Mapping Language to Vision in a Real-World Robotic Scenario

Karla Štepánová, Frederico Belmonte Klein, Angelo Cangelosi, and Michal Vavrečka

***Abstract*—Language has evolved over centuries and was gradually enriched and improved. The question, how people find assignment between meanings and referents, remains unanswered. There are many of computational models based on the statistical co-occurrence of meaning-reference pairs. Unfortunately, these mapping strategies show poor performance in an environment with a higher number of objects or noise. Therefore, we propose a more robust noise-resistant algorithm. We tested the performance of this novel algorithm with simulated and physical iCub robots. We developed a testing scenario consisting of objects with varying visual properties presented to the robot accompanied by utterances describing the given object. The results suggest that the proposed mapping procedure is robust, resistant against noise and shows better performance than one-step mapping for all levels of noise in the linguistic input, as well as slower performance degradation with increasing noise. Furthermore, the proposed procedure increases the clustering accuracy of both modalities.**

***Index Terms*—Cognitive modeling, cross-situational learning, iCub robot, language acquisition, symbol grounding.**

I. Introduction

THE ESSENTIAL (and still not fully answered) question in language acquisition is how percepts are anchored in some arbitrary symbols, in other words, how words (symbols) get their meanings. This is the so-called symbol grounding problem [20]. For many years, cognitive modeling, neuroscience, psychology, and machine learning have jointly attempted to understand how humans can solve this "problem" [6]. The ability to learn language through perception and especially through visual grounding is not only important for understanding human cognition but is also applicable in many areas, such as verbal control of interactive robots [28], automatic sports commentators [15], car navigation systems, for the visually impaired, situated speech understanding in computer games [19], automated generation of weather forecasts [18], tutoring children in foreign languages [22], etc.

Despite the extensive research in the area of language acquisition, the question how the word-to-meaning mapping is learned remains unanswered. There are basically two approaches for mapping language to other modalities. A classic approach for separate modality-dependent representations is advocated by Barsalou [1]; while an example of a system where language is mapped to an intermediate modal representation that can be derived by multiple modalities is [27]. Using the unsupervised approach, Li *et al.* [24] designed the DevLex model, consisting of two self-organizing networks that are bidirectionally connected. Gliozzi *et al.* [17] proposed an alternative with a multimodal representation layer: their unsupervised feature-based model was used to account for early category formation in young infants. This approach postulates the unsupervised role of linguistic labels that can affect categorization during the acquisition process, which has also been supported by Taniguchi *et al.* [47]. Vavrečka and Farkaš [51] recently introduced a multimodal architecture for the grounding of spatial words using a biologically inspired approach (separate "what" and "where" visual subsystems) in which the visual scenes (two objects in 2-D space in a spatial relation) are associated with their linguistic descriptions, thus leading to the integration of modalities.

However, a fully unsupervised architecture, which would be able to deal with language grounding [46], particularly language grounding in a case where sentences have variable structure and when there is more than one object in a scene, is not available. The current state-of-the-art on variable-length sentences is very restricted and deals only with static scenes [29]. Most of the recent models based on deep networks are oriented toward application in image-to-text [21] or video-to-text [11] mapping and do not take into account the psychological aspects of language acquisition (e.g., mutual exclusivity). Moreover, these systems are trained in a supervised manner without the advantage of transfer learning.

The difficulty of the task was described in a well-known experiment performed by Quine [35] who imagined the anthropologist meeting a native who pointed at the scene and said "gavagai." When the anthropologist is stimulated in a situation by seeing a rabbit, he will suppose that the word represents the running rabbit in front of him, even though it could also mean "ground," "sun," "hello," or whatever else. This problem is related to language relativity, as there are several

Manuscript received June 27, 2017; revised December 6, 2017; accepted February 24, 2018. This work was supported in part by the European EU FP7 Research Project TRADR under Grant 609763, in part by TACR CAK under Grant TE01020197, in part by the CAPES Foundation, Ministry of Education of Brazil under Grant BEX 1084/13-5, in part by the CNPq Brazil under Grant 232590/2014-1, and in part by the U.K. EPSRC Project BABEL under Grant EP/J004561/1 and Grant EP/J00457X/1. *(Corresponding author: Karla Štepánová.)*

K. Štepánová and M. Vavrečka are with the Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, 16000 Prague, Czech Republic (e-mail: karla.stepanova@ciirc.cvut.cz).

F. B. Klein and A. Cangelosi are with the School of Computing, Electronics and Mathematics, Plymouth University, Plymouth PL4 8AA, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2819359

objects and their features that are described by words [33]. A simplified version of this problem consists of a simple visual scene and separate words that are grounded based on statistical co-occurrence (cross-situational learning).

From a neuroscientific point of view, symbol grounding can be viewed as a process of finding mappings between primary unimodal visual and language brain areas. Where exactly the integration is performed is still the subject of research, and existing literature provides only incomplete accounts of the cortical location of this convergence. For example, the study by Büchel *et al.* [4] provides evidence for the involvement of the left basal posterior temporal lobe (BA37) in the integration of language and visual information. Other studies (e.g., [43]) propose that access to verbal meaning depends on anterior and posterior heteromodal cortical systems within the temporal lobe. The grounding of actions and motoric primitives is associated with the activity in the dorsal stream and the premotor cortex [8].

How language can be developed in an unsupervised manner is also an important task in developmental robotics as most language acquisition in humans is fully unsupervised. One of the main long-term objectives of many teams worldwide is building conversational robots, which will be able to participate in cooperative tasks mediated by a natural language. It has been shown how robots can learn new symbols using already grounded ones and their combination [5] and how to transfer knowledge between agents [52]. Cangelosi *et al.* [5] presented their research on language emergence and grounding in sensorimotor agents and robots. This model was further extended by Tikhanoff *et al.* [49], who performed iCub simulation experiments and focused on the integration of speech and action. The grounding of higher-order concepts in action was also explored by Stramandinoli *et al.* [44], who made use of recurrent neural networks. Sugita and Tani [45] described an experiment dealing with semantic compositionality: the capability of a robot to use the compositional structure to generalize novel word combinations. Daoutis and Mavridis [9] summarized desiderata for grounded semantic compositionality. Despite all the progress in language grounding, however, the current state-of-the-art on grounding variable-length sentences is very restricted and deals only with static scenes [29], [38].

In this paper, we present this paper in the area of language acquisition using a real-world robotic scenario. We implemented a hierarchical cognitive architecture for language acquisition that includes visual and language processing. In particular, we chose to extend current models of cross-situational learning by allowing vision-to-language mapping in the case of a nonequal number of classes and by taking into account situation-time dynamics.

We show that this can be accomplished more efficiently by replacing one-shot mapping with sequential mapping and adding inhibitory mechanisms to the connections. The best mapped classes are gradually eliminated, and the clusterization is adaptively changed. We see this paper as an extension of the McMurray *et al.* [29] model, and we compare it with other single-step mapping models. The mapping strategy presented in this paper was shown to be very robust as it not only can find the mapping in circumstances of very noisy real-world input but also increases the clustering accuracy of both modalities. Recently, we tested the proposed algorithm also on the task of clustering body parts from simultaneous tactile and linguistic input [59]. In that case, sequential mapping showed slower degradation with increasing noise level in the linguistic input and outperformed one-step mapping for all dataset sizes and all levels of noise.

The rest of this paper is structured as follows. In Section II, we compare different mapping algorithms used in cross-situational learning. In particular, in Section II-B we provide a mathematical formulation of the newly proposed sequential mapping algorithm, and in Section II-C we describe the whole cognitive architecture which incorporates unimodal processing of vision and language and finding their association through the mapping algorithm. The performance of the proposed method on data from an iCub humanoid robot and from an iCub simulator is evaluated in Section III. Finally, results are discussed in Section IV with suggestions for future work.

II. Materials and Methods

In this section, we first present one-step and newly proposed sequential mapping algorithms (Sections II-A and II-B). Then, we describe in detail the whole cognitive architecture used to process data from individual modalities (vision and language). Finally, we describe the iCub robotic platform and the iCub simulator in Section II-D and provide a description of the evaluation in Section II-F.

A. One-Step Mapping in Cross-Situational Learning

One possible way how to establish mapping between visual concepts and linguistic elements is to use frequencies of referent and meaning co-occurrences; that is, the ones with the highest co-occurrence are mapped together [40], [41], [54]. This method is usually called cross-situational learning and supposes the availability of the ideal associative learner who can keep track and store all co-occurrences in all trials, internally memorizing and representing the word–object co-occurrence matrix of the input. This allows the learner to subsequently choose the most strongly associated referent [55], [56]. These models do not see the mapping as dynamic competition but operate only with the static state. Although some use likelihoods of different words and referents to perform Bayesian inference [16], [54], they do not take into account how the similarity of two-word forms can affect learning although it has been shown that the similarity affects learning in children [36]. Another shortcoming of these strategies is that they do not address how these similarities affect learning in a dynamic competition.

The simplest one-step word-to-referent learning algorithm simply accumulates word-referent pairs. This can be viewed as Hebbian learning: the connection between a word and an object is strengthened if the pair co-occurs in a trial. To extend this basic idea, we can enable also forgetting by introducing the parameter η, which can capture the memory decay. This so-called dumb associative-learning model was implemented by Yu and Smith [56]. Supposing that at each trial t we observe an object $\mathbf{o}_t^n$ and hear a corresponding word $\mathbf{w}_t^n$ (N_t possible associations), we can describe the update of the strength of the association between word model $L(i)$ and object model $K(j)$

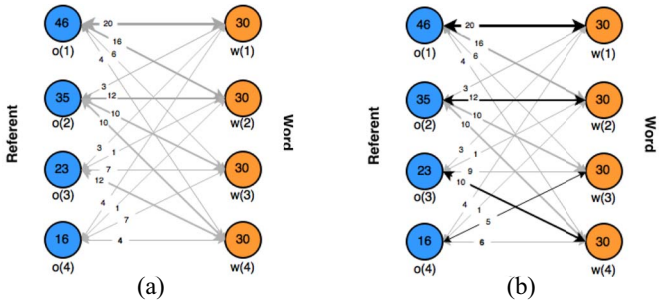


Fig. 1. One-step mapping. (a) In the first stage, weights between objects (referents) and words are changed using Hebbian learning. The connection between a word and an object is increased if the pair co-occurs in a trial [56]. (b) Then, word-to-referent mapping is found in one step: objects and words with the highest co-occurrence are mapped together. The number on each connection between word $w(i)$ and object $o(j)$ refers to the number of co-occurrences of $o(j)$ and $w(i)$. In this example, we suppose that there are 30 occurrences of each word.

as follows:

$$A(i,j) = \sum_{t=1}^{R} \eta(t) \sum_{n=1}^{N_t} \delta\left(\mathbf{w}_t^n, i\right) \delta\left(\mathbf{o}_t^n, j\right) \tag{1}$$

where R is the number of trials, δ is the Kronecker delta function (equal to 1 when both arguments are identical and 0 otherwise), $\mathbf{w}_t^n$ and $\mathbf{o}_t^n$ indicate the nth word–object association that the model attends to and attempts to learn in the trial t, and $\eta(t)$ is the parameter controlling the gain of the strength of the association.

Now let us assume that word $\mathbf{w}(i)$ is modeled by model L_i in the language domain and object (referent) $\mathbf{o}(j)$ is modeled by the model $K_{m(i)}$ in the visual domain. Our goal is to find the corresponding model $K_{m(i)}$ from the visual subdomain for each model L_i from the language domain to assign them together. Indices $m(i)$ are found as follows:

$$\forall i : m(i) = \underset{i}{\operatorname{argmax}} A(i,j) \tag{2}$$

where $\mathbf{A}$ is the co-occurrence matrix computed in (1) [element $A(i,j)$ captures co-occurrence between word $\mathbf{w}(i)$ and object $\mathbf{o}(j)$].

In Fig. 1, one-step mapping is visualized as it was implemented in this paper.

Modifications of the basic model include Regier's [36] work. He proposed a mapping model, which stems from competition models and incorporates two-way associations between words and referents. This enables the model to capture selective attention to individual words and referents, as well as to provide a probability distribution over associated referents/words. Regier [36] also showed that for his model, learning of a novel word is most effective when memory interference is minimized.

B. Proposed Sequential Mapping

Because we know that learning is not static but is a dynamic process, it seems reasonable to extend the basic idea of one-step cross-situational learning by incorporating dynamic competition mechanisms between words and referents in the model. To capture the dynamic competition among models, we extend the basic one-step mapping algorithm with the sequential addition of inhibitory connections.

In this case, the process of finding word-referent associations resembles Hebbian learning with inhibitory connections. Once the word is associated with a corresponding object (referent), links from this referent to other words are inhibited. This idea also corresponds to the fact that children prefer mapping where an object has only one label to multiple labels, the so-called mutual exclusivity bias [25]. The inhibitory mechanisms and situation-time dynamics were already partially included in the model of cross-situational learning proposed by McMurray *et al.* [29].

Even though our model shares some similarities with the model proposed by McMurray *et al.* [29], our model stems from different computational mechanisms. The proposed sequential mapping is able to capture nondiscrete assignment to individual clusters, as well as dynamic competition mechanisms. The first mechanism is incorporated into the model by considering likelihoods that the observed data were generated by a given model instead of 1/0 assignment to models. In this way, similarities of individual meanings and referents, as well as the likelihood of their recognition in each trial, is taken into account. The second mechanism (dynamic competition) facilitates the sequential mapping as the best-mapped classes are gradually justified with inhibitory connections to other classes (i.e., after a reliable assignment between a language and a tactile model is found, inhibitory connections among this tactile model and all other language models are added). Thanks to this mechanism, the mutual exclusivity principle (the fact that children prefer mapping where an object has only one label to multiple labels [25]) is guaranteed. The assignment between visual models K_j and language models L_i is found using the following iterative procedure.

1) Visual and language data are clustered separately and the corresponding posterior probabilities are found

$$p\left(L_i|\,\mathbf{w}_t^n\right) = \frac{p\left(\mathbf{w}_t^n|L_i\right)p(L_i)}{\sum_{i'} p\left(\mathbf{w}_t^n|\,L_{i'}\right)p(L_{i'})}, \quad \forall i \in \{1,\ldots,I\}$$
$$\forall t \in \{1,\ldots,T\}, \quad \forall n \in \{1,\ldots,N_t\} \tag{3}$$

$$p\left(K_j|\,\mathbf{o}_t^n\right) = \frac{p\left(\mathbf{o}_t^n|K_j\right)p\left(K_j\right)\mathbf{k}(j)}{\sum_{j'} p\left(\mathbf{w}_t^n|K_{j'}\right)p\left(K_{j'}\right)}, \quad \forall j \in \{1,\ldots,J\}$$
$$\forall t \in \{1,\ldots,T\}, \quad \forall n \in \{1,\ldots,N_t\} \tag{4}$$

where I is the number of language models, J is the number of visual models, T is the number of trials, and N_t is the number of possible object-word associations in trial t.

2) The most probable cluster to which the data point is assigned is selected (for each data point)

$$a_t^n = \underset{i}{\operatorname{argmax}}\, p\left(L_i|\,\mathbf{w}_t^n\right) \tag{5}$$

$$b_t^n = \underset{j}{\operatorname{argmax}}\, p\left(K_j|\,\mathbf{o}_t^n\right)$$
$$\forall t \in \{1,\ldots,T\}, \forall n \in \{1,\ldots,N_t\}. \tag{6}$$

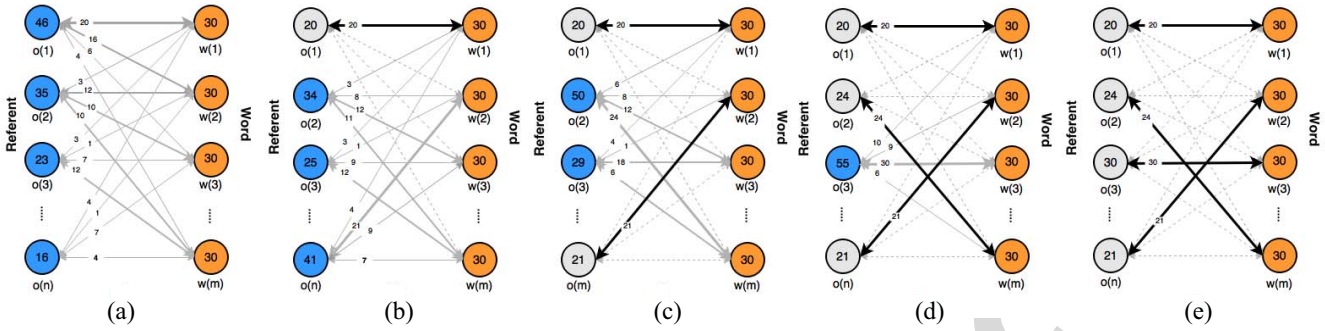


Fig. 2. Sequential mapping. The toy example of sequential mapping is shown to clarify the mechanism of finding the object-word assignment. In this example, we suppose that there are 30 occurrences of each word. The dotted line marks the inhibitory connection between object $o(j)$ and word $w(i)$ and the black line corresponds to the already found mapping. The number on each connection between word $w(j)$ and object $o(i)$ refers to the number of co-occurrences of $o(j)$ and $w(i)$. Objects $o(j)$ and words $w(i)$ are assigned to corresponding models based on the given clustering mechanism.

3) Co-occurrence matrix $A(i,j)$ is computed

$$A(i,j) = \zeta(i,j) \sum_{t=1}^{R} \eta(t) \sum_{n=1}^{N_t} \delta\left(a_t^n, i\right) \delta\left(b_t^n, j\right) \tag{7}$$

where R is the number of trials, $\zeta(i,j)$ is the matrix storing the strength of the connections between visual model K_j and language model L_i, and $\eta(t)$ is the parameter controlling the gain of the strength of association.

4) The best assignment is selected

$$[im, m(im)] = \underset{i}{\operatorname{argmax}} \underset{j}{\operatorname{argmax}} A(i,j). \tag{8}$$

In this step, the visual model $K_{m(im)}$ is assigned to the language model L_{im}.

5) Inhibition connections are added between the assigned visual model $K_{m(im)}$ and all language models other than L_{im} (mutual exclusivity)

$$\zeta(i, m(im)) = \zeta(i, m(im))(1 - z_1), \quad \forall i \neq im \tag{9}$$

where z_i is the parameter capturing the strength of the inhibition (in our experiment, this is set to 1, which corresponds to the total inhibition of the given connection).

6) Inhibition is added to the assigned visual model $K_{m(im)}$ (a prior probability of the model is changed)

$$k(m(im)) = k(m(im))(1 - z_2) \tag{10}$$

where z_2 is the parameter capturing the inhibition of the assigned visual model (in our experiment, this parameter is set to 1, which corresponds to total inhibition of the given model).

7) The assigned points are deleted from the dataset (data points which belong to model K_{jm} and L_{im})

$$X = X \backslash \left\{ \left(\mathbf{o}_t^n, \mathbf{w}_t^n\right) \mid \underset{j}{\operatorname{argmax}}\, p\left(K_j | \mathbf{o}_t^n\right) == m(im) \right. \tag{11}$$

$$\left. \wedge \underset{i}{\operatorname{argmax}}\, p\left(L_i | \mathbf{w}_j^n\right) == im \right\}. \tag{12}$$

8) If the dataset is not empty ($X \notin \emptyset$) or $\|\mathbf{k}\| > 0$ (some of the visual models are not totally inhibited), go to step 1, else stop.

The proposed algorithm where words are assigned to corresponding referents in a sequential manner is visualized in Fig. 2.

In the ideal case, unambiguous mapping between the two clusterizations will be found. In the real case (where the clusterization in visual and language layers is not optimal), none or more than one model from the visual layer will be assigned to one cluster L_i in the language layer or vice versa.

C. Specific Architecture

Our multimodal hierarchical architecture consists of multimodal and unimodal parts. The unimodal part has two layers performing separate processing of localist inputs: 1) visual objects and 2) auditory word-forms. Both unimodal layers are subsequently mapped one to each other in the upper multimodal layer (see Fig. 3).

1) Visual Layer: Each data point (object $\mathbf{o}_t^n$) can be considered as a triplet of continuous-valued vectors for each visual feature: $\mathbf{o}_t^n = (\mathbf{x}_{t,n}^{\text{size}}, \mathbf{x}_{t,n}^{\text{color}}, \mathbf{x}_{t,n}^{\text{shape}})$. This enables us to write the visual dataset as $X^{\text{vis}} = [X^{\text{size}} X^{\text{color}} X^{\text{shape}}]$ and process data for each visual feature separately. For processing visual data, the Gaussian mixture model (GMM) was used, which is a convex mixture of d-dimensional Gaussian densities $l(\mathbf{x}^k | \theta_j^k)$, where $k \in \{\text{size, color, shape}\}$. In this case, each visual model K_j^k is described by a set of parameters θ_j^k. The posterior probabilities $p(\theta_j^k | \mathbf{x}^k)$ are computed as follows:

$$p\left(\theta_j^k | \mathbf{x}^k\right) = \sum_{j=1}^{J_k} r_j^k l\left(\mathbf{x}^k | \theta_j^k\right) \tag{13}$$

$$l\left(\mathbf{x}^k | \theta_j^k\right) = \frac{1}{\sqrt{(2\pi)^d} \sqrt{|\mathbf{S}_j^k|}} \exp\left[-\frac{1}{2} \left(\mathbf{x}^k - \mathbf{m}_j^k\right)^T \left(\mathbf{S}_j^k\right)^{-1} \times \left(\mathbf{x}^k - \mathbf{m}_j^k\right) \right] \tag{14}$$

where $k \in \{\text{size, color, shape}\}$, $\mathbf{x}^k$ is a set of d-dimensional continuous-valued data vectors, r_j^k are the mixture weights, J_k is the number of visual models for each visual feature k, and parameters θ_j^k are cluster centers $\mathbf{m}_j^k$ and covariance matrices $\mathbf{S}_j^k$.

The Gaussian mixture is trained by the expectation-maximization algorithm [10]. An output of this layer for

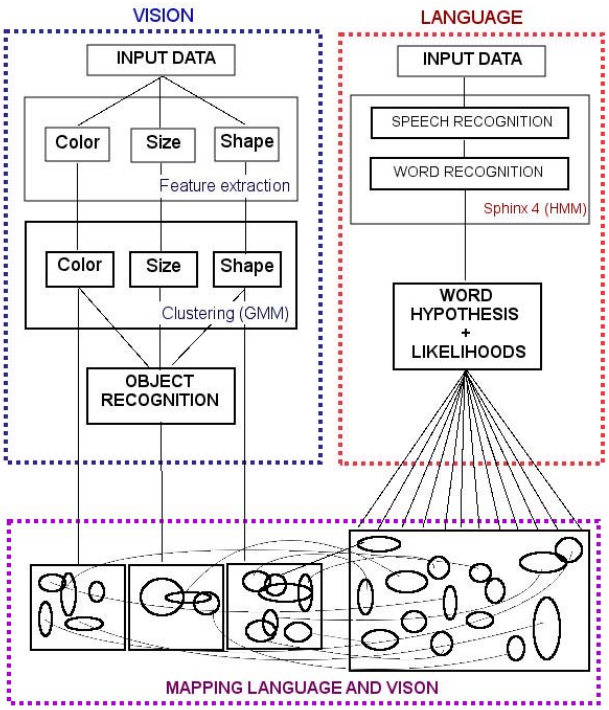


Fig. 3. Proposed multimodal architecture.

each data point $\mathbf{x}_i^k$ is the vector $\mathbf{y}_i^k$ of J_k output parameters describing the data point (the likelihood that the data point belongs to each individual cluster in a mixture). This corresponds to the fuzzy memberships (distributed representation).

For a simpler evaluation, we used a localist representation (winner-takes-all), where only the cluster with the highest cluster membership probability is considered for further processing [see (5) and (6)]

$$M\left(K_j^k|O\right) = \begin{cases} 1 & \text{if } j = \text{argmax}_{j'} f\left(K_{j'}^k|O\right) \\ 0 & \text{if } j \neq \text{argmax}_{j'} f\left(K_{j'}^k|O\right) \end{cases} \quad (15)$$

where $k \in \{\text{size, color, shape}\}$, $j \in \{1, \ldots, J_k\}$.

2) Language Layer: The auditory word-forms are extracted from the language input which are sentences describing the image in the format <size> <color> <shape> (e.g., "small red triangle"). Afterward, individual word-forms are extracted from the sentences and compared to prelearned language models, and the log-scale score $p(\mathbf{w}_t^n|L_i)$ of the audio matching the model is computed. Based on these data, posterior probability can be computed

$$p\left(L_i|\,\mathbf{w}_t^n\right) = \frac{p\left(\mathbf{w}_t^n|L_i\right)p(L_i)}{\sum_{i'} p\left(\mathbf{w}_t^n|L_{i'}\right)p(L_{i'})}, \quad \forall i \in \{1, \ldots, I\}$$
$$\forall t \in \{1, \ldots, T\}, \quad \forall n \in \{1, \ldots, N_t\} \quad (16)$$

where I is the number of language models, T is the number of trials (sentences) and N_t is the number of word-forms in the trial (sentence) t.

An output of this layer for each data point $\mathbf{w}_t^n$ is the vector $\mathbf{y}_i^k$ of I output parameters describing the data point (the likelihood that the data point belongs to each individual language model). This corresponds to the fuzzy memberships (distributed representation). Linguistic and visual inputs are processed simultaneously.

Algorithm 1 Sequential Mapping—Fixed Grammar

Inputs:
language clusters L_i ($i \in 1 : I$), visual clusters $K_j^k \sim N(\mathbf{m}_j^k, \mathbf{S}_j^k)$, $j \in 1 : J_k$, input data $\mathbf{x}^k$ for each feature $k \in \{\text{size, color, shape}\}$, number of clusters J^k for each feature k
Output:
mapping between all visual classes K_j^k and language classes L_i

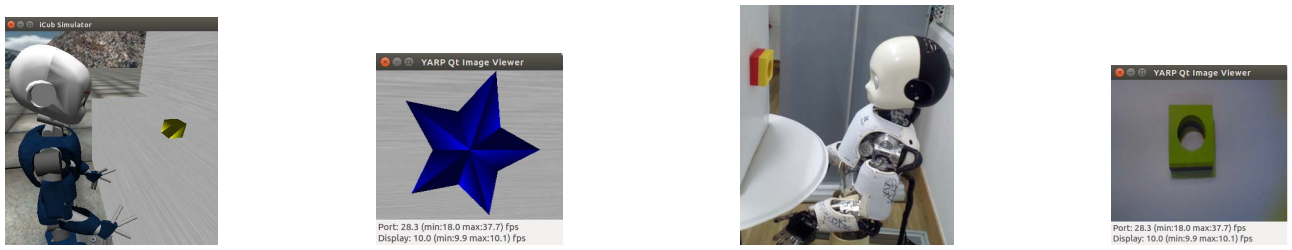
for $k \in \{\text{size, color, shape}\}$ **do**
  $NCl \leftarrow J^k$
  while $NCl > 0$ and $\mathbf{X}^k$ is not empty **do**
    assign each data point from $\mathbf{x}^k$ to visual and language cluster (Winner-takes-all, see Eq. (15))
    for $j = 1 : NCl$ **do**
      for $i = 1 : I$ **do**
        $A_{ij} \leftarrow$ how many times was class i classified as j
      end for
    end for
    $[im, jm] \leftarrow \text{argmax}_i \text{argmax}_j A_{ij}$
    $\mathbf{X}_{del}^k \leftarrow$ data points assigned to K_{jm}^k and L_{im}
    $\Theta_{new}^k \leftarrow N(\mathbf{x}_{del}^k)$ learn Gaussian on the to be deleted data
    $\mathbf{X}^k \leftarrow \mathbf{X^k} \setminus \mathbf{X_{del}^k}$ delete data assigned to both K_{jm}^k and L_{im}
    $NCl \leftarrow NCl - 1$
    relearn $K^k \sim N(\mathbf{m}^k, \mathbf{S}^k)$ on new data $\mathbf{x}^k$ with NCl clusters
  end while
end for
cluster visual data using new θ_{new}^k parameters (cluster centers $\mathbf{m}^k$, covariance matrices $\mathbf{S}^k$) and perform one-step mapping (Model 1)

3) Mapping—Models 1 and 2: After the visual and language data are clustered, the mapping between the two layers must be found. For each cluster L_i in the language layer, a corresponding cluster K_j^k in the visual layer (for each feature $k \in \{\text{size, color, shape}\}$) is found. The mapping is found as follows. For each j and k, we find cluster $L_{\text{kmax}_{jk}}$ from the language layer which will be assigned to cluster K_j^k from the visual layer. In this paper, we compare two different models to find indices $k\text{max}_{jk}$. We compared one-step mapping (see Section II-A) and newly proposed sequential mapping (see Section II-B).

The exact algorithm used to find the mapping between visual and language models in a sequential manner is described in detail in Algorithm 1. Indices $m(i)$ are found sequentially. In each step, the best-mapped data are excluded, and the rest of data are reclustered using GMMs. Then, one-step mapping is performed (see Algorithm 1). An extension of the algorithm for a variable-length sentence is described in the Appendix. Results for a variable length sentence using a fully artificial dataset with a controlled noise level are described in detail in [58].

D. iCub Robotic Platform and iCub Simulator

For the experiment, we used a simulated [48] and a physical stationary [30] iCub robot. The iCub [Fig. 1(c)] is an open-source humanoid robot the size of a three-and-a-half-year-old child, with the fully articulated hands and a head-and-eye system which makes him ideal for cognitive experiments. The iCub simulator has been designed to reproduce, as accurately as possible, the physics and dynamics of the robot and its environment [48]. The simulator and the actual robot have the same interface supporting YARP [31] which is a robot platform for interprocess communication and control of the physical and simulated robots in real-time.



(a) (b) (c) (d)

Fig. 4. Experiment design and corresponding input data. (a) iCub simulator. (b) Blender object. (c) Physical iCub. (d) Real object.

E. Input Data Description and Preprocessing

The input for our model consisted of visual and language data. The visual scene was composed of an object in the center of the scene with a variable position. The visual features (size, shape, and color) of the object also varied. We developed two separate datasets for training and testing purposes. A real-world dataset has visual sensory data acquired from the cameras of the physical iCub robot which observed simple objects placed on the white board in front of his eyes at a distance of 1 m [see Fig. 4(c) and (d)] (210 instances, three sizes, five colors, and seven shapes, in total 70 individual objects; each seen three times in slightly different placement). The simulated dataset is made in the iCub simulator [see Fig. 4(a) and (b)] as Blender-generated virtual objects (432 instances, three sizes, six colors, and six shapes, in total 108 individual objects; each seen four times in slightly different placement).

The spoken language input consisted of sentences pronounced by a non-native English speaker describing the image in the format <size> <color> <shape> (e.g., small red triangle) and was processed simultaneously with the visual input. In the real-world dataset, the tutor was talking to a robot from an approximate 2-m distance with natural background noise. The linguistic input was captured using an external microphone.

1) Speech Recognition: CMU Sphinx (an open-source flexible Markov model-based speech recognizer system) was used for speech recognition [23]. Sphinx itself offers a large vocabulary, but we created our own task-specific smaller vocabulary using the online IMtool that produces a dictionary based on a CMU dictionary and matches its language model.

There is a probabilistic output from the CMU Sphinx. The ten best hypotheses for a matching model with corresponding scores were saved for each utterance (they are log-scale scores of the audio matching the model). Because the scores for the hypothesis of each word in the sentence were needed for further evaluation, the words were pronounced with large pauses, and the end of the sentence was marked by the word "STOP." An output of the language layer is a I-dimensional continuous valued vector, where I is the number of language clusters (corresponding to the number of possible utterances). This vector contains ten nonzero values, and the rest is zero.

2) Image Processing: The image inputs are processed using standard MATLAB functions. First, the image is morphologically opened with a disk-shaped structuring element (*imopen*) to remove the noisy background of the image; then all grayish pixels are removed, and the image is converted from the true color RGB to the gray-scale intensity image by eliminating the hue and saturation information while retaining the luminance (*rgb2gray*). Finally, the intensity image is converted to a binary image using the threshold computed with Otsu's [34] method (*threshold*). An example of the preprocessed image is shown in Fig. 5.

Fig. 5. Image processing: original image, removal of the background, converting to BW image, and filling the holes.

Afterward, the properties of the image regions are measured using the function *regionprops*. Individual visual features (shape, color, and size) are subsequently processed separately. The following features were used: color (three features: average RGB of the selected region), size (six features: perimeter of an object, distance from the centroid to the left corner of the bounding box, and width and length of the bounding box), and shape (13 features: area, centroid, major axis length, eccentricity, orientation, convexArea, FilledArea, EulerNumber, EquivDiameter, solidity, extent, and perimeter). To obtain the shape features, we automatically cropped and resized the image to equalize the size of the objects.

Although the visual model is mainly mathematical and implemented in a very "machine vision" sort of way, the bases of its processing follow biological correlates of mammal vision. More specifically, from the neuroanatomical point of view, this corresponds to the processing of the visual input in the separate higher visual centers in the brain, specifically to the independent processing of the information about the position and identification of an object in the ventral (what) and dorsal (where) neural pathways, respectively [32]. Individual object properties are identified in the separate visual centers of the occipital lobe.

F. Evaluation

In the case of the supervised GMM, growing when required neural gas (GWR), and SOM algorithms, data were divided into training and validation datasets in the ratio 70:30. For the unsupervised GMM, hidden Markov model, and k-means algorithms, we computed the accuracy differently. After the data

TABLE I
COMPARISON OF CLUSTERIZATION AND CLASSIFICATION ACCURACY OF VISUAL DATA. THE MEAN AND STANDARD DEVIATION FROM 100 REPETITIONS ARE VISUALIZED

	Real data			Blender		
Accuracy [%]	Size	Color	Shape	Size	Color	Shape
GMM sup.	83.3 ± 0.0	99.0 ± 0.0	81.4 ± 0.0	98.6 ± 0.0	97.9 ± 0.0	93.1 ± 0.0
GMM unsup.	76.2 ± 6.8	76.1 ± 9.1	56.1 ± 6.2	74.2 ± 10.1	60.9 ± 9.0	64.3 ± 7.2
K-means	67.8 ± 6.2	81.2 ± 1.1	53.1 ± 4.2	66.3 ± 0.2	77.1 ± 10.7	72.8 ± 6.9
SOM	69.6 ± 5.6	78.9 ± 6.8	54.2 ± 4.1	66.1 ± 4.2	81.7 ± 7.6	59.3 ± 6.2
GWR	89.9 ± 2.1	99.5 ± 0.4	76.6 ± 1.4	88.9 ± 0.7	98.1 ± 0.9	94.2 ± 0.6

TABLE II
COMPARISON OF ONE-STEP MAPPING AND SEQUENTIAL MAPPING FOR DATA FROM THE ICUB SIMULATOR (BLENDER) AND THE PHYSICAL ICUB (REAL DATA). THE MEAN AND STANDARD DEVIATION FROM 100 REPETITIONS ARE VISUALIZED

	Real data			Blender		
Accuracy [%]	Size	Color	Shape	Size	Color	Shape
Vision	76.2 ± 6.8	76.1 ± 9.1	56.1 ± 6.2	74.2 ± 10.1	60.9 ± 9.0	64.3 ± 7.6
Language	70.6 ± 0.0	82.4 ± 0.0	77.5 ± 0.0	98.1 ± 0.0	96.5 ± 0.0	98.1 ± 0.0
One-step mapping	54.1 ± 4.1	58.2 ± 10.3	52.2 ± 4.9	67.3 ± 8.2	56.2 ± 6.1	61.9 ± 3.2
Sequential mapping	74.2 ± 15.1	87.1 ± 10.2	72.9 ± 5.1	96.1 ± 31.2	95.2 ± 1.2	92.1 ± 0.9

is clustered, each cluster is assigned to the class that appears most frequently in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned data points (compared to the manual true labels) and dividing this number by the total number of data points.

The accuracy of the learned mapping is calculated in the following manner. We cluster output activations from the visual layer and assign each data point to the most probable cluster. Then, we find indices *m(i)* for all clusters as defined in (2) for one-step mapping and (8) for sequential mapping. Based on this mapping, we can assign each data point to the language label. These language labels are subsequently compared to the ground truth. Accuracy is then computed as

$$\text{acc} = \frac{\text{TP}}{N} \tag{17}$$

where TP (true positive) is the number of correctly assigned data points, and N is the number of all data points.

III. EXPERIMENTAL RESULTS

The first part of the results is dedicated to the performance of the model in the real-world scenario. The robot interacts with a human in a noisy condition that distorts speech input.

A. Vision

In the first stage, we evaluated the Vision subpart of the model. Several algorithms were compared: namely the GMM algorithm, supervised GMM algorithm, k-means, SOM, and GWR algorithm [2], [26]. The SOM and GWR had 100 nodes. The results for the real-world dataset and the simulated dataset with Blender objects can be seen in Table I. Although the SOM and GWR algorithms are considered unsupervised algorithms, we adopted a technique for labeling inputs; thus, they should be compared with supervised algorithms. The number of clusters is also overestimated (the number of nodes corresponds to the number of clusters). It indicates that these algorithms partly overfit the data so we divided the set into testing and validation data.

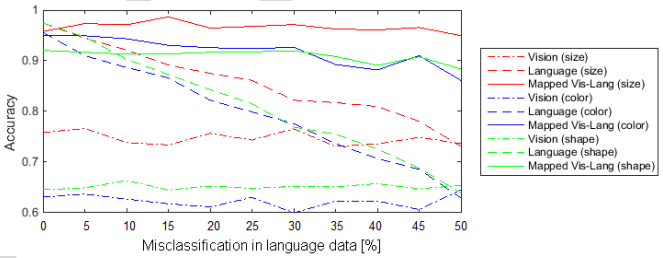


Fig. 6. Dependence of mapping accuracy on the misclassification in the language data for a fixed-length sentence (mean values over 50 repetitions are visualized). Different colors correspond to different visual features (red: size, blue: color, and green: shape). Visual data are generated in Blender and acquired through the iCub simulator, and language data are processed using Sphinx 4.

B. Mapping

The performance of one-step mapping (vision and language are mapped in one step based on the frequency of co-occurrence) and sequential mapping (see Algorithm 1) is shown in Table II. We calculated the accuracy for real-world data from physical iCub and for the Blender objects placed in the iCub simulator. Language accuracy for Blender dataset is much higher compared to the real-world data as a tutor was speaking directly into the microphone.

Tolerance of the sequential mapping to misclassification in the language data is visualized in Fig. 6 for visual data from the iCub simulator in combination with the language data processed by Sphinx 4. The misclassification is added to the language data subsequently and evenly to all classes (the given proportion of the language inputs was randomly changed to random words). The misclassification (a synthetic error) was added artificially but can be interpreted either as white noise added to the input data or mistakes in labeling perceived

objects. We grouped them together into a misclassification variable. The visual data are left intact so the only cause of the observed variations in the accuracy is initialization. As can be seen, the accuracy of sequential mapping remains very stable even though the accuracy of the language decreases and outperforms the language and vision for almost all values of the misclassification.

IV. Conclusion

Current models of vision-to-language mapping often make use of cross-situational learning while relying directly on statistical co-occurrence of meaning-referent pairs: the ones with the highest co-occurrence are mapped together (e.g., [41]). This approach shows inferior performance in cases where the clustering of data is difficult (e.g., due to the high number of objects, high noise level or overlapping classes, etc.). Therefore, we extended this basic model and introduced a new more robust and noise-resistant mapping procedure. Our approach incorporates situation-time dynamics and mutual exclusivity and is able to deal with a nonequal number of classes in individual subdomains.

Mathematical formulation of the newly proposed mapping is provided (see Section II-B), and results for simulated and real-world data from the iCub robot are compared to one-step mapping (see Table II). It was shown that the method is able to find mapping between language and vision, and the method improves the accuracy of both individual subdomains and shows very good resistance to noise or misclassification in language (see Fig. 6). How to map in an unsupervised manner several clusterings (e.g., for vision, action, and language) is an important question not only in cognitive modeling but also in general machine learning, where data acquired from different sensors or in different situations can be independently clustered and mapped to each other. A more detailed discussion of the results follows.

The trivial one-step mapping can be imagined as basic Hebbian learning. Our extension can be likened to Hebbian learning with inhibitory connections. In recent years, several approaches finding an alternative to the basic approach appeared [29], [57]. Our model partially stems from the McMurray *et al.* [29] approach, who showed that associative learning can be sufficient for language acquisition and that the main components of this type of learning are an online competition of models and pruning incorrect associations which enable gradual improvement of associations between models.

The mutual exclusivity principle is guaranteed in our method thanks to the inhibitory connections which are gradually created among models. Once the mapping between the referent and the meaning is found, the connection from a given meaning to other referents is inhibited. Dynamic competition is addressed in the following way: when any association of meaning and referent is found, other models compete again for resources. First, well-mapped data are deleted, and then, the resting data are reclustered. Furthermore, the likelihoods can associate each data point to many separate models instead of binary membership. In this way, our algorithm also emulates how the similarity of two-word forms can affect learning as it occurs in children's development [36]. The algorithm also enables mapping together data in a case where we have an uneven number of clusters in both subdomains.

However, the principle of mutual exclusivity is not suitable for further stages of language acquisition, namely, the learning of polysemic words. A polyseme is a word or phrase with different but related senses (e.g., wood as a piece of tree or the area with trees). The homonyms are a subset of the polysemes, but the difference between the homonyms and the polysemes is subtle and fuzzy. Homonyms represent a group of words that share similar spelling (homographs) and the same sound (homophones) but have different and unrelated meanings, for example, the homograph bank standing for embankment or place where money is kept. The learning of polysemic words violates the principle of mutual exclusivity as the dynamic competition does not allow to map one word to different visual models. We are aware of this problem, and we would like to extend it in the future iteration of the model. Similar to humans who have to learn polysemic words as an exception, we will incorporate this principle into the next version of our model.

In the next part, we analyze the ability of our architecture to deal with ambiguous inputs. The mapping will find reliable labeling for the visual input data (more generally for data from any other modality) with a possibility of incorporating the fuzziness of this mapping. For some concepts, finding an unambiguous mapping is very easy; for others, it is much more difficult or impossible (such as abstract words, e.g., the love has no dominant color, but the sky is usually blue). Since the mapping is established only among the clusters where it makes sense, dealing with a lot of redundant information is avoided. A similar idea is used in classification algorithms which use sparse matrices (e.g., [12] and [13]). We also analyzed the strong and weak points of the algorithms adopted in our architecture. First, the ability of different algorithms to classify unimodal visual data was compared. As expected, for data which are well separated and mainly spherically distributed (this is generally a case for simulated and artificially generated data), the k-means algorithm outperformed the GMM algorithm. However, for nonspherical real data, the GMM algorithm generally performed better (see Table I for the comparison of the performance on simulated data placed in an iCub simulator and data from real iCub robot cameras). As the unsupervised algorithms are highly dependent on the initialization, it can be seen that the standard deviation of the data is quite high even though 20 repetitions were averaged. We should conclude that the performance of the algorithms in our tasks reflects their fundamental advances and limitations. We are still missing the algorithm that is able to cope with highly variable datasets in terms of their statistical properties.

The most significant differences between data acquired from the real and the simulated iCub can be found in the visual subdomain. Affected by natural light and slightly altered points of view, the shading of objects and their bevel differed. This resulted in a different projection of objects to iCub cameras and in the less accurate classification of shapes for the real-world dataset. For colors, different lighting changed the true color of the objects. Nevertheless, as can be seen in Table II, this did not affect the ability of the algorithm to discriminate among individual colors. The real-world objects had a

variable size which resulted in a higher overlap of clusters for individual sizes compared to the Blender dataset.

Afterward, we focused on the mapping between vision and language and compared two different approaches: one-step mapping to sequential mapping which in a stepwise manner finds the best-mapped clusters while constantly relearning clusterization of visual data. As can be seen in the results in Table II, the novel sequential mapping led to an improvement in effectiveness compared to the method which maps vision to language in one step. This can be seen in more accurate mapping which leads to a better estimation of the clustered data labels and consequently to lower classification error for all of the evaluated datasets.

The accuracy of multimodal mapping outperforms vision, language or both. This is an important finding, because sequential mapping improves not only the accuracy of visual clustering but can also fix mistakes in language recognition, which provides the labels. Furthermore, this result suggests that we are able not only to find mapping between more clusterings, but we can also improve the clusterization accuracy by combining individual classifiers. This is not easily seen in the presented dataset as there is high recognition accuracy of Sphinx software (especially in the case of sentences recorded for the simulated dataset). Therefore, we also tested whether the misclassification (or noise) in the language data affects the correct mapping between vision and language. The result can be seen in Fig. 6. Although a synthetic error (misclassification) is added to the language data, the ability to find the mapping remains nearly intact. The mapping accuracy decreases only very slightly and remains at around 90% because the accuracy of language recognition drops from the original approximately 95% to approximately 70% (depending on the specific visual feature). In the future, we would like to explore the effect of noise in individual modalities on classification accuracy in more detail by adding a controlled amount of noise directly to the input. The error propagation from individual input modalities to the multimodal layer was explored, for example, in [37]. They showed that the visual context can steer speech recognition and vice versa.

We also analyzed how the complexity of environment affects the accuracy of the our architecture. We suppose that the performance of one-step mapping will decrease with the increasing complexity of the task (more clusters, higher overlap, and higher dimensionality) as it is more difficult to find reliable clustering of the data. This hypothesis is supported by our preliminary results for clustering body parts from simultaneous tactile and linguistic input [59] and by the results presented in this paper in Table II. The quality of one-step mapping correlates with the quality of the visual data clustering. For Blender data, the worst performance was achieved for mapping words to the visual feature shape ($52 \pm 5\%$) and for physical iCub for the feature color ($62 \pm 3\%$). The feature shape has the highest number of clusters (10), and the feature color has the second highest number of clusters (9) and the highest overlap of the clusters for the physical iCub. Real-world tasks are much more complex, and we can expect tens of different object shapes. In that case, the clustering performance is crucial, and one-step mapping would not be able to provide reliable mapping. It can be seen from the results that the proposed mapping which enables gradual re-estimation of the model parameters and works in a dynamic fashion achieves much higher accuracy even for the cases where one-step mapping fails. We suppose that the mapping accuracy of the proposed method decreases more slowly with the decreasing accuracy of clustering of individual modalities. Unfortunately, this factor was not studied in our restricted scenario, but the preliminary results on mapping tactile and linguistic input [59] support this hypothesis. We plan to investigate this phenomenon more deeply in future research.

The language dataset differs considerably from natural language. However, the dataset reflects some characteristics from the findings of Werker and McLeod [53] as infant-directed words are usually kept short with large pauses between words. Moreover, Brent and Siskind [3] showed that frequency of exposure to a word in isolation predicts better whether that word will be learned than the total frequency of exposure to that word. In addition, Snow [42] found that mother's speech to two-year-old is much simpler and less redundant than their speech to ten-year-old. This indicates that young children have available a sample of speech which is simpler, more redundant and less confusing than normal adult speech.

The proposed algorithm was tested on language to vision mapping and on language to tactile mapping [59], and it can be easily extended to language to any other modalities mapping. The mapping between multiple modalities and words was researched in [47]. Fazly *et al.* [14] proposed a probabilistic model of cross-situational learning where they considered sentences that contain objects and their motion. Monaghan *et al.* [33] studied differences between cross-situational learning of nouns and verbs on human participants as an extension of Tomasello and Akhtar's [50] work and Schwartz and Terrell's [39] work. Monaghan *et al.* [33] considered learning of verbs as difficult as learning of nouns when presented in a syntactic context. The authors noticed that nouns are learned quicker, but verbs and nouns can be acquired simultaneously.

An important direction for extensions of the proposed model deals with the grounding of language related not just to static sequences but also to events which require a temporal axis. For example, Crick and Scassellati [7] studied the interconnection between verbal narratives and episodes of intentional relative motion with the goal that the robot can learn from observing a game the rules of the game, relationships between players and their goals and intentions and then participate in the play.

The main goal of this paper is to analyze mapping between modalities. Thus, we kept processing of the individual modalities deliberatively simple for better understanding of cross-situational learning. We performed further experiments with an artificially generated dataset where the visual features, as well as noise in the linguistic and the visual domain, were under full control. These experiments are described in detail in [58].

APPENDIX

Here we describe an extension of the proposed algorithm presented in Section II-B for a case where we have sentences with variable structure. This means that we cannot

Algorithm 2 Sequential Mapping—Variable Sentence

Inputs:
language clusters L_i ($i \in 1:I$), visual clusters $K_j^k \sim N(\mathbf{m}^k, \mathbf{S}^k)$, $j \in 1:J^k$ for each feature k, visual input $\mathbf{x_t} = \{\mathbf{x}_t^1, \ldots, \mathbf{x}_t^K\}$ and corresponding language data $W_t = \{\mathbf{w}_t^1, \ldots, \mathbf{w}_t^{N_t}\}$ for each trial t and number of clusters J^k
Output:
mapping between all visual classes K_j^k and language classes L_i

while $\sum J^k > 0$ and $\mathbf{X}$ is not empty **do**
  $l_t^n \leftarrow$ assign each word $\mathbf{w}_t^n$ from each sentence t to a language cluster (Winner-takes-all, see Eq. (15), $l_t^n = \text{argmax}_i(P(w_t^n|L_i))$
  for $k \in$ {size, color, orientation, texture, shape} **do**
    $\mathbf{v}_t^k \leftarrow$ assign each data point $\mathbf{x}_t^k$ to a visual cluster (Winner-takes-all, see Eq. (15), $v_t^k = \text{argmax}_j(P(\mathbf{x}_t^k|K_j^k))$
    for $j = 1 : J^k$ **do**
      for $i = 1 : I$ **do**
        $T_{ij}^k \leftarrow$ how many times did visual class i co-occur with language class j ($T_{ij}^k = \sum_{t;\mathbf{v}_t^k==j} \sum_n \delta(l_t^n, i)$), where δ is Kronecker delta
      end for
    end for
  end for
  $[km, im, jm] \leftarrow \text{argmax}_k \text{argmax}_i \text{argmax}_j T_{i,j}^k$ (the visual cluster K_{jm}^{km} is mapped to the language cluster L_{im})
  $\mathbf{X}_{del}^{km} \leftarrow$ data points assigned to K_{jm}^{km} and L_{im}
  $\theta_{new,J^k}^k \leftarrow N(\mathbf{x}_{del}^k)$ learn Gaussian on the to be deleted data
  $\mathbf{X}^{km} \leftarrow \mathbf{X}^{km} \setminus \mathbf{X}_{del}^{km}$ delete data assigned to K_{jm}^{km} and L_{im}
  $J^{km} \leftarrow J^{km} - 1$
  relearn $K_j^{km} \sim N(\mathbf{m}^k, \mathbf{S}^k)$ on new data $\mathbf{x}^{km}$ with J^{km} clusters
end while
cluster visual data using new θ_{new}^k parameters (cluster centers $\mathbf{m}^k$, covariance matrices $\mathbf{S}^k$) and perform one-step mapping (Model 1)

directly associate words from the sentence with individual visual features, and therefore, associations to all visual features must be taken into account. The whole algorithm is described in Algorithm 2 in the form of a pseudocode.

Algorithm 2 can be further extended when we incorporate the language model and corresponding probabilities of sequences of individual visual features (see [58]).

REFERENCES

[1] L. W. Barsalou, "Grounded cognition," *Annu. Rev. Psychol.*, vol. 59, pp. 617–645, Jan. 2008.

[2] F. B. Klein. (2016). *GWR and GNG Classifier—File Exchange—MATLAB Central*. [Online]. Available: http://uk.mathworks.com/matlabcentral/fileexchange/57798-gwr-and-gng-classifier

[3] M. R. Brent and J. M. Siskind, "The role of exposure to isolated words in early vocabulary development," *Cognition*, vol. 81, no. 2, pp. B33–B44, 2001.

[4] C. Büchel, C. Price, and K. Friston, "A multimodal language region in the ventral visual pathway," *Nature*, vol. 394, no. 6690, pp. 274–277, 1998.

[5] A. Cangelosi, A. Greco, and S. Harnad, "From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories," *Connection Sci.*, vol. 12, no. 2, pp. 143–162, 2000.

[6] S. Coradeschi, A. Loutfi, and B. Wrede, "A short review of symbol grounding in robotic and intelligent systems," *Künstliche Intelligenz*, vol. 27, no. 2, pp. 129–136, 2013, doi: 10.1007/s13218-013-0247-2.

[7] C. Crick and B. Scassellati, "Controlling a robot with intention derived from motion," *Topics Cogn. Sci.*, vol. 2, no. 1, pp. 114–126, 2010.

[8] J. C. Culham and K. F. Valyear, "Human parietal cortex in action," *Current Opin. Neurobiol.*, vol. 16, no. 2, pp. 205–212, 2006.

[9] M. Daoutis and N. Mavridis, "Towards a model for grounding semantic composition," in *Proc. Artif. Intell. Simulat. Behav. (AISB)*, 2014.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.

[11] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 2625–2634.

[12] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 1873–1879.

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[14] A. Fazly, A. Alishahi, and S. Stevenson, "A probabilistic computational model of cross-situational word learning," *Cogn. Sci.*, vol. 34, no. 6, pp. 1017–1063, 2010.

[15] M. Fleischman and D. Roy, "Grounded language modeling for automatic speech recognition of sports video," in *Proc. ACL*, Columbus, OH, USA, 2008, pp. 121–129.

[16] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using speakers' referential intentions to model early cross-situational word learning," *Psychol. Sci.*, vol. 20, no. 5, pp. 578–585, 2009.

[17] V. Gliozzi, J. Mayor, J. F. Hu, and K. Plunkett, "Labels as features (not names) for infant categorization: A neurocomputational approach," *Cogn. Sci.*, vol. 33, no. 4, pp. 709–738, 2009.

[18] E. Goldberg, N. Driedger, and R. I. Kittredge, "Using natural-language processing to produce weather forecasts," *IEEE Expert*, vol. 9, no. 2, pp. 45–53, Apr. 1994.

[19] P. Gorniak and D. Roy, "Speaking with your sidekick: Understanding situated speech in computer role playing games," in *Proc. AIIDE*, Marina Del Rey, CA, USA, 2005, pp. 57–62.

[20] S. Harnad, "The symbol grounding problem," *Phys. D Nonlin. Phenomena*, vol. 42, nos. 1–3, pp. 335–346, 1990.

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3128–3137.

[22] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing robot embodiments in a guided discovery learning interaction with children," *Int. J. Soc. Robot.*, vol. 7, no. 2, pp. 293–308, 2015.

[23] P. Lamere *et al.*, "The CMU SPHINX-4 speech recognition system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1. Hong Kong, 2003, pp. 2–5.

[24] P. Li, I. Farkas, and B. MacWhinney, "Early lexical development in a self-organizing neural network," *Neural Netw.*, vol. 17, nos. 8–9, pp. 1345–1362, 2004.

[25] E. M. Markman, "Constraints children place on word meanings," *Cogn. Sci.*, vol. 14, no. 1, pp. 57–77, 1990.

[26] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Netw.*, vol. 15, nos. 8–9, pp. 1041–1058, 2002.

[27] N. Mavridis, "Grounded situation models for situated conversational assistants," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.

[28] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *J. Robot. Auton. Syst.*, vol. 63, pp. 22–35, Jan. 2015.

[29] B. McMurray, J. S. Horst, and L. K. Samuelson, "Word learning emerges from the interaction of online referent selection and slow associative learning," *Psychol. Rev.*, vol. 119, no. 4, pp. 831–877, 2012.

[30] G. Metta *et al.*, "The iCub humanoid robot: An open platform for research in embodied cognition," in *Proc. ACM 8th Workshop Perform. Metrics Intell. Syst.*, Gaithersburg, MD, USA, 2008, pp. 50–56.

[31] G. Metta, P. Fitzpatrick, and L. Natale, "YARP: Yet another robot platform," *Int. J. Adv. Robot. Syst.*, vol. 3, no. 1, pp. 43–48, 2006.

[32] M. Mishkin, L. G. Ungerleider, and K. A. Macko, "Object vision and spatial vision: Two cortical pathways," *Trends Neurosci.*, vol. 6, no. 10, pp. 414–417, 1983.

[33] P. Monaghan, K. Mattock, R. A. I. Davies, and A. C. Smith, "Gavagai is as Gavagai does: Learning nouns and verbs from cross-situational statistics," *Cogn. Sci.*, vol. 39, no. 5, pp. 1099–1112, 2015, doi: 10.1111/cogs.12186.

[34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[35] W. V. Quine, "On the reasons for indeterminacy of translation," *J. Philos.*, vol. 67, no. 6, pp. 178–183, 1970.

[36] T. Regier, "The emergence of words: Attentional learning in form and meaning," *Cogn. Sci.*, vol. 29, no. 6, pp. 819–865, 2005.

[37] D. Roy and N. Mukherjee, "Towards situated speech understanding: Visual context priming of language models," *Comput. Speech Lang.*, vol. 19, no. 2, pp. 227–248, 2005.

[38] D. K. Roy, "Learning visually grounded words and syntax for a scene description task," *Comput. Speech Lang.*, vol. 16, no. 3, pp. 353–385, 2002.

[39] R. G. Schwartz and B. Y. Terrell, "The role of input frequency in lexical acquisition," *J. Child Lang.*, vol. 10, no. 1, pp. 57–64, 1983.

[40] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, nos. 1–2, pp. 39–91, 1996.

[41] K. Smith, A. D. M. Smith, R. A. Blythe, and P. Vogt, *Cross-Situational Learning: A Mathematical Approach* (LCNS 4211). Heidelberg, Germany: Springer, 2006, pp. 31–44.

[42] C. E. Snow, "Mothers' speech to children learning language," *Child Dev.*, vol. 43, no. 2, pp. 549–565, 1972.

[43] G. Spitsyna, J. E. Warren, S. K. Scott, F. E. Turkheimer, and R. J. Wise, "Converging language streams in the human temporal lobe," *J. Neurosci.*, vol. 26, no. 28, pp. 7328–7336, 2006.

[44] F. Stramandinoli, D. Marocco, and A. Cangelosi, "The grounding of higher order concepts in action and language: A cognitive robotics model," *Neural Netw.*, vol. 32, pp. 165–173, Aug. 2012.

[45] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adapt. Behav.*, vol. 13, no. 1, pp. 33–52, 2005.

[46] M. Taddeo and L. Floridi, "Solving the symbol grounding problem: A critical review of fifteen years of research," *J. Exp. Theor. Artif. Intell.*, vol. 17, no. 4, pp. 419–445, 2005.

[47] A. Taniguchi, T. Taniguchi, and A. Cangelosi, "Multiple categorization by iCub: Learning relationships between multiple modalities and words," in *Proc. IROS Workshop Mach. Learn. Methods High Level Cogn. Capabilities Robot.*, 2016.

[48] V. Tikhanoff *et al.*, "An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator," in *Proc. ACM 8th Workshop Perform. Metrics Intell. Syst.*, Gaithersburg, MD, USA, 2008, pp. 57–61.

[49] V. Tikhanoff, A. Cangelosi, and G. Metta, "Integration of speech and action in humanoid robots: iCub simulation experiments," *IEEE Trans. Auton. Mental Develop.*, vol. 3, no. 1, pp. 17–29, Mar. 2011.

[50] M. Tomasello and N. Akhtar, "Two-year-olds use pragmatic cues to differentiate reference to objects and actions," *Cogn. Dev.*, vol. 10, no. 2, pp. 201–224, 1995.

[51] M. Vavrečka and I. Farkaš, "A multimodal connectionist architecture for unsupervised grounding of spatial language," *Cogn. Comput.*, vol. 6, no. 1, pp. 101–112, 2014.

[52] P. Vogt, "Language evolution and robotics: Issues on symbol grounding," in *Artificial Cognition Systems*. Hershey, PA, USA: IGI Glob., 2006, p. 176.

[53] J. F. Werker and P. J. McLeod, "Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness," *Can. J. Psychol. Revue Can. De Psychol.*, vol. 43, no. 2, pp. 230–246, 1989.

[54] F. Xu and J. B. Tenenbaum, "Word learning as Bayesian inference," *Psychol. Rev.*, vol. 114, no. 2, pp. 245–272, 2007.

[55] C. Yu and L. B. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychol. Sci.*, vol. 18, no. 5, pp. 414–420, 2007.

[56] C. Yu and L. B. Smith, "Modeling cross-situational word-referent learning: Prior questions," *Psychol. Rev.*, vol. 119, no. 1, pp. 21–39, 2012.

[57] D. Yurovsky, C. Yu, and L. B. Smith, "Competitive processes in cross-situational word learning," *Cogn. Sci.*, vol. 37, no. 5, pp. 891–921, 2013.

[58] K. Štěpánová, "Hierarchical probabilistic model of language acquisition," Ph.D. dissertation, Dept. Cybern., Czech Tech. Univ. Prague, Prague, Czech Republic, 2016.

[59] K. Štěpánová *et al.*, "Where is my forearm? Clustering body parts from simultaneous tactile and linguistic input," in *Proc. Cogn. Artif. Life (KUZ XVII)*, 2017.

Karla Štepánová received the master's degree in condensed matter physics from Charles University, Prague, Czech Republic, and the Ph.D. degree in artificial intelligence and biocybernetics from Czech Technical University in Prague, Prague, in 2017.

She is a Researcher with the Czech Institute of Informatics, Robotics, and Cybernetics, Prague. She was a Visiting Researcher with Plymouth University, Plymouth, U.K., in 2016. Her current research interests include probabilistic models of cognition, unsupervised learning, language acquisition, and multimodal integration.

Frederico Belmonte Klein was born in Porto Alegre, Brazil, in 1982. He received the Electrical Engineering degree and the medical degree from the Federal University of Rio Grande do Sul, Porto Alegre, in 2006 and 2014, respectively. He is currently pursuing the Ph.D. degree in cognitive robotics with the University of Plymouth, Plymouth, U.K.

Angelo Cangelosi received the degree in psychology and cognitive science from the University of Rome La Sapienza, Rome, Italy, and the University of Genoa, Genoa, Italy.

He is a Professor of artificial intelligence and cognition and the Director of the Centre for Robotics and Neural Systems with Plymouth University, Plymouth, U.K. and was a Visiting Scholar with the University of California at San Diego, San Diego, CA, USA, and the University of Southampton, Southampton, U.K. His current research interests include language grounding and embodiment in humanoid robots, developmental robotics, human–robot interaction, and on the application of neuromorphic systems for robot learning.

Michal Vavrečka received the Ph.D. degree in general psychology from the Faculty of Social Studies, in 2008.

He is with the Czech Institution for Informatics Robotics and Cybernetics, Czech Technical University in Prague, Prague, Czech Republic. He develops cognitive architecture for symbol grounding, language acquisition, and knowledge representation.

AUTHOR QUERIES
AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ1: Please confirm/give details of funding source.
AQ2: Please provide the postal code for "Plymouth University Plymouth, U.K."
AQ3: Please provide the page range for References [9], [47], and [59].
AQ4: Please provide the department name for Reference [27].
AQ5: Please confirm if the location and publisher information for Reference [41] is correct as set.
AQ6: Please specify the type of the degree and year attained by the author "A. Cangelosi."
AQ7: Please provide the organization name and location for the Ph.D. degree obtained by "M. Vavrečka" biography.

Mapping Language to Vision in a Real-World Robotic Scenario

Karla Štepánová, Frederico Belmonte Klein, Angelo Cangelosi, and Michal Vavrečka

***Abstract*—Language has evolved over centuries and was gradually enriched and improved. The question, how people find assignment between meanings and referents, remains unanswered. There are many of computational models based on the statistical co-occurrence of meaning-reference pairs. Unfortunately, these mapping strategies show poor performance in an environment with a higher number of objects or noise. Therefore, we propose a more robust noise-resistant algorithm. We tested the performance of this novel algorithm with simulated and physical iCub robots. We developed a testing scenario consisting of objects with varying visual properties presented to the robot accompanied by utterances describing the given object. The results suggest that the proposed mapping procedure is robust, resistant against noise and shows better performance than one-step mapping for all levels of noise in the linguistic input, as well as slower performance degradation with increasing noise. Furthermore, the proposed procedure increases the clustering accuracy of both modalities.**

***Index Terms*—Cognitive modeling, cross-situational learning, iCub robot, language acquisition, symbol grounding.**

I. Introduction

THE ESSENTIAL (and still not fully answered) question in language acquisition is how percepts are anchored in some arbitrary symbols, in other words, how words (symbols) get their meanings. This is the so-called symbol grounding problem [20]. For many years, cognitive modeling, neuroscience, psychology, and machine learning have jointly attempted to understand how humans can solve this "problem" [6]. The ability to learn language through perception and especially through visual grounding is not only important for understanding human cognition but is also applicable in many areas, such as verbal control of interactive robots [28], automatic sports commentators [15], car navigation systems, for the visually impaired, situated speech understanding in computer games [19], automated generation of weather forecasts [18], tutoring children in foreign languages [22], etc.

Despite the extensive research in the area of language acquisition, the question how the word-to-meaning mapping is learned remains unanswered. There are basically two approaches for mapping language to other modalities. A classic approach for separate modality-dependent representations is advocated by Barsalou [1]; while an example of a system where language is mapped to an intermediate modal representation that can be derived by multiple modalities is [27]. Using the unsupervised approach, Li *et al.* [24] designed the DevLex model, consisting of two self-organizing networks that are bidirectionally connected. Gliozzi *et al.* [17] proposed an alternative with a multimodal representation layer: their unsupervised feature-based model was used to account for early category formation in young infants. This approach postulates the unsupervised role of linguistic labels that can affect categorization during the acquisition process, which has also been supported by Taniguchi *et al.* [47]. Vavrečka and Farkaš [51] recently introduced a multimodal architecture for the grounding of spatial words using a biologically inspired approach (separate "what" and "where" visual subsystems) in which the visual scenes (two objects in 2-D space in a spatial relation) are associated with their linguistic descriptions, thus leading to the integration of modalities.

However, a fully unsupervised architecture, which would be able to deal with language grounding [46], particularly language grounding in a case where sentences have variable structure and when there is more than one object in a scene, is not available. The current state-of-the-art on variable-length sentences is very restricted and deals only with static scenes [29]. Most of the recent models based on deep networks are oriented toward application in image-to-text [21] or video-to-text [11] mapping and do not take into account the psychological aspects of language acquisition (e.g., mutual exclusivity). Moreover, these systems are trained in a supervised manner without the advantage of transfer learning.

The difficulty of the task was described in a well-known experiment performed by Quine [35] who imagined the anthropologist meeting a native who pointed at the scene and said "gavagai." When the anthropologist is stimulated in a situation by seeing a rabbit, he will suppose that the word represents the running rabbit in front of him, even though it could also mean "ground," "sun," "hello," or whatever else. This problem is related to language relativity, as there are several

Manuscript received June 27, 2017; revised December 6, 2017; accepted February 24, 2018. This work was supported in part by the European EU FP7 Research Project TRADR under Grant 609763, in part by TACR CAK under Grant TE01020197, in part by the CAPES Foundation, Ministry of Education of Brazil under Grant BEX 1084/13-5, in part by the CNPq Brazil under Grant 232590/2014-1, and in part by the U.K. EPSRC Project BABEL under Grant EP/J004561/1 and Grant EP/J00457X/1. *(Corresponding author: Karla Štepánová.)*

K. Štepánová and M. Vavrečka are with the Czech Institute of Informatics, Robotics, and Cybernetics, Czech Technical University in Prague, 16000 Prague, Czech Republic (e-mail: karla.stepanova@ciirc.cvut.cz).

F. B. Klein and A. Cangelosi are with the School of Computing, Electronics and Mathematics, Plymouth University, Plymouth PL4 8AA, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2819359

objects and their features that are described by words [33]. A simplified version of this problem consists of a simple visual scene and separate words that are grounded based on statistical co-occurrence (cross-situational learning).

From a neuroscientific point of view, symbol grounding can be viewed as a process of finding mappings between primary unimodal visual and language brain areas. Where exactly the integration is performed is still the subject of research, and existing literature provides only incomplete accounts of the cortical location of this convergence. For example, the study by Büchel *et al.* [4] provides evidence for the involvement of the left basal posterior temporal lobe (BA37) in the integration of language and visual information. Other studies (e.g., [43]) propose that access to verbal meaning depends on anterior and posterior heteromodal cortical systems within the temporal lobe. The grounding of actions and motoric primitives is associated with the activity in the dorsal stream and the premotor cortex [8].

How language can be developed in an unsupervised manner is also an important task in developmental robotics as most language acquisition in humans is fully unsupervised. One of the main long-term objectives of many teams worldwide is building conversational robots, which will be able to participate in cooperative tasks mediated by a natural language. It has been shown how robots can learn new symbols using already grounded ones and their combination [5] and how to transfer knowledge between agents [52]. Cangelosi *et al.* [5] presented their research on language emergence and grounding in sensorimotor agents and robots. This model was further extended by Tikhanoff *et al.* [49], who performed iCub simulation experiments and focused on the integration of speech and action. The grounding of higher-order concepts in action was also explored by Stramandinoli *et al.* [44], who made use of recurrent neural networks. Sugita and Tani [45] described an experiment dealing with semantic compositionality: the capability of a robot to use the compositional structure to generalize novel word combinations. Daoutis and Mavridis [9] summarized desiderata for grounded semantic compositionality. Despite all the progress in language grounding, however, the current state-of-the-art on grounding variable-length sentences is very restricted and deals only with static scenes [29], [38].

In this paper, we present this paper in the area of language acquisition using a real-world robotic scenario. We implemented a hierarchical cognitive architecture for language acquisition that includes visual and language processing. In particular, we chose to extend current models of cross-situational learning by allowing vision-to-language mapping in the case of a nonequal number of classes and by taking into account situation-time dynamics.

We show that this can be accomplished more efficiently by replacing one-shot mapping with sequential mapping and adding inhibitory mechanisms to the connections. The best mapped classes are gradually eliminated, and the clusterization is adaptively changed. We see this paper as an extension of the McMurray *et al.* [29] model, and we compare it with other single-step mapping models. The mapping strategy presented in this paper was shown to be very robust as it not only can find the mapping in circumstances of very noisy real-world input but also increases the clustering accuracy of both modalities. Recently, we tested the proposed algorithm also on the task of clustering body parts from simultaneous tactile and linguistic input [59]. In that case, sequential mapping showed slower degradation with increasing noise level in the linguistic input and outperformed one-step mapping for all dataset sizes and all levels of noise.

The rest of this paper is structured as follows. In Section II, we compare different mapping algorithms used in cross-situational learning. In particular, in Section II-B we provide a mathematical formulation of the newly proposed sequential mapping algorithm, and in Section II-C we describe the whole cognitive architecture which incorporates unimodal processing of vision and language and finding their association through the mapping algorithm. The performance of the proposed method on data from an iCub humanoid robot and from an iCub simulator is evaluated in Section III. Finally, results are discussed in Section IV with suggestions for future work.

II. Materials and Methods

In this section, we first present one-step and newly proposed sequential mapping algorithms (Sections II-A and II-B). Then, we describe in detail the whole cognitive architecture used to process data from individual modalities (vision and language). Finally, we describe the iCub robotic platform and the iCub simulator in Section II-D and provide a description of the evaluation in Section II-F.

A. One-Step Mapping in Cross-Situational Learning

One possible way how to establish mapping between visual concepts and linguistic elements is to use frequencies of referent and meaning co-occurrences; that is, the ones with the highest co-occurrence are mapped together [40], [41], [54]. This method is usually called cross-situational learning and supposes the availability of the ideal associative learner who can keep track and store all co-occurrences in all trials, internally memorizing and representing the word–object co-occurrence matrix of the input. This allows the learner to subsequently choose the most strongly associated referent [55], [56]. These models do not see the mapping as dynamic competition but operate only with the static state. Although some use likelihoods of different words and referents to perform Bayesian inference [16], [54], they do not take into account how the similarity of two-word forms can affect learning although it has been shown that the similarity affects learning in children [36]. Another shortcoming of these strategies is that they do not address how these similarities affect learning in a dynamic competition.

The simplest one-step word-to-referent learning algorithm simply accumulates word-referent pairs. This can be viewed as Hebbian learning: the connection between a word and an object is strengthened if the pair co-occurs in a trial. To extend this basic idea, we can enable also forgetting by introducing the parameter η, which can capture the memory decay. This so-called dumb associative-learning model was implemented by Yu and Smith [56]. Supposing that at each trial t we observe an object $\mathbf{o}_t^n$ and hear a corresponding word $\mathbf{w}_t^n$ (N_t possible associations), we can describe the update of the strength of the association between word model $L(i)$ and object model $K(j)$

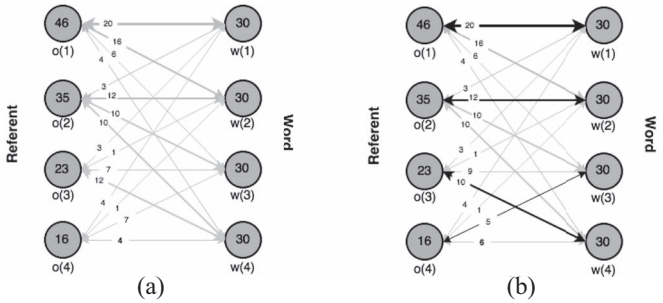


Fig. 1. One-step mapping. (a) In the first stage, weights between objects (referents) and words are changed using Hebbian learning. The connection between a word and an object is increased if the pair co-occurs in a trial [56]. (b) Then, word-to-referent mapping is found in one step: objects and words with the highest co-occurrence are mapped together. The number on each connection between word $w(i)$ and object $o(j)$ refers to the number of co-occurrences of $o(j)$ and $w(i)$. In this example, we suppose that there are 30 occurrences of each word.

as follows:

$$A(i,j) = \sum_{t=1}^{R} \eta(t) \sum_{n=1}^{N_t} \delta\left(\mathbf{w}_t^n, i\right) \delta\left(\mathbf{o}_t^n, j\right) \tag{1}$$

where R is the number of trials, δ is the Kronecker delta function (equal to 1 when both arguments are identical and 0 otherwise), $\mathbf{w}_t^n$ and $\mathbf{o}_t^n$ indicate the nth word–object association that the model attends to and attempts to learn in the trial t, and $\eta(t)$ is the parameter controlling the gain of the strength of the association.

Now let us assume that word $\mathbf{w}(i)$ is modeled by model L_i in the language domain and object (referent) $\mathbf{o}(j)$ is modeled by the model $K_{m(i)}$ in the visual domain. Our goal is to find the corresponding model $K_{m(i)}$ from the visual subdomain for each model L_i from the language domain to assign them together. Indices $m(i)$ are found as follows:

$$\forall i : m(i) = \underset{i}{\operatorname{argmax}} A(i,j) \tag{2}$$

where $\mathbf{A}$ is the co-occurrence matrix computed in (1) [element $A(i,j)$ captures co-occurrence between word $\mathbf{w}(i)$ and object $\mathbf{o}(j)$].

In Fig. 1, one-step mapping is visualized as it was implemented in this paper.

Modifications of the basic model include Regier's [36] work. He proposed a mapping model, which stems from competition models and incorporates two-way associations between words and referents. This enables the model to capture selective attention to individual words and referents, as well as to provide a probability distribution over associated referents/words. Regier [36] also showed that for his model, learning of a novel word is most effective when memory interference is minimized.

B. Proposed Sequential Mapping

Because we know that learning is not static but is a dynamic process, it seems reasonable to extend the basic idea of one-step cross-situational learning by incorporating dynamic competition mechanisms between words and referents in the model. To capture the dynamic competition among models, we extend the basic one-step mapping algorithm with the sequential addition of inhibitory connections.

In this case, the process of finding word-referent associations resembles Hebbian learning with inhibitory connections. Once the word is associated with a corresponding object (referent), links from this referent to other words are inhibited. This idea also corresponds to the fact that children prefer mapping where an object has only one label to multiple labels, the so-called mutual exclusivity bias [25]. The inhibitory mechanisms and situation-time dynamics were already partially included in the model of cross-situational learning proposed by McMurray *et al.* [29].

Even though our model shares some similarities with the model proposed by McMurray *et al.* [29], our model stems from different computational mechanisms. The proposed sequential mapping is able to capture nondiscrete assignment to individual clusters, as well as dynamic competition mechanisms. The first mechanism is incorporated into the model by considering likelihoods that the observed data were generated by a given model instead of 1/0 assignment to models. In this way, similarities of individual meanings and referents, as well as the likelihood of their recognition in each trial, is taken into account. The second mechanism (dynamic competition) facilitates the sequential mapping as the best-mapped classes are gradually justified with inhibitory connections to other classes (i.e., after a reliable assignment between a language and a tactile model is found, inhibitory connections among this tactile model and all other language models are added). Thanks to this mechanism, the mutual exclusivity principle (the fact that children prefer mapping where an object has only one label to multiple labels [25]) is guaranteed. The assignment between visual models K_j and language models L_i is found using the following iterative procedure.

1) Visual and language data are clustered separately and the corresponding posterior probabilities are found

$$p\left(L_i|\,\mathbf{w}_t^n\right) = \frac{p\left(\mathbf{w}_t^n|L_i\right)p(L_i)}{\sum_{i'} p\left(\mathbf{w}_t^n|\,L_{i'}\right)p(L_{i'})}, \quad \forall i \in \{1,\ldots,I\}$$
$$\forall t \in \{1,\ldots,T\}, \quad \forall n \in \{1,\ldots,N_t\} \tag{3}$$

$$p\left(K_j|\,\mathbf{o}_t^n\right) = \frac{p\left(\mathbf{o}_t^n|K_j\right)p\left(K_j\right)\mathbf{k}(j)}{\sum_{j'} p\left(\mathbf{w}_t^n|K_{j'}\right)p\left(K_{j'}\right)}, \quad \forall j \in \{1,\ldots,J\}$$
$$\forall t \in \{1,\ldots,T\}, \quad \forall n \in \{1,\ldots,N_t\} \tag{4}$$

where I is the number of language models, J is the number of visual models, T is the number of trials, and N_t is the number of possible object-word associations in trial t.

2) The most probable cluster to which the data point is assigned is selected (for each data point)

$$a_t^n = \underset{i}{\operatorname{argmax}}\, p\left(L_i|\,\mathbf{w}_t^n\right) \tag{5}$$

$$b_t^n = \underset{j}{\operatorname{argmax}}\, p\left(K_j|\,\mathbf{o}_t^n\right)$$
$$\forall t \in \{1,\ldots,T\}, \forall n \in \{1,\ldots,N_t\}. \tag{6}$$

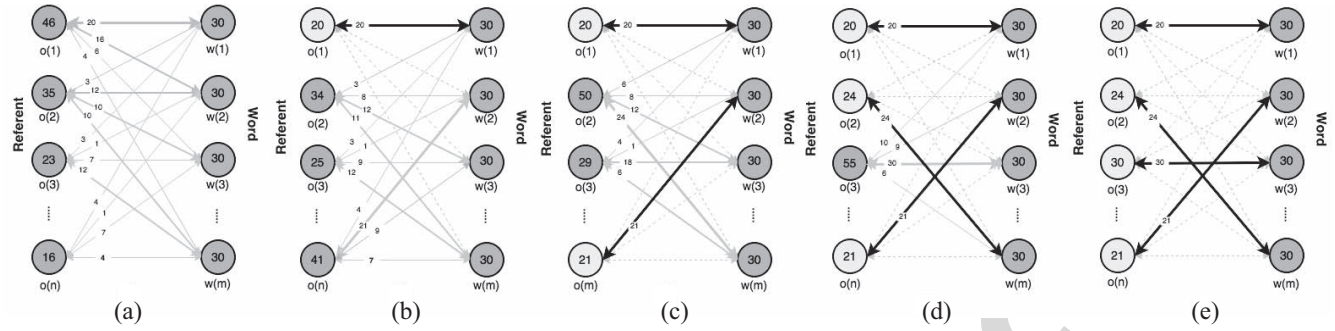


Fig. 2. Sequential mapping. The toy example of sequential mapping is shown to clarify the mechanism of finding the object-word assignment. In this example, we suppose that there are 30 occurrences of each word. The dotted line marks the inhibitory connection between object $o(j)$ and word $w(i)$ and the black line corresponds to the already found mapping. The number on each connection between word $w(j)$ and object $o(i)$ refers to the number of co-occurrences of $o(j)$ and $w(i)$. Objects $o(j)$ and words $w(i)$ are assigned to corresponding models based on the given clustering mechanism.

3) Co-occurrence matrix $A(i,j)$ is computed

$$A(i,j) = \zeta(i,j) \sum_{t=1}^{R} \eta(t) \sum_{n=1}^{N_t} \delta\left(a_t^n, i\right) \delta\left(b_t^n, j\right) \tag{7}$$

where R is the number of trials, $\zeta(i,j)$ is the matrix storing the strength of the connections between visual model K_j and language model L_i, and $\eta(t)$ is the parameter controlling the gain of the strength of association.

4) The best assignment is selected

$$[im, m(im)] = \underset{i}{\operatorname{argmax}} \underset{j}{\operatorname{argmax}} A(i,j). \tag{8}$$

In this step, the visual model $K_{m(im)}$ is assigned to the language model L_{im}.

5) Inhibition connections are added between the assigned visual model $K_{m(im)}$ and all language models other than L_{im} (mutual exclusivity)

$$\zeta(i, m(im)) = \zeta(i, m(im))(1 - z_1), \quad \forall i \neq im \tag{9}$$

where z_i is the parameter capturing the strength of the inhibition (in our experiment, this is set to 1, which corresponds to the total inhibition of the given connection).

6) Inhibition is added to the assigned visual model $K_{m(im)}$ (a prior probability of the model is changed)

$$k(m(im)) = k(m(im))(1 - z_2) \tag{10}$$

where z_2 is the parameter capturing the inhibition of the assigned visual model (in our experiment, this parameter is set to 1, which corresponds to total inhibition of the given model).

7) The assigned points are deleted from the dataset (data points which belong to model K_{jm} and L_{im})

$$X = X \backslash \left\{ \left(\mathbf{o}_t^n, \mathbf{w}_t^n\right) \mid \underset{j}{\operatorname{argmax}}\, p\left(K_j | \mathbf{o}_t^n\right) == m(im) \right. \tag{11}$$

$$\left. \wedge \underset{i}{\operatorname{argmax}}\, p\left(L_i | \mathbf{w}_j^n\right) == im \right\}. \tag{12}$$

8) If the dataset is not empty ($X \notin \emptyset$) or $\|\mathbf{k}\| > 0$ (some of the visual models are not totally inhibited), go to step 1, else stop.

The proposed algorithm where words are assigned to corresponding referents in a sequential manner is visualized in Fig. 2.

In the ideal case, unambiguous mapping between the two clusterizations will be found. In the real case (where the clusterization in visual and language layers is not optimal), none or more than one model from the visual layer will be assigned to one cluster L_i in the language layer or vice versa.

C. Specific Architecture

Our multimodal hierarchical architecture consists of multimodal and unimodal parts. The unimodal part has two layers performing separate processing of localist inputs: 1) visual objects and 2) auditory word-forms. Both unimodal layers are subsequently mapped one to each other in the upper multimodal layer (see Fig. 3).

1) Visual Layer: Each data point (object $\mathbf{o}_t^n$) can be considered as a triplet of continuous-valued vectors for each visual feature: $\mathbf{o}_t^n = (\mathbf{x}_{t,n}^{\text{size}}, \mathbf{x}_{t,n}^{\text{color}}, \mathbf{x}_{t,n}^{\text{shape}})$. This enables us to write the visual dataset as $X^{\text{vis}} = [X^{\text{size}} X^{\text{color}} X^{\text{shape}}]$ and process data for each visual feature separately. For processing visual data, the Gaussian mixture model (GMM) was used, which is a convex mixture of d-dimensional Gaussian densities $l(\mathbf{x}^k | \theta_j^k)$, where $k \in \{\text{size, color, shape}\}$. In this case, each visual model K_j^k is described by a set of parameters θ_j^k. The posterior probabilities $p(\theta_j^k | \mathbf{x}^k)$ are computed as follows:

$$p\left(\theta_j^k | \mathbf{x}^k\right) = \sum_{j=1}^{J_k} r_j^k l\left(\mathbf{x}^k | \theta_j^k\right) \tag{13}$$

$$l\left(\mathbf{x}^k | \theta_j^k\right) = \frac{1}{\sqrt{(2\pi)^d} \sqrt{|\mathbf{S}_j^k|}} \exp\left[-\frac{1}{2} \left(\mathbf{x}^k - \mathbf{m}_j^k\right)^T \left(\mathbf{S}_j^k\right)^{-1} \times \left(\mathbf{x}^k - \mathbf{m}_j^k\right) \right] \tag{14}$$

where $k \in \{\text{size, color, shape}\}$, $\mathbf{x}^k$ is a set of d-dimensional continuous-valued data vectors, r_j^k are the mixture weights, J_k is the number of visual models for each visual feature k, and parameters θ_j^k are cluster centers $\mathbf{m}_j^k$ and covariance matrices $\mathbf{S}_j^k$.

The Gaussian mixture is trained by the expectation-maximization algorithm [10]. An output of this layer for

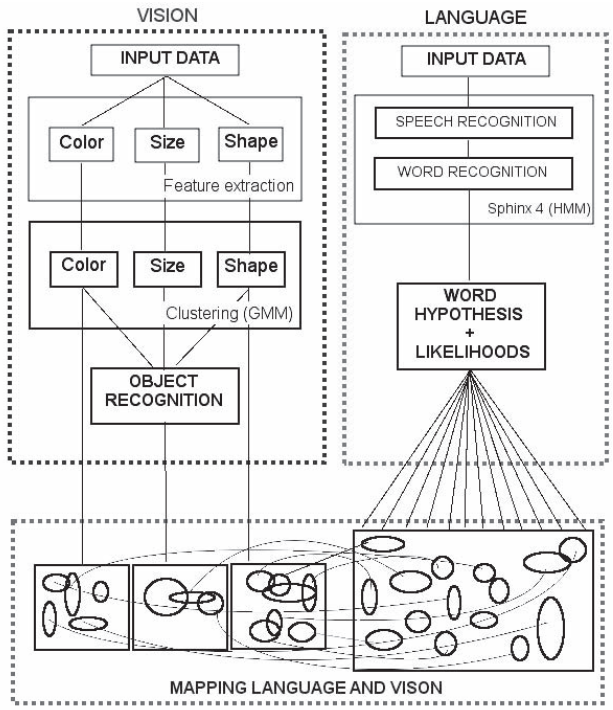


Fig. 3. Proposed multimodal architecture.

each data point $\mathbf{x}_i^k$ is the vector $\mathbf{y}_i^k$ of J_k output parameters describing the data point (the likelihood that the data point belongs to each individual cluster in a mixture). This corresponds to the fuzzy memberships (distributed representation).

For a simpler evaluation, we used a localist representation (winner-takes-all), where only the cluster with the highest cluster membership probability is considered for further processing [see (5) and (6)]

$$M\left(K_j^k|O\right) = \begin{cases} 1 & \text{if } j = \text{argmax}_{j'} f\left(K_{j'}^k|O\right) \\ 0 & \text{if } j \neq \text{argmax}_{j'} f\left(K_{j'}^k|O\right) \end{cases} \quad (15)$$

where $k \in \{\text{size, color, shape}\}$, $j \in \{1, \ldots, J_k\}$.

2) Language Layer: The auditory word-forms are extracted from the language input which are sentences describing the image in the format <size> <color> <shape> (e.g., "small red triangle"). Afterward, individual word-forms are extracted from the sentences and compared to prelearned language models, and the log-scale score $p(\mathbf{w}_t^n|L_i)$ of the audio matching the model is computed. Based on these data, posterior probability can be computed

$$p\left(L_i|\,\mathbf{w}_t^n\right) = \frac{p\left(\mathbf{w}_t^n|L_i\right)p(L_i)}{\sum_{i'} p\left(\mathbf{w}_t^n|L_{i'}\right)p(L_{i'})}, \quad \forall i \in \{1, \ldots, I\}$$
$$\forall t \in \{1, \ldots, T\}, \quad \forall n \in \{1, \ldots, N_t\} \quad (16)$$

where I is the number of language models, T is the number of trials (sentences) and N_t is the number of word-forms in the trial (sentence) t.

An output of this layer for each data point $\mathbf{w}_t^n$ is the vector $\mathbf{y}_i^k$ of I output parameters describing the data point (the likelihood that the data point belongs to each individual language model). This corresponds to the fuzzy memberships (distributed representation). Linguistic and visual inputs are processed simultaneously.

Algorithm 1 Sequential Mapping—Fixed Grammar

Inputs:
language clusters L_i ($i \in 1 : I$), visual clusters $K_j^k \sim N(\mathbf{m}_j^k, \mathbf{S}_j^k)$, $j \in 1 : J_k$, input data $\mathbf{x}^k$ for each feature $k \in \{\text{size, color, shape}\}$, number of clusters J^k for each feature k
Output:
mapping between all visual classes K_j^k and language classes L_i

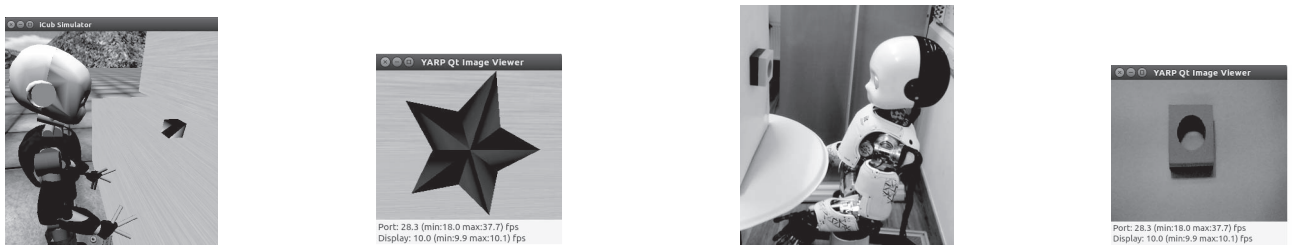
for $k \in \{\text{size, color, shape}\}$ **do**
  $NCl \leftarrow J^k$
  while $NCl > 0$ and $\mathbf{X}^k$ is not empty **do**
    assign each data point from $\mathbf{x}^k$ to visual and language cluster (Winner-takes-all, see Eq. (15))
    for $j = 1 : NCl$ **do**
      for $i = 1 : I$ **do**
        $A_{ij} \leftarrow$ how many times was class i classified as j
      end for
    end for
    $[im, jm] \leftarrow \text{argmax}_i \text{argmax}_j A_{ij}$
    $\mathbf{X}_{del}^k \leftarrow$ data points assigned to K_{jm}^k and L_{im}
    $\Theta_{new}^k \leftarrow N(\mathbf{x}_{del}^k)$ learn Gaussian on the to be deleted data
    $\mathbf{X}^k \leftarrow \mathbf{X^k} \setminus \mathbf{X_{del}^k}$ delete data assigned to both K_{jm}^k and L_{im}
    $NCl \leftarrow NCl - 1$
    relearn $K^k \sim N(\mathbf{m}^k, \mathbf{S}^k)$ on new data $\mathbf{x}^k$ with NCl clusters
  end while
end for
cluster visual data using new θ_{new}^k parameters (cluster centers $\mathbf{m}^k$, covariance matrices $\mathbf{S}^k$) and perform one-step mapping (Model 1)

3) Mapping—Models 1 and 2: After the visual and language data are clustered, the mapping between the two layers must be found. For each cluster L_i in the language layer, a corresponding cluster K_j^k in the visual layer (for each feature $k \in \{\text{size, color, shape}\}$) is found. The mapping is found as follows. For each j and k, we find cluster $L_{\text{kmax}_{jk}}$ from the language layer which will be assigned to cluster K_j^k from the visual layer. In this paper, we compare two different models to find indices $k\text{max}_{jk}$. We compared one-step mapping (see Section II-A) and newly proposed sequential mapping (see Section II-B).

The exact algorithm used to find the mapping between visual and language models in a sequential manner is described in detail in Algorithm 1. Indices $m(i)$ are found sequentially. In each step, the best-mapped data are excluded, and the rest of data are reclustered using GMMs. Then, one-step mapping is performed (see Algorithm 1). An extension of the algorithm for a variable-length sentence is described in the Appendix. Results for a variable length sentence using a fully artificial dataset with a controlled noise level are described in detail in [58].

D. iCub Robotic Platform and iCub Simulator

For the experiment, we used a simulated [48] and a physical stationary [30] iCub robot. The iCub [Fig. 1(c)] is an open-source humanoid robot the size of a three-and-a-half-year-old child, with the fully articulated hands and a head-and-eye system which makes him ideal for cognitive experiments. The iCub simulator has been designed to reproduce, as accurately as possible, the physics and dynamics of the robot and its environment [48]. The simulator and the actual robot have the same interface supporting YARP [31] which is a robot platform for interprocess communication and control of the physical and simulated robots in real-time.



(a) (b) (c) (d)

Fig. 4. Experiment design and corresponding input data. (a) iCub simulator. (b) Blender object. (c) Physical iCub. (d) Real object.

E. Input Data Description and Preprocessing

The input for our model consisted of visual and language data. The visual scene was composed of an object in the center of the scene with a variable position. The visual features (size, shape, and color) of the object also varied. We developed two separate datasets for training and testing purposes. A real-world dataset has visual sensory data acquired from the cameras of the physical iCub robot which observed simple objects placed on the white board in front of his eyes at a distance of 1 m [see Fig. 4(c) and (d)] (210 instances, three sizes, five colors, and seven shapes, in total 70 individual objects; each seen three times in slightly different placement). The simulated dataset is made in the iCub simulator [see Fig. 4(a) and (b)] as Blender-generated virtual objects (432 instances, three sizes, six colors, and six shapes, in total 108 individual objects; each seen four times in slightly different placement).

The spoken language input consisted of sentences pronounced by a non-native English speaker describing the image in the format <size> <color> <shape> (e.g., small red triangle) and was processed simultaneously with the visual input. In the real-world dataset, the tutor was talking to a robot from an approximate 2-m distance with natural background noise. The linguistic input was captured using an external microphone.

1) Speech Recognition: CMU Sphinx (an open-source flexible Markov model-based speech recognizer system) was used for speech recognition [23]. Sphinx itself offers a large vocabulary, but we created our own task-specific smaller vocabulary using the online IMtool that produces a dictionary based on a CMU dictionary and matches its language model.

There is a probabilistic output from the CMU Sphinx. The ten best hypotheses for a matching model with corresponding scores were saved for each utterance (they are log-scale scores of the audio matching the model). Because the scores for the hypothesis of each word in the sentence were needed for further evaluation, the words were pronounced with large pauses, and the end of the sentence was marked by the word "STOP." An output of the language layer is a I-dimensional continuous valued vector, where I is the number of language clusters (corresponding to the number of possible utterances). This vector contains ten nonzero values, and the rest is zero.

2) Image Processing: The image inputs are processed using standard MATLAB functions. First, the image is morphologically opened with a disk-shaped structuring element (*imopen*) to remove the noisy background of the image; then all grayish pixels are removed, and the image is converted from the true color RGB to the gray-scale intensity image by eliminating the hue and saturation information while retaining the luminance (*rgb2gray*). Finally, the intensity image is converted to a binary image using the threshold computed with Otsu's [34] method (*threshold*). An example of the preprocessed image is shown in Fig. 5.

Fig. 5. Image processing: original image, removal of the background, converting to BW image, and filling the holes.

Afterward, the properties of the image regions are measured using the function *regionprops*. Individual visual features (shape, color, and size) are subsequently processed separately. The following features were used: color (three features: average RGB of the selected region), size (six features: perimeter of an object, distance from the centroid to the left corner of the bounding box, and width and length of the bounding box), and shape (13 features: area, centroid, major axis length, eccentricity, orientation, convexArea, FilledArea, EulerNumber, EquivDiameter, solidity, extent, and perimeter). To obtain the shape features, we automatically cropped and resized the image to equalize the size of the objects.

Although the visual model is mainly mathematical and implemented in a very "machine vision" sort of way, the bases of its processing follow biological correlates of mammal vision. More specifically, from the neuroanatomical point of view, this corresponds to the processing of the visual input in the separate higher visual centers in the brain, specifically to the independent processing of the information about the position and identification of an object in the ventral (what) and dorsal (where) neural pathways, respectively [32]. Individual object properties are identified in the separate visual centers of the occipital lobe.

F. Evaluation

In the case of the supervised GMM, growing when required neural gas (GWR), and SOM algorithms, data were divided into training and validation datasets in the ratio 70:30. For the unsupervised GMM, hidden Markov model, and k-means algorithms, we computed the accuracy differently. After the data

TABLE I
COMPARISON OF CLUSTERIZATION AND CLASSIFICATION ACCURACY OF VISUAL DATA. THE MEAN AND STANDARD DEVIATION FROM 100 REPETITIONS ARE VISUALIZED

	Real data			Blender		
Accuracy [%]	Size	Color	Shape	Size	Color	Shape
GMM sup.	83.3 ± 0.0	99.0 ± 0.0	81.4 ± 0.0	98.6 ± 0.0	97.9 ± 0.0	93.1 ± 0.0
GMM unsup.	76.2 ± 6.8	76.1 ± 9.1	56.1 ± 6.2	74.2 ± 10.1	60.9 ± 9.0	64.3 ± 7.2
K-means	67.8 ± 6.2	81.2 ± 1.1	53.1 ± 4.2	66.3 ± 0.2	77.1 ± 10.7	72.8 ± 6.9
SOM	69.6 ± 5.6	78.9 ± 6.8	54.2 ± 4.1	66.1 ± 4.2	81.7 ± 7.6	59.3 ± 6.2
GWR	89.9 ± 2.1	99.5 ± 0.4	76.6 ± 1.4	88.9 ± 0.7	98.1 ± 0.9	94.2 ± 0.6

TABLE II
COMPARISON OF ONE-STEP MAPPING AND SEQUENTIAL MAPPING FOR DATA FROM THE ICUB SIMULATOR (BLENDER) AND THE PHYSICAL ICUB (REAL DATA). THE MEAN AND STANDARD DEVIATION FROM 100 REPETITIONS ARE VISUALIZED

	Real data			Blender		
Accuracy [%]	Size	Color	Shape	Size	Color	Shape
Vision	76.2 ± 6.8	76.1 ± 9.1	56.1 ± 6.2	74.2 ± 10.1	60.9 ± 9.0	64.3 ± 7.6
Language	70.6 ± 0.0	82.4 ± 0.0	77.5 ± 0.0	98.1 ± 0.0	96.5 ± 0.0	98.1 ± 0.0
One-step mapping	54.1 ± 4.1	58.2 ± 10.3	52.2 ± 4.9	67.3 ± 8.2	56.2 ± 6.1	61.9 ± 3.2
Sequential mapping	74.2 ± 15.1	87.1 ± 10.2	72.9 ± 5.1	96.1 ± 31.2	95.2 ± 1.2	92.1 ± 0.9

is clustered, each cluster is assigned to the class that appears most frequently in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned data points (compared to the manual true labels) and dividing this number by the total number of data points.

The accuracy of the learned mapping is calculated in the following manner. We cluster output activations from the visual layer and assign each data point to the most probable cluster. Then, we find indices *m(i)* for all clusters as defined in (2) for one-step mapping and (8) for sequential mapping. Based on this mapping, we can assign each data point to the language label. These language labels are subsequently compared to the ground truth. Accuracy is then computed as

$$\text{acc} = \frac{\text{TP}}{N} \tag{17}$$

where TP (true positive) is the number of correctly assigned data points, and N is the number of all data points.

III. EXPERIMENTAL RESULTS

The first part of the results is dedicated to the performance of the model in the real-world scenario. The robot interacts with a human in a noisy condition that distorts speech input.

A. Vision

In the first stage, we evaluated the Vision subpart of the model. Several algorithms were compared: namely the GMM algorithm, supervised GMM algorithm, *k*-means, SOM, and GWR algorithm [2], [26]. The SOM and GWR had 100 nodes. The results for the real-world dataset and the simulated dataset with Blender objects can be seen in Table I. Although the SOM and GWR algorithms are considered unsupervised algorithms, we adopted a technique for labeling inputs; thus, they should be compared with supervised algorithms. The number of clusters is also overestimated (the number of nodes corresponds to the number of clusters). It indicates that these algorithms partly overfit the data so we divided the set into testing and validation data.

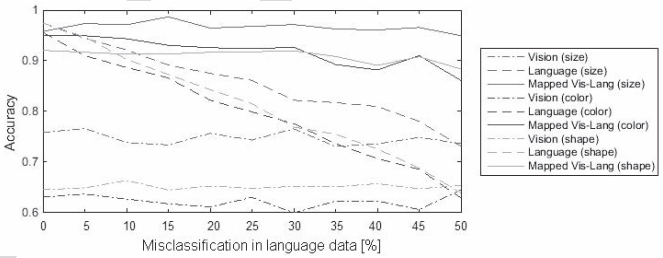


Fig. 6. Dependence of mapping accuracy on the misclassification in the language data for a fixed-length sentence (mean values over 50 repetitions are visualized). Different colors correspond to different visual features (red: size, blue: color, and green: shape). Visual data are generated in Blender and acquired through the iCub simulator, and language data are processed using Sphinx 4.

B. Mapping

The performance of one-step mapping (vision and language are mapped in one step based on the frequency of co-occurrence) and sequential mapping (see Algorithm 1) is shown in Table II. We calculated the accuracy for real-world data from physical iCub and for the Blender objects placed in the iCub simulator. Language accuracy for Blender dataset is much higher compared to the real-world data as a tutor was speaking directly into the microphone.

Tolerance of the sequential mapping to misclassification in the language data is visualized in Fig. 6 for visual data from the iCub simulator in combination with the language data processed by Sphinx 4. The misclassification is added to the language data subsequently and evenly to all classes (the given proportion of the language inputs was randomly changed to random words). The misclassification (a synthetic error) was added artificially but can be interpreted either as white noise added to the input data or mistakes in labeling perceived

objects. We grouped them together into a misclassification variable. The visual data are left intact so the only cause of the observed variations in the accuracy is initialization. As can be seen, the accuracy of sequential mapping remains very stable even though the accuracy of the language decreases and outperforms the language and vision for almost all values of the misclassification.

IV. Conclusion

Current models of vision-to-language mapping often make use of cross-situational learning while relying directly on statistical co-occurrence of meaning-referent pairs: the ones with the highest co-occurrence are mapped together (e.g., [41]). This approach shows inferior performance in cases where the clustering of data is difficult (e.g., due to the high number of objects, high noise level or overlapping classes, etc.). Therefore, we extended this basic model and introduced a new more robust and noise-resistant mapping procedure. Our approach incorporates situation-time dynamics and mutual exclusivity and is able to deal with a nonequal number of classes in individual subdomains.

Mathematical formulation of the newly proposed mapping is provided (see Section II-B), and results for simulated and real-world data from the iCub robot are compared to one-step mapping (see Table II). It was shown that the method is able to find mapping between language and vision, and the method improves the accuracy of both individual subdomains and shows very good resistance to noise or misclassification in language (see Fig. 6). How to map in an unsupervised manner several clusterings (e.g., for vision, action, and language) is an important question not only in cognitive modeling but also in general machine learning, where data acquired from different sensors or in different situations can be independently clustered and mapped to each other. A more detailed discussion of the results follows.

The trivial one-step mapping can be imagined as basic Hebbian learning. Our extension can be likened to Hebbian learning with inhibitory connections. In recent years, several approaches finding an alternative to the basic approach appeared [29], [57]. Our model partially stems from the McMurray *et al.* [29] approach, who showed that associative learning can be sufficient for language acquisition and that the main components of this type of learning are an online competition of models and pruning incorrect associations which enable gradual improvement of associations between models.

The mutual exclusivity principle is guaranteed in our method thanks to the inhibitory connections which are gradually created among models. Once the mapping between the referent and the meaning is found, the connection from a given meaning to other referents is inhibited. Dynamic competition is addressed in the following way: when any association of meaning and referent is found, other models compete again for resources. First, well-mapped data are deleted, and then, the resting data are reclustered. Furthermore, the likelihoods can associate each data point to many separate models instead of binary membership. In this way, our algorithm also emulates how the similarity of two-word forms can affect learning as it occurs in children's development [36]. The algorithm also enables mapping together data in a case where we have an uneven number of clusters in both subdomains.

However, the principle of mutual exclusivity is not suitable for further stages of language acquisition, namely, the learning of polysemic words. A polyseme is a word or phrase with different but related senses (e.g., wood as a piece of tree or the area with trees). The homonyms are a subset of the polysemes, but the difference between the homonyms and the polysemes is subtle and fuzzy. Homonyms represent a group of words that share similar spelling (homographs) and the same sound (homophones) but have different and unrelated meanings, for example, the homograph bank standing for embankment or place where money is kept. The learning of polysemic words violates the principle of mutual exclusivity as the dynamic competition does not allow to map one word to different visual models. We are aware of this problem, and we would like to extend it in the future iteration of the model. Similar to humans who have to learn polysemic words as an exception, we will incorporate this principle into the next version of our model.

In the next part, we analyze the ability of our architecture to deal with ambiguous inputs. The mapping will find reliable labeling for the visual input data (more generally for data from any other modality) with a possibility of incorporating the fuzziness of this mapping. For some concepts, finding an unambiguous mapping is very easy; for others, it is much more difficult or impossible (such as abstract words, e.g., the love has no dominant color, but the sky is usually blue). Since the mapping is established only among the clusters where it makes sense, dealing with a lot of redundant information is avoided. A similar idea is used in classification algorithms which use sparse matrices (e.g., [12] and [13]). We also analyzed the strong and weak points of the algorithms adopted in our architecture. First, the ability of different algorithms to classify unimodal visual data was compared. As expected, for data which are well separated and mainly spherically distributed (this is generally a case for simulated and artificially generated data), the k-means algorithm outperformed the GMM algorithm. However, for nonspherical real data, the GMM algorithm generally performed better (see Table I for the comparison of the performance on simulated data placed in an iCub simulator and data from real iCub robot cameras). As the unsupervised algorithms are highly dependent on the initialization, it can be seen that the standard deviation of the data is quite high even though 20 repetitions were averaged. We should conclude that the performance of the algorithms in our tasks reflects their fundamental advances and limitations. We are still missing the algorithm that is able to cope with highly variable datasets in terms of their statistical properties.

The most significant differences between data acquired from the real and the simulated iCub can be found in the visual subdomain. Affected by natural light and slightly altered points of view, the shading of objects and their bevel differed. This resulted in a different projection of objects to iCub cameras and in the less accurate classification of shapes for the real-world dataset. For colors, different lighting changed the true color of the objects. Nevertheless, as can be seen in Table II, this did not affect the ability of the algorithm to discriminate among individual colors. The real-world objects had a

variable size which resulted in a higher overlap of clusters for individual sizes compared to the Blender dataset.

Afterward, we focused on the mapping between vision and language and compared two different approaches: one-step mapping to sequential mapping which in a stepwise manner finds the best-mapped clusters while constantly relearning clusterization of visual data. As can be seen in the results in Table II, the novel sequential mapping led to an improvement in effectiveness compared to the method which maps vision to language in one step. This can be seen in more accurate mapping which leads to a better estimation of the clustered data labels and consequently to lower classification error for all of the evaluated datasets.

The accuracy of multimodal mapping outperforms vision, language or both. This is an important finding, because sequential mapping improves not only the accuracy of visual clustering but can also fix mistakes in language recognition, which provides the labels. Furthermore, this result suggests that we are able not only to find mapping between more clusterings, but we can also improve the clusterization accuracy by combining individual classifiers. This is not easily seen in the presented dataset as there is high recognition accuracy of Sphinx software (especially in the case of sentences recorded for the simulated dataset). Therefore, we also tested whether the misclassification (or noise) in the language data affects the correct mapping between vision and language. The result can be seen in Fig. 6. Although a synthetic error (misclassification) is added to the language data, the ability to find the mapping remains nearly intact. The mapping accuracy decreases only very slightly and remains at around 90% because the accuracy of language recognition drops from the original approximately 95% to approximately 70% (depending on the specific visual feature). In the future, we would like to explore the effect of noise in individual modalities on classification accuracy in more detail by adding a controlled amount of noise directly to the input. The error propagation from individual input modalities to the multimodal layer was explored, for example, in [37]. They showed that the visual context can steer speech recognition and vice versa.

We also analyzed how the complexity of environment affects the accuracy of the our architecture. We suppose that the performance of one-step mapping will decrease with the increasing complexity of the task (more clusters, higher overlap, and higher dimensionality) as it is more difficult to find reliable clustering of the data. This hypothesis is supported by our preliminary results for clustering body parts from simultaneous tactile and linguistic input [59] and by the results presented in this paper in Table II. The quality of one-step mapping correlates with the quality of the visual data clustering. For Blender data, the worst performance was achieved for mapping words to the visual feature shape ($52 \pm 5\%$) and for physical iCub for the feature color ($62 \pm 3\%$). The feature shape has the highest number of clusters (10), and the feature color has the second highest number of clusters (9) and the highest overlap of the clusters for the physical iCub. Real-world tasks are much more complex, and we can expect tens of different object shapes. In that case, the clustering performance is crucial, and one-step mapping would not be able to provide reliable mapping. It can be seen from the results that the proposed mapping which enables gradual re-estimation of the model parameters and works in a dynamic fashion achieves much higher accuracy even for the cases where one-step mapping fails. We suppose that the mapping accuracy of the proposed method decreases more slowly with the decreasing accuracy of clustering of individual modalities. Unfortunately, this factor was not studied in our restricted scenario, but the preliminary results on mapping tactile and linguistic input [59] support this hypothesis. We plan to investigate this phenomenon more deeply in future research.

The language dataset differs considerably from natural language. However, the dataset reflects some characteristics from the findings of Werker and McLeod [53] as infant-directed words are usually kept short with large pauses between words. Moreover, Brent and Siskind [3] showed that frequency of exposure to a word in isolation predicts better whether that word will be learned than the total frequency of exposure to that word. In addition, Snow [42] found that mother's speech to two-year-old is much simpler and less redundant than their speech to ten-year-old. This indicates that young children have available a sample of speech which is simpler, more redundant and less confusing than normal adult speech.

The proposed algorithm was tested on language to vision mapping and on language to tactile mapping [59], and it can be easily extended to language to any other modalities mapping. The mapping between multiple modalities and words was researched in [47]. Fazly *et al.* [14] proposed a probabilistic model of cross-situational learning where they considered sentences that contain objects and their motion. Monaghan *et al.* [33] studied differences between cross-situational learning of nouns and verbs on human participants as an extension of Tomasello and Akhtar's [50] work and Schwartz and Terrell's [39] work. Monaghan *et al.* [33] considered learning of verbs as difficult as learning of nouns when presented in a syntactic context. The authors noticed that nouns are learned quicker, but verbs and nouns can be acquired simultaneously.

An important direction for extensions of the proposed model deals with the grounding of language related not just to static sequences but also to events which require a temporal axis. For example, Crick and Scassellati [7] studied the interconnection between verbal narratives and episodes of intentional relative motion with the goal that the robot can learn from observing a game the rules of the game, relationships between players and their goals and intentions and then participate in the play.

The main goal of this paper is to analyze mapping between modalities. Thus, we kept processing of the individual modalities deliberatively simple for better understanding of cross-situational learning. We performed further experiments with an artificially generated dataset where the visual features, as well as noise in the linguistic and the visual domain, were under full control. These experiments are described in detail in [58].

APPENDIX

Here we describe an extension of the proposed algorithm presented in Section II-B for a case where we have sentences with variable structure. This means that we cannot

Algorithm 2 Sequential Mapping—Variable Sentence

Inputs:
language clusters L_i ($i \in 1:I$), visual clusters $K_j^k \sim N(\mathbf{m}^k, \mathbf{S}^k)$, $j \in 1:J^k$ for each feature k, visual input $\mathbf{x_t} = \{\mathbf{x}_t^1, \ldots, \mathbf{x}_t^K\}$ and corresponding language data $W_t = \{\mathbf{w}_t^1, \ldots, \mathbf{w}_t^{N_t}\}$ for each trial t and number of clusters J^k
Output:
mapping between all visual classes K_j^k and language classes L_i

while $\sum J^k > 0$ and $\mathbf{X}$ is not empty **do**
  $l_t^n \leftarrow$ assign each word $\mathbf{w}_t^n$ from each sentence t to a language cluster (Winner-takes-all, see Eq. (15), $l_t^n = \text{argmax}_i(P(w_t^n|L_i))$
  for $k \in$ {size, color, orientation, texture, shape} **do**
    $\mathbf{v}_t^k \leftarrow$ assign each data point $\mathbf{x}_t^k$ to a visual cluster (Winner-takes-all, see Eq. (15), $v_t^k = \text{argmax}_j(P(\mathbf{x}_t^k|K_j^k))$
    for $j = 1 : J^k$ **do**
      for $i = 1 : I$ **do**
        $T_{ij}^k \leftarrow$ how many times did visual class i co-occur with language class j ($T_{ij}^k = \sum_{t;\mathbf{v}_t^k==j} \sum_n \delta(l_t^n, i)$), where δ is Kronecker delta
      end for
    end for
  end for
  $[km, im, jm] \leftarrow \text{argmax}_k \text{argmax}_i \text{argmax}_j T_{i,j}^k$ (the visual cluster K_{jm}^{km} is mapped to the language cluster L_{im})
  $\mathbf{X}_{del}^{km} \leftarrow$ data points assigned to K_{jm}^{km} and L_{im}
  $\theta_{new,J^k}^k \leftarrow N(\mathbf{x}_{del}^k)$ learn Gaussian on the to be deleted data
  $\mathbf{X}^{km} \leftarrow \mathbf{X}^{km} \setminus \mathbf{X}_{del}^{km}$ delete data assigned to K_{jm}^{km} and L_{im}
  $J^{km} \leftarrow J^{km} - 1$
  relearn $K_j^{km} \sim N(\mathbf{m}^k, \mathbf{S}^k)$ on new data $\mathbf{x}^{km}$ with J^{km} clusters
end while
cluster visual data using new θ_{new}^k parameters (cluster centers $\mathbf{m}^k$, covariance matrices $\mathbf{S}^k$) and perform one-step mapping (Model 1)

directly associate words from the sentence with individual visual features, and therefore, associations to all visual features must be taken into account. The whole algorithm is described in Algorithm 2 in the form of a pseudocode.

Algorithm 2 can be further extended when we incorporate the language model and corresponding probabilities of sequences of individual visual features (see [58]).

REFERENCES

[1] L. W. Barsalou, "Grounded cognition," *Annu. Rev. Psychol.*, vol. 59, pp. 617–645, Jan. 2008.

[2] F. B. Klein. (2016). *GWR and GNG Classifier—File Exchange—MATLAB Central*. [Online]. Available: http://uk.mathworks.com/matlabcentral/fileexchange/57798-gwr-and-gng-classifier

[3] M. R. Brent and J. M. Siskind, "The role of exposure to isolated words in early vocabulary development," *Cognition*, vol. 81, no. 2, pp. B33–B44, 2001.

[4] C. Büchel, C. Price, and K. Friston, "A multimodal language region in the ventral visual pathway," *Nature*, vol. 394, no. 6690, pp. 274–277, 1998.

[5] A. Cangelosi, A. Greco, and S. Harnad, "From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories," *Connection Sci.*, vol. 12, no. 2, pp. 143–162, 2000.

[6] S. Coradeschi, A. Loutfi, and B. Wrede, "A short review of symbol grounding in robotic and intelligent systems," *Künstliche Intelligenz*, vol. 27, no. 2, pp. 129–136, 2013, doi: 10.1007/s13218-013-0247-2.

[7] C. Crick and B. Scassellati, "Controlling a robot with intention derived from motion," *Topics Cogn. Sci.*, vol. 2, no. 1, pp. 114–126, 2010.

[8] J. C. Culham and K. F. Valyear, "Human parietal cortex in action," *Current Opin. Neurobiol.*, vol. 16, no. 2, pp. 205–212, 2006.

[9] M. Daoutis and N. Mavridis, "Towards a model for grounding semantic composition," in *Proc. Artif. Intell. Simulat. Behav. (AISB)*, 2014.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.

[11] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 2625–2634.

[12] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 1873–1879.

[13] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[14] A. Fazly, A. Alishahi, and S. Stevenson, "A probabilistic computational model of cross-situational word learning," *Cogn. Sci.*, vol. 34, no. 6, pp. 1017–1063, 2010.

[15] M. Fleischman and D. Roy, "Grounded language modeling for automatic speech recognition of sports video," in *Proc. ACL*, Columbus, OH, USA, 2008, pp. 121–129.

[16] M. C. Frank, N. D. Goodman, and J. B. Tenenbaum, "Using speakers' referential intentions to model early cross-situational word learning," *Psychol. Sci.*, vol. 20, no. 5, pp. 578–585, 2009.

[17] V. Gliozzi, J. Mayor, J. F. Hu, and K. Plunkett, "Labels as features (not names) for infant categorization: A neurocomputational approach," *Cogn. Sci.*, vol. 33, no. 4, pp. 709–738, 2009.

[18] E. Goldberg, N. Driedger, and R. I. Kittredge, "Using natural-language processing to produce weather forecasts," *IEEE Expert*, vol. 9, no. 2, pp. 45–53, Apr. 1994.

[19] P. Gorniak and D. Roy, "Speaking with your sidekick: Understanding situated speech in computer role playing games," in *Proc. AIIDE*, Marina Del Rey, CA, USA, 2005, pp. 57–62.

[20] S. Harnad, "The symbol grounding problem," *Phys. D Nonlin. Phenomena*, vol. 42, nos. 1–3, pp. 335–346, 1990.

[21] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 3128–3137.

[22] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing robot embodiments in a guided discovery learning interaction with children," *Int. J. Soc. Robot.*, vol. 7, no. 2, pp. 293–308, 2015.

[23] P. Lamere *et al.*, "The CMU SPHINX-4 speech recognition system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1. Hong Kong, 2003, pp. 2–5.

[24] P. Li, I. Farkas, and B. MacWhinney, "Early lexical development in a self-organizing neural network," *Neural Netw.*, vol. 17, nos. 8–9, pp. 1345–1362, 2004.

[25] E. M. Markman, "Constraints children place on word meanings," *Cogn. Sci.*, vol. 14, no. 1, pp. 57–77, 1990.

[26] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," *Neural Netw.*, vol. 15, nos. 8–9, pp. 1041–1058, 2002.

[27] N. Mavridis, "Grounded situation models for situated conversational assistants," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.

[28] N. Mavridis, "A review of verbal and non-verbal human–robot interactive communication," *J. Robot. Auton. Syst.*, vol. 63, pp. 22–35, Jan. 2015.

[29] B. McMurray, J. S. Horst, and L. K. Samuelson, "Word learning emerges from the interaction of online referent selection and slow associative learning," *Psychol. Rev.*, vol. 119, no. 4, pp. 831–877, 2012.

[30] G. Metta *et al.*, "The iCub humanoid robot: An open platform for research in embodied cognition," in *Proc. ACM 8th Workshop Perform. Metrics Intell. Syst.*, Gaithersburg, MD, USA, 2008, pp. 50–56.

[31] G. Metta, P. Fitzpatrick, and L. Natale, "YARP: Yet another robot platform," *Int. J. Adv. Robot. Syst.*, vol. 3, no. 1, pp. 43–48, 2006.

[32] M. Mishkin, L. G. Ungerleider, and K. A. Macko, "Object vision and spatial vision: Two cortical pathways," *Trends Neurosci.*, vol. 6, no. 10, pp. 414–417, 1983.

[33] P. Monaghan, K. Mattock, R. A. I. Davies, and A. C. Smith, "Gavagai is as Gavagai does: Learning nouns and verbs from cross-situational statistics," *Cogn. Sci.*, vol. 39, no. 5, pp. 1099–1112, 2015, doi: 10.1111/cogs.12186.

[34] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[35] W. V. Quine, "On the reasons for indeterminacy of translation," *J. Philos.*, vol. 67, no. 6, pp. 178–183, 1970.

[36] T. Regier, "The emergence of words: Attentional learning in form and meaning," *Cogn. Sci.*, vol. 29, no. 6, pp. 819–865, 2005.

[37] D. Roy and N. Mukherjee, "Towards situated speech understanding: Visual context priming of language models," *Comput. Speech Lang.*, vol. 19, no. 2, pp. 227–248, 2005.

[38] D. K. Roy, "Learning visually grounded words and syntax for a scene description task," *Comput. Speech Lang.*, vol. 16, no. 3, pp. 353–385, 2002.

[39] R. G. Schwartz and B. Y. Terrell, "The role of input frequency in lexical acquisition," *J. Child Lang.*, vol. 10, no. 1, pp. 57–64, 1983.

[40] J. M. Siskind, "A computational study of cross-situational techniques for learning word-to-meaning mappings," *Cognition*, vol. 61, nos. 1–2, pp. 39–91, 1996.

[41] K. Smith, A. D. M. Smith, R. A. Blythe, and P. Vogt, *Cross-Situational Learning: A Mathematical Approach* (LCNS 4211). Heidelberg, Germany: Springer, 2006, pp. 31–44.

[42] C. E. Snow, "Mothers' speech to children learning language," *Child Dev.*, vol. 43, no. 2, pp. 549–565, 1972.

[43] G. Spitsyna, J. E. Warren, S. K. Scott, F. E. Turkheimer, and R. J. Wise, "Converging language streams in the human temporal lobe," *J. Neurosci.*, vol. 26, no. 28, pp. 7328–7336, 2006.

[44] F. Stramandinoli, D. Marocco, and A. Cangelosi, "The grounding of higher order concepts in action and language: A cognitive robotics model," *Neural Netw.*, vol. 32, pp. 165–173, Aug. 2012.

[45] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adapt. Behav.*, vol. 13, no. 1, pp. 33–52, 2005.

[46] M. Taddeo and L. Floridi, "Solving the symbol grounding problem: A critical review of fifteen years of research," *J. Exp. Theor. Artif. Intell.*, vol. 17, no. 4, pp. 419–445, 2005.

[47] A. Taniguchi, T. Taniguchi, and A. Cangelosi, "Multiple categorization by iCub: Learning relationships between multiple modalities and words," in *Proc. IROS Workshop Mach. Learn. Methods High Level Cogn. Capabilities Robot.*, 2016.

[48] V. Tikhanoff *et al.*, "An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator," in *Proc. ACM 8th Workshop Perform. Metrics Intell. Syst.*, Gaithersburg, MD, USA, 2008, pp. 57–61.

[49] V. Tikhanoff, A. Cangelosi, and G. Metta, "Integration of speech and action in humanoid robots: iCub simulation experiments," *IEEE Trans. Auton. Mental Develop.*, vol. 3, no. 1, pp. 17–29, Mar. 2011.

[50] M. Tomasello and N. Akhtar, "Two-year-olds use pragmatic cues to differentiate reference to objects and actions," *Cogn. Dev.*, vol. 10, no. 2, pp. 201–224, 1995.

[51] M. Vavrečka and I. Farkaš, "A multimodal connectionist architecture for unsupervised grounding of spatial language," *Cogn. Comput.*, vol. 6, no. 1, pp. 101–112, 2014.

[52] P. Vogt, "Language evolution and robotics: Issues on symbol grounding," in *Artificial Cognition Systems*. Hershey, PA, USA: IGI Glob., 2006, p. 176.

[53] J. F. Werker and P. J. McLeod, "Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness," *Can. J. Psychol. Revue Can. De Psychol.*, vol. 43, no. 2, pp. 230–246, 1989.

[54] F. Xu and J. B. Tenenbaum, "Word learning as Bayesian inference," *Psychol. Rev.*, vol. 114, no. 2, pp. 245–272, 2007.

[55] C. Yu and L. B. Smith, "Rapid word learning under uncertainty via cross-situational statistics," *Psychol. Sci.*, vol. 18, no. 5, pp. 414–420, 2007.

[56] C. Yu and L. B. Smith, "Modeling cross-situational word-referent learning: Prior questions," *Psychol. Rev.*, vol. 119, no. 1, pp. 21–39, 2012.

[57] D. Yurovsky, C. Yu, and L. B. Smith, "Competitive processes in cross-situational word learning," *Cogn. Sci.*, vol. 37, no. 5, pp. 891–921, 2013.

[58] K. Štěpánová, "Hierarchical probabilistic model of language acquisition," Ph.D. dissertation, Dept. Cybern., Czech Tech. Univ. Prague, Prague, Czech Republic, 2016.

[59] K. Štěpánová *et al.*, "Where is my forearm? Clustering body parts from simultaneous tactile and linguistic input," in *Proc. Cogn. Artif. Life (KUZ XVII)*, 2017.

Karla Štepánová received the master's degree in condensed matter physics from Charles University, Prague, Czech Republic, and the Ph.D. degree in artificial intelligence and biocybernetics from Czech Technical University in Prague, Prague, in 2017.

She is a Researcher with the Czech Institute of Informatics, Robotics, and Cybernetics, Prague. She was a Visiting Researcher with Plymouth University, Plymouth, U.K., in 2016. Her current research interests include probabilistic models of cognition, unsupervised learning, language acquisition, and multimodal integration.

Frederico Belmonte Klein was born in Porto Alegre, Brazil, in 1982. He received the Electrical Engineering degree and the medical degree from the Federal University of Rio Grande do Sul, Porto Alegre, in 2006 and 2014, respectively. He is currently pursuing the Ph.D. degree in cognitive robotics with the University of Plymouth, Plymouth, U.K.

Angelo Cangelosi received the degree in psychology and cognitive science from the University of Rome La Sapienza, Rome, Italy, and the University of Genoa, Genoa, Italy.

He is a Professor of artificial intelligence and cognition and the Director of the Centre for Robotics and Neural Systems with Plymouth University, Plymouth, U.K. and was a Visiting Scholar with the University of California at San Diego, San Diego, CA, USA, and the University of Southampton, Southampton, U.K. His current research interests include language grounding and embodiment in humanoid robots, developmental robotics, human–robot interaction, and on the application of neuromorphic systems for robot learning.

Michal Vavrečka received the Ph.D. degree in general psychology from the Faculty of Social Studies, in 2008.

He is with the Czech Institution for Informatics Robotics and Cybernetics, Czech Technical University in Prague, Prague, Czech Republic. He develops cognitive architecture for symbol grounding, language acquisition, and knowledge representation.

AUTHOR QUERIES
AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ1: Please confirm/give details of funding source.
AQ2: Please provide the postal code for "Plymouth University Plymouth, U.K."
AQ3: Please provide the page range for References [9], [47], and [59].
AQ4: Please provide the department name for Reference [27].
AQ5: Please confirm if the location and publisher information for Reference [41] is correct as set.
AQ6: Please specify the type of the degree and year attained by the author "A. Cangelosi."
AQ7: Please provide the organization name and location for the Ph.D. degree obtained by "M. Vavrečka" biography.