hidden unit (i.e. to increase hidden unit's level of activation). Each hidden unit receives inputs multiplied by their respective weights. The sum of those products is again added to a bias, and the result is passed through the activation algorithm that produces one output for each hidden unit.

When one RBM is stacked on another RBM, the hidden layer of the lower RBM becomes a visible layer for the upper RBM. That is, the output of hidden layer 1 would be passed as input to hidden layer 2, and so on, if the network has more layers. The number of layers as well as the number of units within layers are important parameters that might influence the neural network's capacity to learn and form representations (Zorzi, Testolin, & Stoianov, 2013). On the one hand, a larger number of units in a visual module might allow encoding of particular characteristics of the object visual features, perhaps to the expense of general features such as shape and size. On the other hand, a smaller number of units might lead to a greater compression of emerging representations, increasing the generality of the learned features. In addition, a large number of units in an associative (top) layer (i.e., sensorimotor and lexico-semantic systems) might help to unfold categories and facilitate the construction of object-action associations. We explored several possible architectures, varying the number of layers (1- and 2-layers) and the number of units (250 or 500 units), within the visual layer, the sensorimotor and lexico-semantic systems (500, 1000, or 2000 units).

Training and test sets

Training and test sets included objects from four categories (i.e., chair, car, motorbike, and umbrella). As discussed earlier, object shape and size were processed separately. To create object shape inputs, pixel images depicting object silhouette

were extracted from CalTech 101 Silhouettes Data Set. Each image was 28 x 28 pixel size and depicted a filled black outline on a white background. As it can be seen in Figure 2, each object category has a highly distinctive category-specific features. These images were converted into binary vectors of size 784. To create object size inputs, we roughly estimated the objects sizes appropriate for 2-3 year old children, as well as sizes of their miniature replicas. We assumed the following sizes for chairs: 40 x 24 x 10 cm (full-sized), 6 x 5 x 6 cm (miniature); for cars: 75 x 42 x 85 cm (full-sized) and 15 x 8 x 17 cm (miniature); for motorbikes: 73 x 42 x 59 cm (full-sized) and 12 x 7 x 5 cm (miniature), and finally, for umbrellas: 68 x 85 x 98 cm (full-sized) and 10 x 13 x 15 cm (miniature). The decimal numbers were subsequently converted into binary representations with 7 bits per dimension, resulting in binary vectors of size 21.

In total, the training set included 616 paired binary vectors of shape and size: 100 of chairs, 196 of cars, 200 of motorbikes, and 120 of umbrellas. These were equally distributed among full-sized and miniature sizes. The test dataset included 154 samples: 24 chairs, 50 cars, 50 motorbikes, and 30 umbrellas, again equally distributed among full-sized and miniature sizes.

For training the associative systems, that is, the sensorimotor and lexico-semantic systems, each visual representation was paired with one action label based on the object category and size (i.e., a full-sized chair was paired with a different action label than a miniature chair) and with one name label based on the object category (i.e., a full-sized chair was paired with the same name label as a miniature chair).

Training procedure

The training process of the neural networks shown in Figure 1 was run in steps, with each step being seen as training a single RBM. The values of the connection weights

within an RBM were first initialized with small random values drawn from a zero-mean Gaussian distribution with a standard deviation of 0.1 and subsequently adjusted during training using the contrastive divergence algorithm (Hinton, 2002). During training, the probability that the network assigns to a given input data can be raised by adjusting the weights to lower the energy of that input and to raise the energy of other inputs. Hence, low energy states of the network correspond to the training patterns. After training, similar patterns have energy states closer to each other, whereas two orthogonal patterns have energy states more distant from one another.

Figure 3 illustrates the training and testing phases of DDMAN. First, the visual module was trained independently on 616 paired shape and size vectors (see Figure 3a); these training sets were divided further into 62 mini-batches. The layer of weights in the visual module was trained for 100 sweeps through the training set (called "epochs"). The weights were modified after each mini-batch.

In the case of the 2-layer visual module, once training of the first layer was completed, the activations of the hidden layer were used as inputs to the second hidden layer. The second layer of weights was trained for 100 epochs.

Subsequently, the sensorimotor system was trained on the hidden activations of the visual module and on the action labels, while the lexico-semantic system was trained on the hidden activations of the visual module and on the action and name labels (Figure 3b). The labels were represented by turning on one unit in a "softmax" group of 8 units for action and 4 units for name.

To ensure that observable learning effects did not arise due to random effects, the training procedure was repeated 10 times for each tested configuration, each time initializing the connections within the network with different (small) random weights.

In all our simulations, the learning rate was set to 0.1, momentum to 0.5, and, weight decay to 0.0002.

Testing procedure

After each training epoch of the associative systems (i.e., sensorimotor and lexico-semantic systems), we tested the abilities of DDMAN to generate appropriate actions for test objects (i.e., paired object pixel silhouettes and size vectors). Similarly to what was done in Hinton et al. (2006), we used free energy of the associative systems to select appropriate actions and names. To this end, we first propagated visual input through the visual module and clamped the states of the visible layer of the associative systems. Subsequently, we turned on, in turn, each of the action label units, and computed the exact free energy of the resulting vector. The action labels that generated the lowest free energy were selected as the network's response.

In case of the information flow via indirect semantic route, we turned on, in turn, each of the action label units and name label units, and computed the exact free energy of the resulting vector. Again, the action and name labels that generated the lowest free energy were selected as DDMAN's response.

Coding of scale errors

A scale error in child studies is defined as any instance of a child seriously attempting to perform an action on an object that is impossible because of the object size (e.g., Casler et al., 2011; DeLoache et al., 2004). In accordance with empirical studies, a scale error in our computational simulations occurred when, for a miniature object, DDMAN selected an action that corresponds to a full-sized object; or vice versa, when, for a full-sized object, DDMAN selected an action that corresponds to a miniature object. For instance, a scale error would occur when DDMAN selects an action associated with the full-sized object (i.e., sitting) for a miniature chair.

Results

We emulated the learning/developmental processes of action selection via the visual and semantic routes. First, we examined several different neural network architectures to determine: (1) whether any of the instantiations would produce scale errors; and if so, (2) whether scale errors depend on the specifications of the architecture used (e.g., do scale errors occur when there are fewer units in the visual module and associative systems?).

To answer our first question, we explored several possible architectures, varying the number of layers (1- or 2-layers) and the number of units in the visual module (250 or 500 units) as well as in the associative systems (500, 1000, or 2000 units). Each architecture was initialized with small random weights and trained for the period of 100 epochs; this process was repeated 10 times resulting in 10 computational models per architecture. Then, we examined the action error rate and the proportion of scale errors within action errors of the models produced after the training process was completed. All actions that were incorrectly selected, for instance, the action "sitting" selected for an umbrella or for a miniature chair (i.e., a scale error) were counted as action errors. Overall, instantiations of most architectures achieved good convergence as their action error rate was lower than 20%. Five instantiations of the architecture, which had 250 units in 2-layered visual module and 1000 units in the associative systems, failed to converge. Similarly, three instantiations of a similar architecture but with 500 units in the associative systems exceeded 20% action error rate. Hence, the results of these two architectures were discarded altogether in the subsequent analysis. Only one model of a 2-layered network with 500 units in each layer in the visual module and 2000 units in the associative system failed to converge, therefore, the

results of this model were removed from the cohort, leaving 9 computational models of this architecture for further analysis.

Overall, across all models, we observed on average 4.21% (*SD* = 4.56%) action errors, out of whom 31.1% (*SD* = 36.75%) were scale errors; the proportion of scale errors was higher than the "chance level" of committing scale errors (i.e., 13.56%)[1]. The least error-prone models were those with 1-layer architecture with 500 units in visual module and 1000 units in the associative systems. The mean action error rate in the visual route was equal to 1.49% (*SD* = 0.43%) and in the semantic route to 1.36% (*SD* = 0.2%). The most error-prone models were those with 2-layer architectures with 500 units in each layer of visual module, and 2000 units in the associative systems. The mean action error rate in the visual route amounted to 9.16% (*SD* = 7.58%) and in the semantic route to 7.43% (*SD* = 5.85%).

Subsequently, we examined how the action error rates and the proportions of scale errors changed during training. Figure 4 illustrates the mean action error rate (left side) and, out of whom, the mean proportion of scale errors (right side) produced in the visual and semantic routes during 100 epochs. The mean action error rates are initially large due to the models being untrained; these action error rates rapidly decrease over training and drop below 20% by epoch 20. Similarly, the proportions of scale errors are large at the beginning of training (epochs 1 – 20) across all considered architectures, and as in case of action errors, decrease over time.

[1] To estimate the "chance level" of committing scale errors, we first initialized all tested architectures with small random values drawn from a zero-mean Gaussian distribution with a STD of 0.1, and subsequently examined the number of action errors and proportions of scale errors that such untrained networks produce. On average, the untrained networks produced 88.09% (SD = 3.98%) action errors and 13.56% (SD = 5.5%) of these errors were scale errors. Hence, we assumed 13.56% as the "chance level" of committing scale errors in our simulations.

An ANOVA with the action error rates and the proportions of scale errors as DVs, and Epoch (i.e. 3 vs 100) and Route (i.e. visual vs semantic) as IVs yielded significant effects of Epoch. The action error rates at the beginning of training (epoch 3) (M = 44.69%, SD = 6.05%) were significantly higher than the action error rates after training (epoch 100) (M = 4.21%, SD = 4.56) ($F(1, 230) = 3317.05$, p <.000, η^2=.935), which was expected since we only retained converging models. Importantly, the proportions of scale errors produced at the beginning of training (M = 93.53%, SD = 4.07%) were significantly higher than the proportions of scale errors after training (M = 31.09%, SD = 36.75%) ($F(1, 230) = 330.81$, p = <.000, η^2=.59). High proportions of scale errors at the beginning of training demonstrate that most action errors were scale errors. These results also indicate that scale errors decrease linearly over training, and do not follow an inverted U-shape curve.

To examine the effects of naming on scale errors we compared the action error rates and the proportions of scale errors produced by direct (visual) and indirect (semantic) routes. The action error rates produced in the visual route did not differ from the action error rates produced in the semantic route ($F(1, 230) = .55$, p = .46, n.s.). Similarly, the proportions of scale errors did not differ between the two routes ($F(1, 230) = .05$, p = .82, n.s.). The interaction between Epoch and Route was also not significant for the action error rates as well as for the proportions of scale errors, demonstrating that the error patterns produced by these two routes are similar.

Next, we examined the emerging neural representations to determine possible origins of scale errors. It could be that these representations initially collapse over size, leading to similar neural activations when a full-sized and a miniature object are presented as input. For this analysis, we recorded the activations of units in the visual module and associative systems for all test objects presented as input at the beginning

of training (epoch 3) and after training (epoch 100). We selected a computational model that resulted in the lowest action selection error after training (i.e., 1-layer architecture with 500 units in the visual module and 1000 units in the associative systems). To visualize the activation patterns in the visual module, first we used Principal Component Analysis (PCA) dimensionality reduction to reduce the 500 features (i.e., each corresponding to activations of one unit) to 30 dimensions, and then t-Distributed Stochastic Neighbor Embedding (T-SNE) (van der Maaten & Hinton, 2008) to further project these into a 3D space. Results, illustrated in Figure 5a, showed that the activations, already at the beginning of training, were organized largely by shape, without a clear distinction of object size. Interestingly, after training the activations remained organized by shape. A similar approach was used to visualize the activations in the sensorimotor system. Again, as Figure 5b illustrates these were largely organized by shape at the beginning and after training. If object representations are dominated by shape, how did DDMAN learn over time to select appropriate actions for differently sized objects?

To answer this question, we calculated the Euclidean distance between the neural activations for the full-sized and miniature test objects per each shape (i.e., chair, car, motorbike, and umbrella). As it can be seen in Figure 6, these distances were smaller at the beginning of training when DDMAN produced many scale errors. Over training epochs, however, these distances became larger, indicating that more nuanced size information had been learned.

Discussion

The primary aim of this study was to explore the possible reasons of scale error production in the action selection processes in children. We based our computational model (DDMAN) on the dual-route of action selection theory, and emulated the

developmental/learning process of action selection via direct visual route and indirect semantic route. We examined several architectures, varying the number of layers and the number of units in the visual module and the associative systems. DDMAN produced scale errors across all successfully trained network architectures: on average 31.1% out of 4.21% of all action errors were scale errors. Importantly, the largest proportion of scale errors occurred at the beginning of training (Epoch 3), and amounted to 93.53% (out of 44.69% of action errors). In addition, scale errors were observed both in the visual and semantic routes, across all considered architectures, demonstrating the robustness of the observed phenomenon in our simulations.

Inspection of emerging object representations in the visual module and the associative systems revealed that these were coarsely organized by shape; hence, during initial action selection, shape outweighs size information, leading to the selection of actions based on object shape rather than size (scale errors). With experience, even though such shape-dominated organisation did not change much over training, DDMAN gradually learned to pay attention to object size and selected actions considering both object shape and size. Overall, our computer simulations demonstrate that scale errors are produced during the process of learning to act on objects.

The second goal of our study was to determine the shape of the function of scale errors with age (here training epochs), as the literature reports inverted-U shape curves with a peak at 20 – 24 (DeLoache et al., 2004; Ware, Uttal, & DeLoache, 2010), while others showed a linear decrease from infancy to 40 months (Brownell et al., 2007; Grzyb et al., 2018; Rosengren et al., 2009). In our simulations, scale errors decreased linearly during training across all considered architectures and no inverted-U shape development was observed. As discussed earlier, the linear decrease of scale errors with age suggests that development of several concurrent abilities (object

representation, body knowledge, inhibitory control, action selection etc) may contribute to the prevalence of scale errors in young children. The results of our computer simulations highlight the role that object representation and action selection system play in children's scale errors – although other possible factors such as body knowledge and inhibitory control, that were not represented in our model, could contribute to the generation of scale errors.

Although the number of action errors, and the number of scale errors decreased linearly over training, these did not disappear completely. After training, the computational models (across all tested architectures) sporadically produced action errors, including scale errors. This result, again, is in line with the empirical results with adults showing that action mistakes (called also "action slips") occasionally happen in everyday life (e.g., Reason, 1984). Examples include an adult using an air freshener as a hairspray, where an inappropriate action (or inappropriate object) is selected because of the high similarity in object shapes. Similarly, Casler et al. (2014) demonstrated that adults sometimes make scale errors with tools.

The third aim of our simulations was to examine the influence of object naming on the number of scale errors. The acquisition of object names has important consequences for the children's conception of objects, as it influences the way they categorize and individuate objects (e.g., Balaban & Waxman, 1997). Based on the results by Hunley and Hahn (2016), as well as Oláh et al. (2016), who showed effects of object labelling on scale errors, and by Grzyb et al. (2018) who showed that early talkers tend to produce more errors than late talkers, we expected DDMAN to produce more scale errors when actions were selected via the indirect semantic route (which incorporates a link between object name and action) as compared to when actions were selected via the direct visual route. However, this was not confirmed by

the results from our simulation, for two possible reasons. First, object naming may not only operate at the level of associative memory (modelled here), but also may be important for perceptual processing, for instance, by directing the attention to object shape. This is illustrated in the well-established 'shape bias', that is, children's tendency to extend the object name to new exemplars based on their shape (Landau, Smith, Jones, 1988; Samuelson & Smith, 2003; Smith, 2003; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). However, with the type of artificial neural network we used we could not observe the influence of object name learning on the processing of object features in the visual module. As described earlier, learning in deep belief networks is carried out in stages with the lowest layers trained first and the top layers trained last. Therefore once the layers of the visual module were trained, they did not change during training of the lexico-semantic system. In other words, emerging object-name associations during training of the lexico-semantic system could not alter already existing object representations in the visual module. Yet we can speculate that, if object naming could influence the visual module during training, this would cause the emerging object representations to be even more strongly organized by shape, as similarly shaped object would be given the same name. That would cause stronger associations between objects and actions based on object shape, making it even harder for the action selection system to pick up the size information, therefore, contributing to a larger number of scale errors.

The second possibility for not observing any object naming effects on scale errors, might be simply because in this study we used only four object categories/names. To obtain an effect of naming on the prevalence of scale errors, a much larger vocabulary might be needed, first because at the age where scale errors can be observed, that is from the age of 18 months, children have already acquired a large vocabulary (e.g.

Fenson, Marchman, Thal, Dale, & Reznick, 2007). Second the fact that early talkers produce more scale errors than late talkers (Grzyb et al., 2018) indicates that the effect may be due to a quantitative change in word acquisition, rather than a qualitative change. Our model was designed to evaluate the effect of qualitative changes - that is, what changes when an object is linked to a name - and not for evaluating the effect of quantitative changes - that is, what changes when many words are learned versus few. It will be for future computational studies to examine the role of vocabulary growth in action errors more closely.

The results of our simulations provide important insights into the existing explanations of scale errors. As mentioned earlier, according to DeLoache et al. (2004), scale errors result from children's failures in the integration of visual information processed by the two visual streams coupled with a failure in inhibitory control. In this view, whenever a child encounters a miniature object from a highly familiar category, the visual features of the object (e.g., its shape, color, texture) activate the representation of the general category of the object, which in turn, activates the action that is linked with this object category. For instance, seeing a miniature chair activates the child's representation of the general category of typical chairs, which is linked to an action for interacting with the full-sized chair. Unsurprisingly, this is what we observe in our computer simulations: the visual input from a miniature chair activates the representation of chairs and the action that is associated with the full-sized chair. More interestingly, inspection of the developing object representations in DDMAN revealed the origins of such mistakes: these representations form clusters that are organized by shape while largely collapsing size (Figure 5). The action selection system, therefore, has to gradually learn to attend to a more fine-grained size information. This indicates that the origins of scale errors lies

in an incomplete representation of objects, where size if not fully represented; in other words, children do not fail to perceive size per se, but they ignore it once the object has been identified.

The results of our computer simulations are more difficult to reconcile with the teleofunctional explanation put forth by Casler et al. (2011). In this view, children are teleofunctional thinkers who rapidly form tight object-function pairings; these pairings constitute the basic blocks of the semantic knowledge used for action selection. Scale errors occur because the sight of an object automatically elicits a function associated with that object, despite the object having inappropriate size to be effective. However, in our model function and names were confounded, as it is actually the case for most object names (Landau, Smith, & Jones, 1998): objects that have the same function tend to have the same name. Casler et al.'s teleofunctional explanation would be supported if the model produced more scale errors when function (here, name) is clamped onto the objects. However the proportion of scale errors was similar whether we interrogated the direct visual route or the indirect lexico-semantic route. That does not discard the possibility that the functional link plays a role in the production of scale errors, but perhaps points to another limitation of our model. In future simulations it might prove useful to distinguish between name and function in order to evaluate more directly Casler et al.'s explanation for scale errors.

Overall, the results of our simulations suggest that scale errors are caused by emerging object representations that are largely organized by shape, and by the immature action planning system that initially selects actions based on object shape rather than size.

References

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64*, 3 – 26.

Brownell, C., Zerwas, S., Ramani, G. B. (2007). "So big": the development of body self-awareness in toddlers. *Child Development, 78*(5), 1426–1440.

Casler, K., Eshleman, A., Greene, K., Terziyan, T. (2011). Children's scale errors with tools. *Developmental Psychology*, *47*(3), 857 - 866.

Casler, K., Hoffman, K., & Eshleman, A. (2014). Do adults make scale errors too? How function sometimes trumps size, *Journal of Experimental Psychology*: *General*, *143*(4), 1690-1700. doi: 10.1037/a0036309.

DeLoache, J. S., LoBue, V., Vanderborght, M., Chiong, C. (2013). On the validity and robustness of the scale error phenomenon in early childhood. *Infant Behavior and Development*, *36*(1), 63 – 70. doi: 10.1016/j.infbeh.2012.10.007

DeLoache, J. S., Uttal, D. H., Rosengren, K. S. (2004). Scale errors offer evidence for a perception-action dissociation early in life. *Science, 304* (5673), 1027 - 1029.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., & Reznick, J. S. (2007). *MacArthur-Bates communicative development inventories: User's guide and technical manual*. Baltimore, MD: Brookes.

Freund, Y. & Haussler, D. (1992). Unsupervised learning of distributions on binary vectors using two layer networks. In *Advances in Neural Information Processing Systems 4*, pp. 912-919, San Mateo, CA. Morgan Kaufmann.

Glover, S. (2004). What causes scale errors in children? *Trends in Cognitive Sciences*, *8*(10), 440–442.

Grzyb, B.J., Cangelosi, A., Cattani, A., & Floccia, F., *Children's scale errors: A by-product of lexical development? Developmental Science,* doi: 10.1111/desc.12741

Hunley, S. B., Hahn, E. R. (2016). Labels affect preschooler's tool-based scale errors. *Journal of Experimental Child Psychology*. http://dx.doi.org/10.1016/j.jecp.2016.01.007

Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*(8), 1711-1800.

Hinton, G.E., Osidero, S., & Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527 – 1554.

Landau, B., Smith, L., & Jones, S. (1998). Object shape, object function, and object name. *Journal of memory and language*, *38*(1), 1-27.

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*, 436 – 444.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579 – 2605.

Mandler, J.M. & DeLoache J. (2012). *The beginning of conceptual representation*. In S. Pauen (Ed) Early childhood development and later outcome (pp. 9-32). Cambridge: Cambridge University Press.

Milner, A.D., & Goodale, M.A. (1995). *The visual brain in action*. Oxford: Oxford University Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746 – 748.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. (2011). Multimodal deep learning. In *Proceedings of International Conference on Machine Learning,*

Washington.

van Noord, N., & Postma, E. (2017). Learning scale-variant and scale-invariant features for deep image classification. *Computer Vision and Pattern Recognition*, 61, 583 – 592.

Oláh, K., Elekes, F., Pető, R., Peres, K., & Király, I. (2016). 3-Year-Old Children Selectively Generalize Object Functions Following a Demonstration from a Linguistic In-group Member: Evidence from the Phenomenon of Scale Error. *Frontiers in Psychology*, 7:963.

Reason, J.T. (1984) Lapses of attention in everyday life. In W. Parasuraman & R. Davies (Eds.), *Varieties of attention* (pp. 515 – 549). Orlando, FL: Academic Press.

Riddoch, M. J, Humphreys, G. W, & Price, C. J. (1989). Routes to action: Evidence from apraxia. *Cognitive Neuropsychology*, *6*, 437 – 454.

Rizzolatti, G., & Fadiga, L. (1998). Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). In *Sensory Guidance of Movement, Novartis Foundation Symposium 218*, eds G. R. Bock, & J. A. Goode (Chichester: John Wiley and Sons), 81 – 103.

Rizolatti, G., & Craighero, L. (2004). The mirror neuron system. *Annu. Rev. Physiol. 27*, 169 – 192. doi: 10.1146/annurev.neuro.27.070203.144230

Rosengren, K. S., Carmichael, C., Schein, S. S., Anderson, K. N., Gutierrez, I. T. (2009). A method for eliciting scale errors in preschool classrooms. *Infant Behavior and Development*, *32*, 286 – 290.

Rosengren, K. S., Gutierrez, I. T., Anderson, K. N., & Schein, S. (2009). Parental reports of children's scale errors in everyday life. *Child Development*, *80*(6), 1586 - 1591. doi: 10.1111/j.1467-8624.2009.01355.x

Samuelson, L. K., & Smith, L. B. (2000). Children's attention to rigid and deformable shape in naming and non-naming tasks. *Child Development, 71*(6), 1555-1570.

Smith, L. B. (2003). Learning to recognize objects, *Psychological Science*, *14*(3), 244 – 250.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science, 13*(1):13–19.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In Rumelhart, D.E., & McCelland, J.L. (Eds.), *Parallel Distributed Processing, 1*(6), pp. 194 – 281. MIT Press, Cambridge.

Srivastava, N., Salakhutdinov, R. (2014). Multimodal learning with Deep Boltzmann Machines. *Journal of Machine Learning Research*, *15*, 2949 – 2980.

Thill, S., Caligiore, D., Borghi, A. M., Ziemke, T., & Baldassarre, G. (2013). Theories and computational models of affordance and mirror systems: an integrative review. *Neurosci. Biobehav. Rev. 37*, 491 – 521. doi: 10.1016/j.neubiorev.2013.01.012

Ware, E. A., Uttal, D. H., DeLoache, J. S. (2010). Everyday scale errors. *Developmental Science*, *13*(1), 28 – 26. doi: 10.1111/j.1467-7687.2009.00853.x.

Ware, E. A., Uttal, D. H., Wetter, E. K., DeLoache. J. S. (2006). Young children make scale errors when playing with dolls. *Developmental Science, 9*(1), 40 - 45.

Welder, A. N., Graham, S. A. (2001). The influence of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties.

Child Development, *72*, 1653 – 163.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*(3), 223 – 250.

Xu, F., Cote, M., & Baker, A. (2005). Labeling guides objects individuation in 12-month-old infants. *Psychological Science*, *16*(5), 372 – 377.

Xu, Y., Xiao, T., Zhang, J., Yang, K., & Zhang, Z., Scale-invariant convolutional neural networks, arXiv:1411.6369.

Yoon, E. Y., Dietmar, D., Heinke, & Humphreys, G. W. (2002) Modelling direct perceptual constraints on action selection: The naming and action model (NAM). *Visual Cognition*, 9, 615 – 661.

Zorzi, M., Testolin, A., & Stoianov, I. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in Psychology*, *4*(515).

Figure captions

Figure 1. The architecture of the Developmental Deep Model of Action and Naming (DDMAN).

Figure 2. Sample pixel images used for training and testing the DDMAN; images from four object categories (i.e., chair, car, motorbike, umbrella) were extracted from CalTech 101 silhouettes dataset.

Figure 3. Schematic explanation of training and test processes of DDMAN.

Figure 4. Mean proportion of action errors and mean proportion of scale errors within action errors produced in the direct visual route and the indirect semantic routes of DDMAN, initialized 10 times with random small weights and trained over 100 epochs. The number of layers in the visual module (1 or 2 layers) and the number of units in the associative systems was varied across simulations.

Figure 5. Activation patterns in the computational model (500 units in the visual module, and 1000 units in the sensorimotor and lexico-semantic systems) for all test objects at the beginning of training (epoch 3) and after training (epoch 100) in (a) the visual module and (b) the sensorimotor system. Activations were projected in 3D space using T-SNE dimensionality reduction algorithm.

Figure 6. Mean Euclidean distance per object category between the activations of units in the sensorimotor system of the computational model (500 units in the visual module, and 1000 units in the sensorimotor system) for full-sized and miniature test objects.

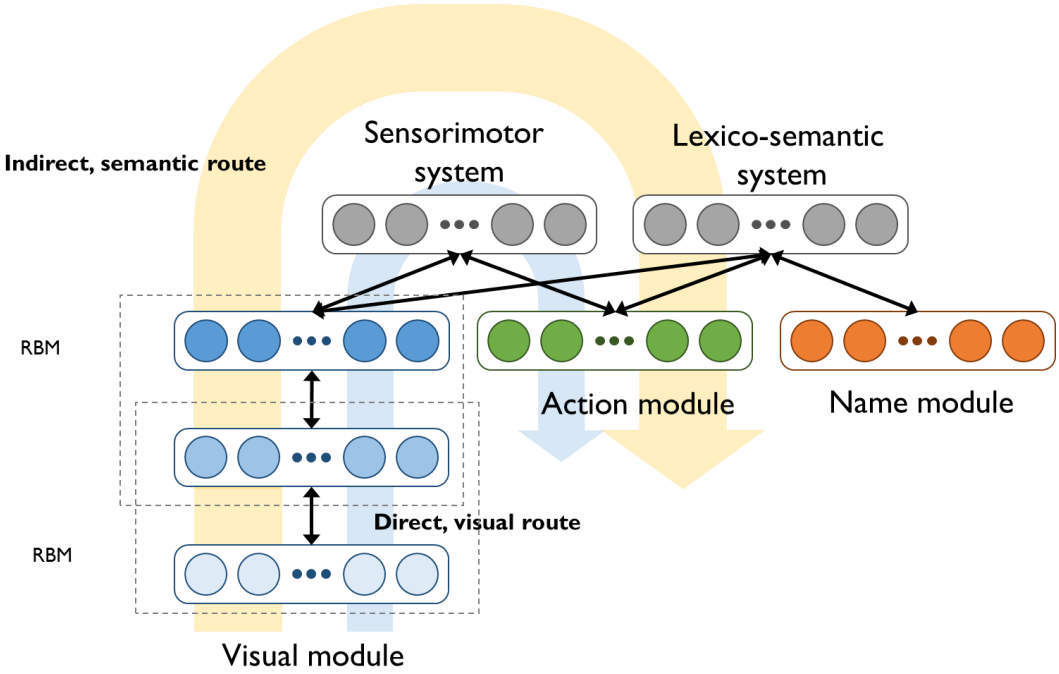
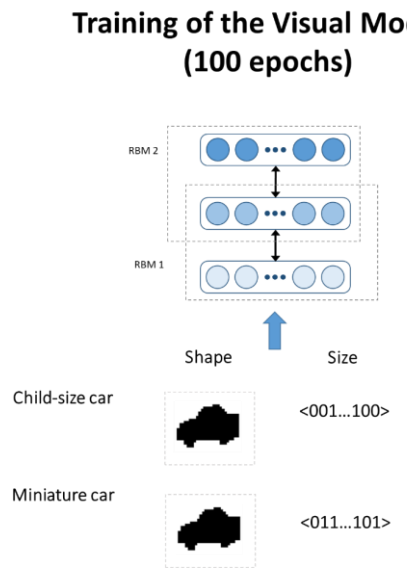

Indirect, semantic route
Sensorimotor system
Lexico-semantic system
RBM
Action module
Name module
Direct, visual route
RBM
Visual module

Figure 1

Figure 2

(a)


Training of the Visual Module
(100 epochs)
RBM 2
RBM 1
Shape
Size
Child-size car
<001…100>
Miniature car
<011…101>

(b)

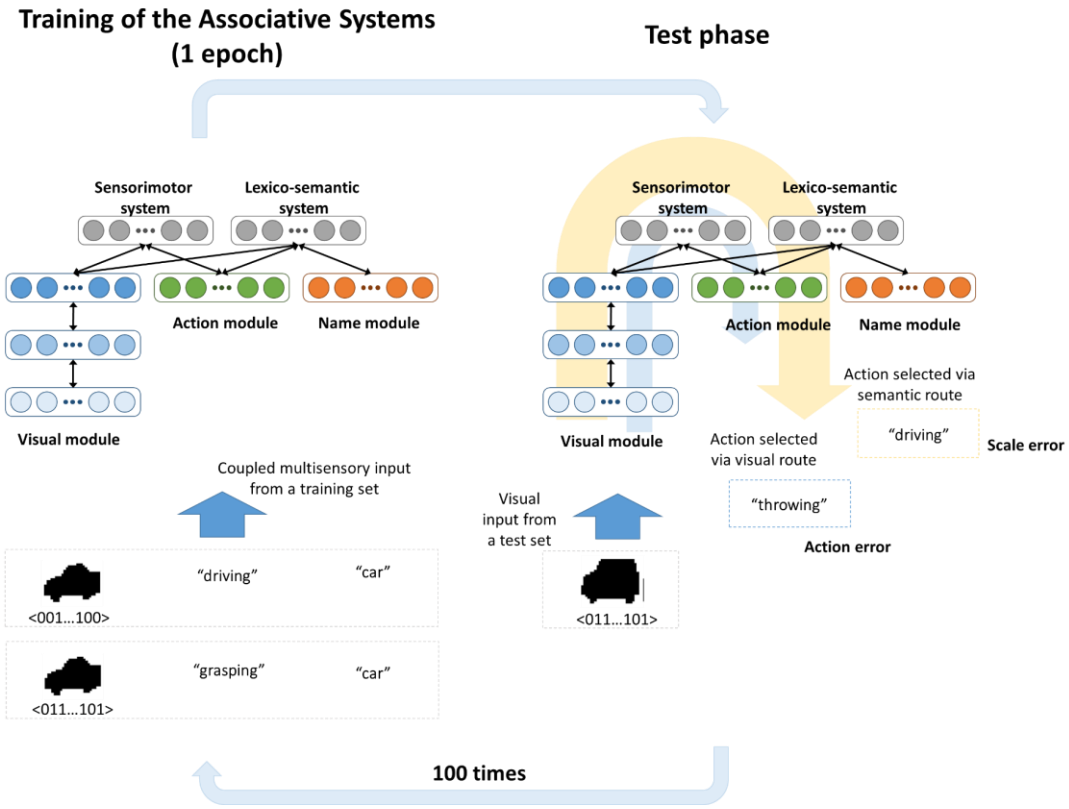

Training of the Associative Systems
(1 epoch)
Test phase
Sensorimotor system
Lexico-semantic system
Action module
Name module
Visual module
Coupled multisensory input from a training set
"driving"
"car"
<001…100>
"grasping"
"car"
<011…101>
Visual input from a test set
Action selected via semantic route
"driving"
Scale error
Action selected via visual route
"throwing"
Action error
100 times

Figure 3

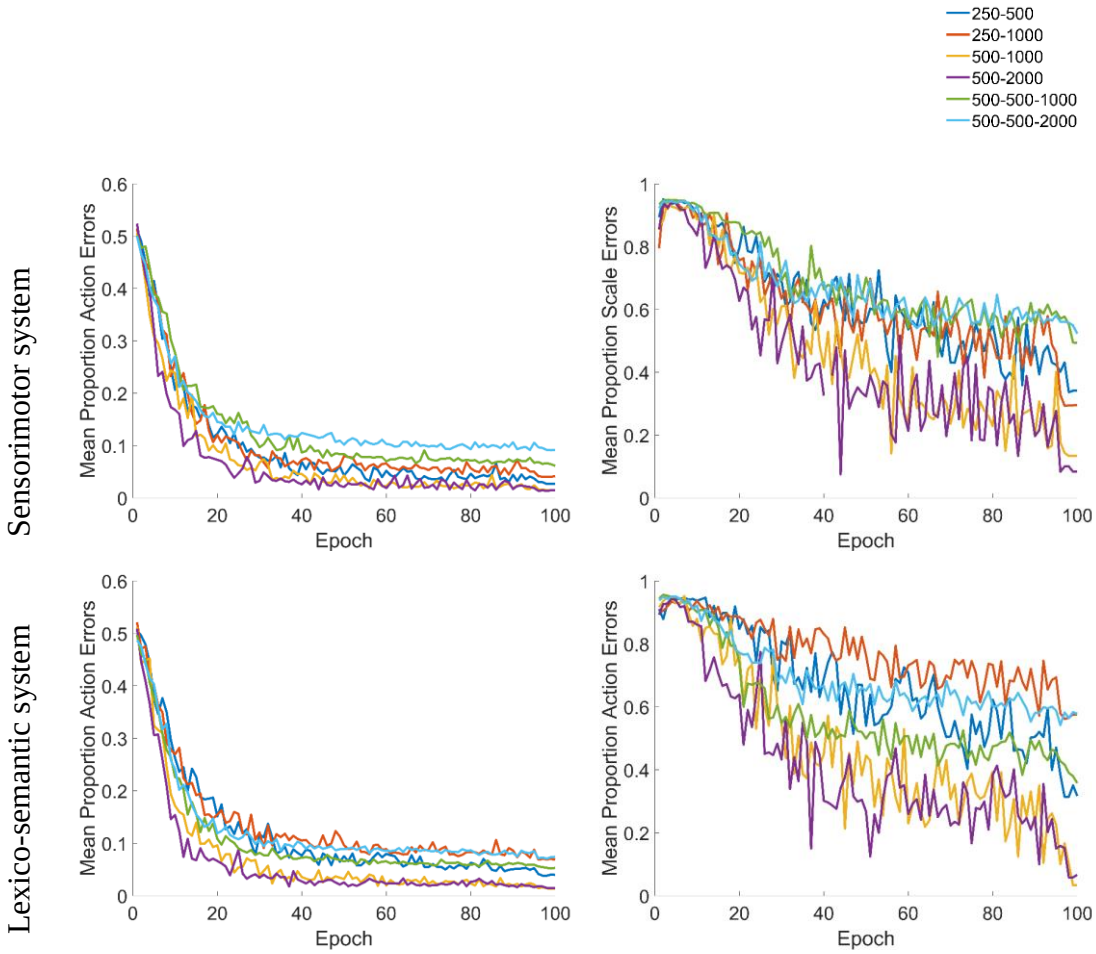
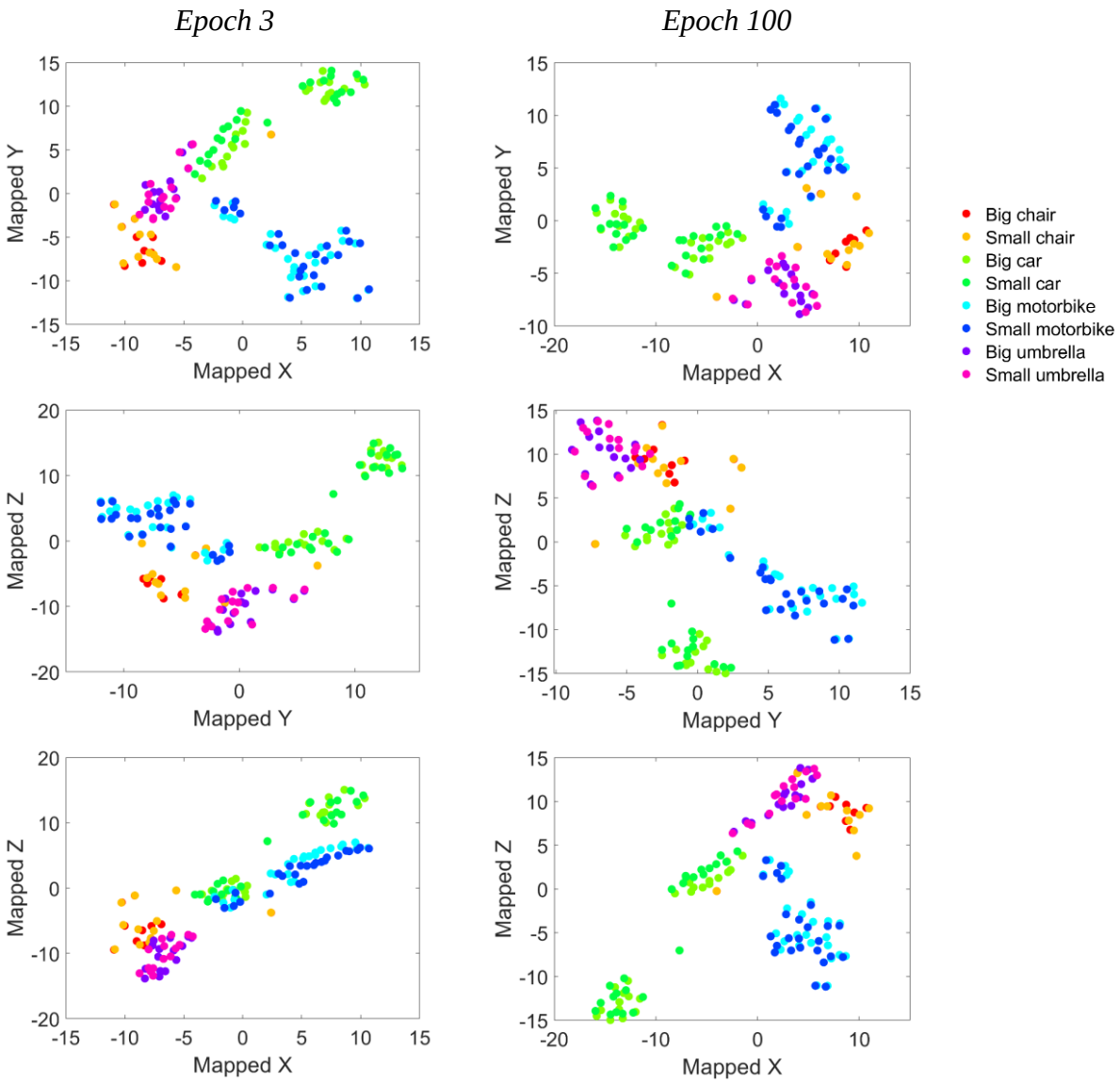

250-500
250-1000
500-1000
500-2000
500-500-1000
500-500-2000
Sensorimotor system
Lexico-semantic system
Mean Proportion Action Errors
Mean Proportion Scale Errors
Epoch

Figure 4

(a)

Visual module



(b)

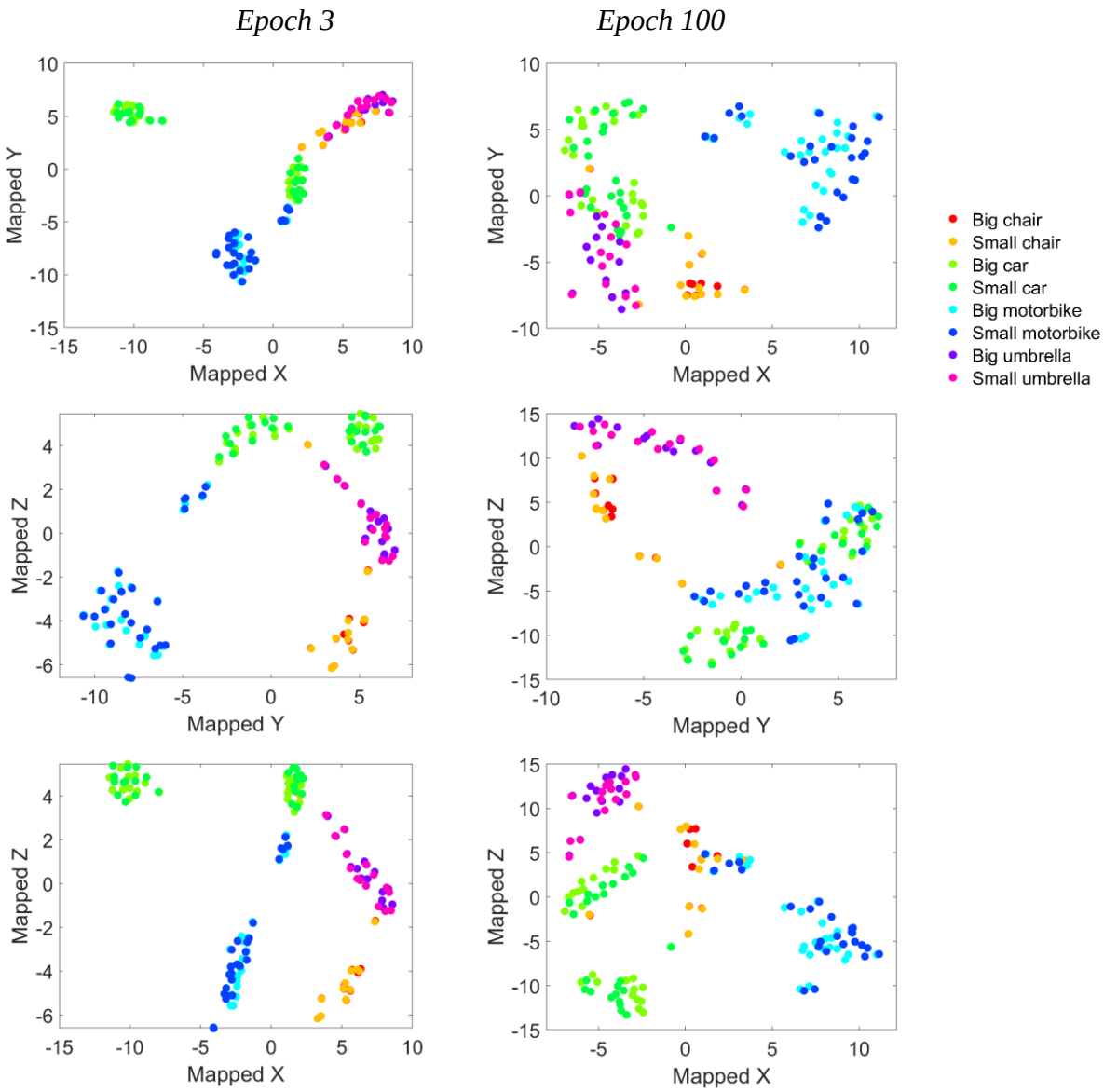

Sensorimotor System
Epoch 3
Epoch 100
Mapped Y
Mapped X
Mapped Z
Big chair
Small chair
Big car
Small car
Big motorbike
Small motorbike
Big umbrella
Small umbrella

Figure 5

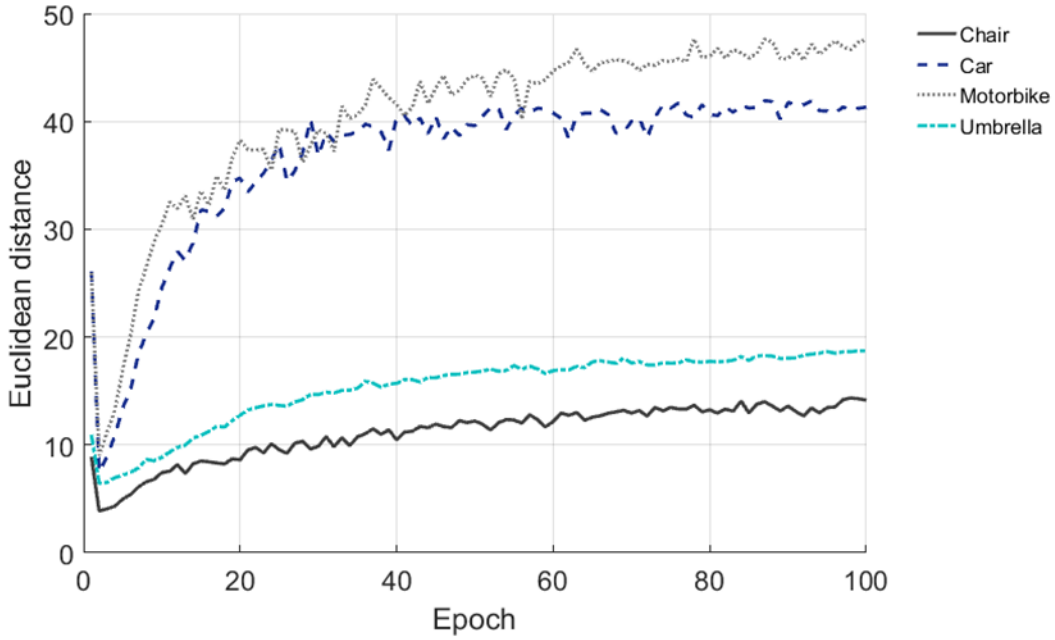

Euclidean distance
Epoch
0
10
20
30
40
50
0
20
40
60
80
100
Chair
Car
Motorbike
Umbrella

Figure 6