2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

The Impact of Motor Action on Language Acquisition and Action-Verb Learning in a Robot

Zakaria Lemhaouri, Laura Cohen, Ann Nowé and Lola Cañamero

***Abstract*—In humans, the acquisition of a new motor skill is associated with the development of a wide range of cognitive areas and can create contexts in which new cognitive capacities develop. Motor development is linked to language development in infants, as crawling and walking promote active exploration of the environment, while manipulating objects and pointing draw the caregiver's attention and help establish joint attention. Together, these motor experiences broaden communication contexts and support the learning of nouns (object-based words) and verbs (action-based words). However, many questions remain unanswered about how children's actions influence language development, qualitatively and quantitatively, and how they help the acquisition of different types of words, particularly the learning of verbs. In this paper, we propose a robot architecture to study how gestures can affect early language learning. The architecture follows the developmental robotics paradigm, i.e. inspired by the way human children develop and acquire language according to multiple developmental theories. The experimental results demonstrate that enabling the robot to produce gestures expands its vocabulary size and facilitates the acquisition of verbs. These results are in line with the finding that verb learning lags behind noun learning since the acquisition of verbs depends more on motor abilities and requires the maturation of motor development.**

***Index Terms*—Developmental robotics, Language acquisition, Verb learning, Symbol grounding, Social affective robotics.**

I. Introduction

Language development in infants is influenced by several interacting factors, including biological predispositions, the quantity and quality of language exposure, interaction with caregivers, socioeconomic context, and cognitive abilities [1]. Among these factors, motor skills also play a significant role. The supportive role of children's actions in language learning was investigated in several studies. Findings show that the names of objects manipulated by children appear more frequently in infants' vocabularies [2]. Learning to walk is associated with a significant increase in receptive and productive language [3] [4]. Active infants with multiple motor skills and mobility exhibit greater environmental exploration and can pick up a greater number of objects, encouraging shared communication between parent and child [5]. Manual motor behavior plays a role in early language development in infants since there is a link between infant object manipulation and caregiver verbal labeling of actions [6] used by the child to learn. This learning typically takes place when an infant's actions prompt the caregiver to engage with them. Children are very sensitive to the caregiver's reaction, and rely on them to acquire new knowledge [7]. In terms of learning the two key grammatical forms of words, namely nouns and verbs, several studies have demonstrated that the learning of verbs lags far behind the learning of nouns [8] [9], and that children's vocabularies contain a higher proportion of nouns than verbs [10]. This early advantage of nouns can be explained by the nature of the two categories: for a 24-month-old child, language acquisition involves mapping new nouns to object categories and new verbs to event and action categories [8]; most nouns words are concrete, conceptually stable and can be learned by observation [8] [9]. Furthermore, since the meaning of a given verb depends on the arguments (nouns) it takes, children may need to develop a repertoire of nouns before they can easily learn verbs [8]. Other possible explanations for this disparity include the fact that learning verbs is embodied and requires the development of children's motor actions, as these provide important clues to the meaning of verbs [11]. Understanding verbs is more directly related to motor development than understanding nouns [12]. Due to their embodiment and ability to perform actions, robot language models can be used to help investigate and validate these conclusions. We propose a robot architecture to validate the findings that highlight the relationship between motor development and language development, and to explain how the former affects the emergence of verb acquisition. The proposed architecture follows the developmental robotics paradigm. This field of robotics uses robots as research tools to study and model the emergence and development of cognition and action [13]. To validate models proposed by neuroscientists and developmental psychologists, the robot must learn in the manner of a human child, i.e. through embodied online interactions with its environment and social interactions [14]. Next section further explores some works using the developmental paradigm to study language acquisition in children.

II. Related work

Learning the semantic aspect of the action verb has received particular attention in robotics research, as it is necessary for a robot to understand commands, and since the vast majority of instructions given to robots are in verbal form (e.g., follow, open, close ....). A robot architecture was proposed in [15] for grounded verbal semantics that aims to associate the

Z. Lemhaouri is with ETIS Lab, CY Cergy Paris University - ENSEA - CNRS UMR8051, France; the Artificial Intelligence Lab, Vrije Universiteit Brussel (VUB), Belgium; and laboratoire Integr'It, Esiee-It, France

L. Cohen is with the ETIS Lab, CY Cergy Paris University - ENSEA - CNRS UMR8051, France

A. Nowé is with the Artificial Intelligence Lab, Vrije Universiteit Brussel (VUB), Belgium.

L. Cañamero is with the ETIS Lab, CY Cergy Paris University - ENSEA - CNRS UMR8051, France

verbal phrase and its corresponding sequences of primitive actions with the robot's actuator. This interactive approach was proposed to address the limitations of techniques that use only demonstration and observation learning [16] [17] to learn the semantic of verbs by making the robot more interactive and learning when to ask questions that guide the robot's learning. In [18], the authors introduced a robot architecture for learning the meaning of action verbs based on human-robot dialogues and natural language explanations, the architecture enables the robot to use these actions in new and different contexts. An algorithm for understanding natural language commands, enabling a human to control a simulated robot to carry out instructions, was introduced in [19], and then used the collected data to train a model of command verb meaning from a corpus of natural language commands. In [20], a machine learning approach was introduced to map object-manipulation verbs onto sensory inputs and motor outputs. After learning, when it receives an instruction, such as to move an object, the robot uses Hidden Markov Models to generate the motion trajectories learned by user demonstration. Although these models investigated how the semantic and pragmatic aspects of natural language can be embedded in motor actions or sensory representation, their aim was only to help the robot understand and execute vocal commands, but not to assess how motor abilities contribute to language learning or to follow the same developmental trajectory as language in humans.

The method we propose here follows models of language development that draw parallels with the way children learn and develop skills. Models such as [21] use interaction with the robot (through description, feedback, demonstration...) to help the robot learn simple grammar and basic language skills. The study in [22] explored how action words can be grounded in sensorimotor experiences using an artificial neural network. In [23], a developmental robot model [24] was used to replicate a study [25] showing how the body can impact the cognitive processes involved in language learning. The results indicate that, as in humans, body posture plays a central role in the name-object correspondence and language learning abilities of robots.

III. Proposed Approach

To evaluate how motor development can affect language learning and support the acquisition of verbs, we propose a robot language learning architecture in which the robot learns language in a functional way. This approach aligns with M.K. Halliday's functionalist theory of language acquisition [26], which posits that children learn language as a means to achieve other goals—such as requesting objects or expressing needs. In this view, infants take on a central and active role in language learning, rather than being regarded as passive learners . Learning also occurs when parents describe verbally their infants' actions [11]. We propose to extend these works [27]–[30] that were limited by the absence of the audio modality. By adding it, we give the robot the ability to map words to actions using a recurrent neural network. In this way, it learns verbs and the names of its actions. The learning methods are also based on connectionism and reinforcement learning. The robot's learning scenario is summarized in Figure 1. To make this approach consistent with the functionalist view of language, the robot is motivated to achieve goals that can be difficult to self-fulfill without human interaction. Thus, the robot uses its proto-word vocabulary to communicate its need. Successful interactions help the robot to ground its vocabulary to the sensory data and to the goal that can be achieved with the acquired word. The robot can also perform actions to manipulate objects and point to them. The caregiver provides verbal labels for these gestures, and the robot uses them to learn the names of the gestures. In this way, the robot learns verbs. The choice of this learning method from the caregiver labeling is justified by the fact that parents tend to respond to and label their children's gestures [31].

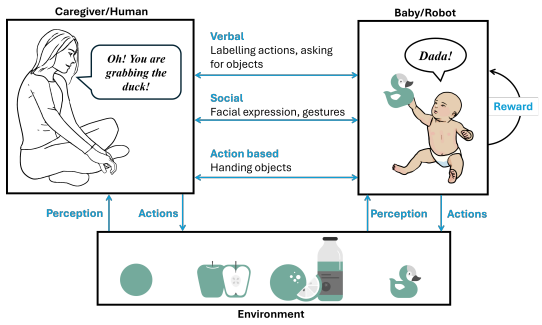


Fig. 1. The robot's language learning scenario. The robot is driven to achieve goals that can be accomplished with human assistance. The robot uses its vocabulary to communicate its needs to a caregiver. Successful interactions enable the robot to ground its vocabulary in sensory data and in the goal that can be achieved with the acquired word. The robot can also explore its environment and perform gestures to manipulate and point at objects. The caregiver provides labels for these gestures, which the robot uses to learn the names of its actions.

IV. Method

The proposed architecture is based on the robot language model proposed by [29] dedicated to learning the name of objects and the associations between internal needs and words in a robot. The extension consists of giving the robot the ability to manipulate and point to objects, and to hear the caregiver's verbal labeling of actions, allowing it to learn the association between the actions and their labels. The present architecture consists of four subsystems. A dynamic model of motivations [32] [33] [34] enabling the robot to have goals. A visual perception module that helps the robot to perceive its environment and learn the names of objects. An action/communication module enabling the robot to express its needs and manipulate or point to objects. An auditory module for speech perception and word-action association. The robot's overall architecture is illustrated in Figure 2. The following sections describe in detail each of the modules and how associations between words/actions, motivations and objects are created via human-robot interaction.

A. The robot cognitive model

We first detail the architecture's motivation, perception, and actions/communication modules, and then outline how

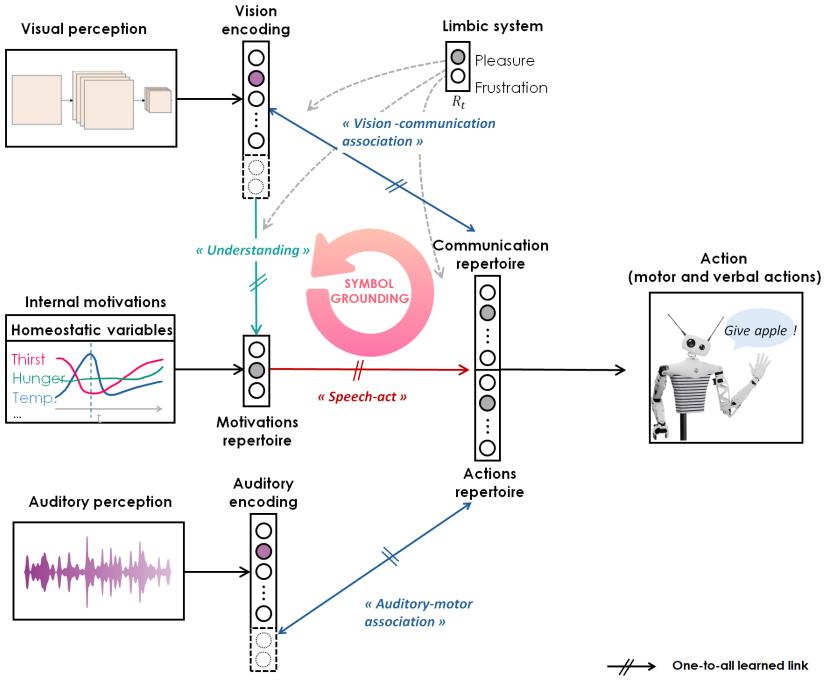


Fig. 2. The robot's architecture consists of a perception module, a motivation module, and a communication/action module. Like a child, the robot learns through interaction with its environment and a human partner to name objects (nouns), identify their affordances (i.e., the motivations they can satisfy), and learn the names of actions (verbs). The robot grounds nouns in the visual appearance of objects and verbs in the motor joint angles.

associative learning enables their integration.

1) Motivation module

We propose to endow the robot with a motivation module (Fig. 2) that modulates its internal drives and sets its goals. These motivations make the robot an active learner and drive it to act or communicate in order to satisfy them. Motivations can include the need to "eat", "drink", or "play", and they can be satisfied by objects present in the robot's environment. When a motivation is triggered, the robot may choose an action or utter a word that prompts a caregiver to assist it. Learning occurs when the motivation is satisfied, and after receiving a reward, the robot learns the optimal action or word to use in similar future contexts.

We used a simplified version of the motivation module from [27], since we are not studying in this work the impact of temporal evolution of motivation on language acquisition. The robot's motivation to satisfy is randomly selected from the motivation space according to a uniform distribution.

2) Perception module

The robot is endowed with a perception system that enhances its ability to detect its environment, including the presence of objects, the caregiver, and verbal feedback. This system comprises two modalities: visual and auditory.

Visual perception

The vision module performs object detection and feature extraction, which are used as input for two neural networks (perceptrons) to create two types of associative learning: between words and referents, and between objects and internal states. The robot's stereovision allows it to estimate the distance to objects and learn that certain actions—such as grasping or reaching—are ineffective if the objects are too far. Depth information is estimated based on the horizontal shift (disparity) of the same object in two images captured at the same time by the robot's cameras [35].

Auditory perception

The auditory system allows the robot to perceive verbal feedback provided by the caregiver. In our work, we consider the robot to be comparable to a 24-month-old baby, so it can only pronounce a limited number of syllables (see next section). When the robot pronounces a proto-word and obtains the correct object that responds to its current motivation, the verbal feedback from the experimenter (the name of the object) is used by the robot to learn the correct word for the object and expand its vocabulary. The caregiver's verbal feedback is also used by the robot to learn the name of its actions. For example, when the robot hands an apple to the caregiver, the latter provides verbal feedback by saying "you give the apple," which allows the robot to learn the name of its action. The verbal labels phrases are converted into text and encoded (one-hot encoding) to serve as input to a recurrent neural network, which associates it with the robot's motor joint angles involved in executing the action.

3) Actions/Communication module

This module is dedicated to the action production of the robot, either verbal or gestural. The robot's actions are preprogrammed and can be executed to manipulate objects (like grasping or taking) or to draw the caregiver's attention to an object (like pointing) in order to establish joint attention. The robot's action space is defined as: $a \in \mathcal{A} = \{\text{pointing, grasping, giving, taking}\}$. These actions correspond to the early gestural production that infants in their

first and early second years use to communicate [31]. Each of this robot's action corresponds to a sequence of the robot's arm motor angles (trajectory) that forms a gesture directed toward an object.

The robot vocabulary is composed of two-syllable words that correspond to 10 of the most frequent syllables of an 8-month-old infant [36]. The communication module (Fig.2) also contains a text-to-speech unit that allows the robot to vocalize its words.

The protolanguage and gestures development are not addressed in this work, we consider that the robot has already developed phonological and motor skills and that the aim is to ground the developed words/gestures to the sensory data, action verbs, and to the goal that can be achieved with these words. Several developmentally inspired models have studied the emergence of proto-words and motor skills in robots through human interaction [37] [38] [39]. Our model is designed to explore the next developmental milestone once the infant has mastered those necessary prerequisites.

B. Learning the associations between modules

In our model, associations are learned online between perceptual, motor and motivation modules through the interaction with the environment and the caregiver. The robot's goal is to obtain the correct object that can meet its current need. This will lead the robot to learn the correct word to use for each context, to learn to name the objects in its visual field, the verb that describes its action, and to understand the internal needs that each object can satisfy, i.e. their affordances.

1) Association visual perception-motivation module

When the robot chooses a word/action and gets the correct object that satisfy the current motivation, the extracted visual features of the object by the visual perception module are used to train a neural network (perceptron) to associate object/internal need.

The update of the synaptic weights of this neural network follows the least mean squares rule:

$$\Delta\omega_{ij} = \epsilon V_i(y_j - \hat{y}_j) \tag{1}$$

with : • ϵ: the learning rate. • V: visual features of the object. • y_j: The internal state satisfied by the object. • $\hat{y}_j$: The predicted object affordance.

2) Association visual perception-word

A second neural network creates an association between the visual features of an object and the pronounced word that allowed the robot to obtain it. This association enables the robot to associate names from its vocabulary with the object it has interacted with. The synaptic weights update of this second neural network follows the same rule as the first one, namely the least mean squares method.

3) Association motivation-action/communication module

The mapping between the robot's vocabulary/actions and the internal state (Fig.2) uses the reinforcement learning approach proposed in [29]. The robot starts by producing a random word/action when one need outweighs the others, and the caregiver/human partner who has no prior knowledge of the robot's internal need responds to the robot's vocalization by choosing an object and giving it to the robot. If the given object meets the robot's need, the motivation associated with this need decreases, and a reward of +1 is given to the robot, which expresses its satisfaction with a happy gesture (Fig. 4b). Otherwise, the word receives a reward of -1, which decreases the probability of reusing the word in the same situation, and the robot expresses its dissatisfaction (Fig.4c).

In reinforcement learning, this problem can be formulated as a contextual multi-arm bandit problem [40], in which the action space corresponds to the words/actions the robot can utter and perform, states are the robot internal needs, the contexts correspond to the presence or absence of the caregiver, the presence of objects, and their proximity to the robot. When the robot encounters a new situation, a new context is added to the Q-table and initialized with the values of the most similar situation. In each context, the Q value of action a is calculated using equation :

$$Q_{n+1}(a) = \frac{h-1}{h} Q_n(a) + R_n \tag{2}$$

with h, a parameter used to prevent divergence of the Q value, and R_n the reward received at time step n. The robot uses a unified greedy policy (Eq. 3) to select a word/action according to context.

$$A_n = \arg\max_a Q_n(a) \tag{3}$$

When the interaction is successful, the caregiver also provides the correct name of the object, allowing the robot to add the new word to its Q-table and assign it a positive reward. The learning method is detailed in the Algorithm 1.

Algorithm 1 Learning algorithm

Require: N, h
  Initialize action value function $Q(s, a)$
  for $n = 1, 2, \ldots, N$ **do**
    wait for an internal state s to trigger
    while The internal state is not satisfied **do**
      Choose action/word a from s using policy derived from Q
      Take action a, observe $R(s, a)$
      $r \leftarrow R(s, a)$
      Update
      $Q(s, a) \leftarrow \frac{h-1}{h} Q(s, a) + r$
    end while
    Create the association between the visual features of the object and the internal state
    If the object is obtained by requesting it from the caregiver: create the second associations between the object visual features and the pronounced word. Add the new word to the Q_table.
    If the object is obtained by a robot action: create the second association between the performed action and the label provided by the caregiver.
  end for

4) Association auditory perception-actions

When the robot manipulates objects and performs actions on them, such as grasping, taking, giving, pointing... The human partner provides the label for this action. This verbal label is converted into a sequence of words, then associated with the sequence of the motor primitive that resulted in the action. We use a recurrent neural network that associates a word sequence with action sequence to learn the verbs that describe the robot's actions. Specifically, we use sequence to sequence LSTM (Long Short Term Memory) [41] for this learning task, since the labels provided by the caregiver are a sequence of words (used as the inputs of the LSTM), and the robot's actions (the outputs) are a sequence of motor joint angles that change over time during the action trajectory.

V. Experimental setup

We employed a simulated version of the humanoid robot Reachy running under the Unity simulation environment (Fig.3) and we endowed the robot with the proposed architecture. The robot is internally driven by three needs: hunger, thirst and boredom. These needs can be satisfied by objects in the scene. When a need is triggered, the robot either chooses an action or utters a word to obtain an object. A human caregiver gives the robot an object when the robot speaks or point at an object and provides labels for the actions performed by the robot. The robot's vocabulary is composed of 10 words, 6 objects are present in the robot's environment (Fig. 3), two of which can satisfy the need to play (a car and a bear), two can satisfy the need to eat (an apple and an orange) and two can satisfy the need to drink (a cup, a milk bottle). The moving average of rewards is used as an evaluation metric. It is defined as the average, at iteration n, of the m rewards previously received:

$$\bar{r}_n = \begin{cases} \frac{1}{n}\sum_{i=1}^{n} R_i & \text{if } n < m \\ \\ \frac{1}{m}\sum_{i=n-m+1}^{n} R_i & \text{Otherwise} \end{cases} \quad (4)$$

Convergence time is defined as the number of iterations required to reach 90% of the final value.

To study how motor development affects language acquisition, we conduct a control experiment in which the robot's outputs are limited to speech, and a second experiment where the robot can both speak and perform actions.

Virtual caregiver

To be able to repeat the experiment many times and with a high number of iterations, we used a virtual caregiver that simulates different caregiving strategies. This caregiver does not know the internal states of the robot, he can only hear/see the actions of the robot, choose and give an object, label the robot actions, and observe the robot feedback. This problem can be formulated as a multi-armed bandit problem [40], the states correspond to the actions heard/seen by the virtual caregiver and the action space corresponds to the choices of objects she/he will give, an action receives a reward if the robot expresses its happiness (using its antennae) after obtaining the desired object, otherwise the action is penalized.

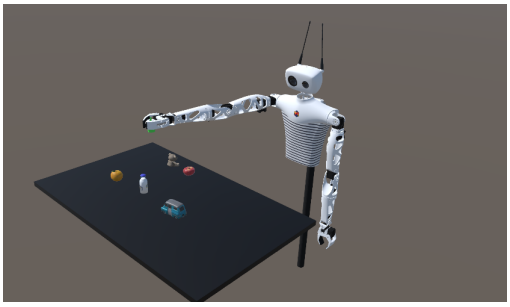


Fig. 3. Experimental setup in the Unity simulator. Reachy is standing in front of a table with objects that can satisfy its internal needs - boredom, hunger and thirst. In the control experiment, for a given internal state, Reachy says a word from its repertoire to obtain the appropriate object. The caregiver tries to guess the desired object and clicks on it. The robot then expresses satisfaction or frustration with its antennae when the given object satisfies the current need. In the second experimental scenario, the robot can discover its environment by manipulating objects and performing actions on them (e.g. giving an object as shown in the figure). The caregiver labels the actions which enable learning of the verbs.

(a) Robot neutral gesture

(b) Robot happy gesture

(c) Robot sad gesture

Fig. 4. The robot's feedback is displayed after it utters a word and receives an object from the caregiver. This allows the caregiver to know whether the object given is the one desired by the robot. As a result, the caregiver will give the same object in the future when he/she hears the same word spoken by the robot.

The choice of the object to be given follows a greedy approach (the action with the highest Q-value). During the training phase, caregiver verbal feedback is provided following the robot's actions. For example, after the robot's action (e.g., robot: [grabs the cup]), the caregiver responds with verbal description (e.g., caregiver: "You take the cup!"). In the test phase, the caregiver issues verbal commands to prompt the robot's actions (e.g., caregiver: "Take the cup!"), and the robot responds accordingly (e.g., robot: [grabs the cup]).

VI. Results

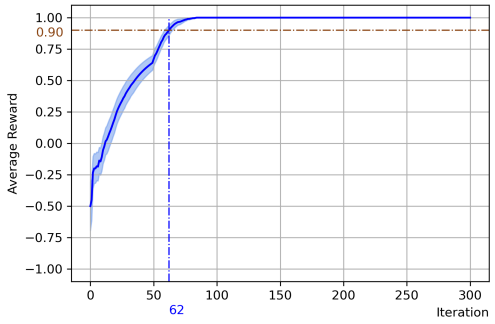
The results were calculated on the average of 20 repetitions of each experiment. The moving average reward at each time step n is computed on the previous $m = 50$ values. The results of each experimental scenario tested are presented below.

Figure 5a illustrates the evolution of the average reward in the control scenario where the robot relies only on words to satisfy its needs. Convergence is reached after 62 iterations.

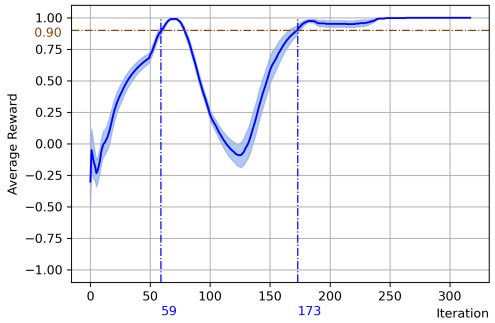
Table I shows the new learned words from the caregiver feedback and the association between the robot's vocabulary and the internal needs. After learning, each need has only one word with a convergent Q-value.

In the second experimental scenario, the robot relies on words and actions to satisfy its needs. Figure 5b shows the

evolution of the average reward in this scenario. The first convergence is reached after 59 iterations. At iteration 70 the caregiver stops responding to the robot's babble, causing the robot to learn to rely on its actions to satisfy its need, which leads to the second convergence in iteration 173 after the robot learned to take objects.



(a) The control scenario, the robot relies only words to communicate its needs to the caregiver. Convergence is reached on iteration 62.



(b) Scenario where the robot relies on words and actions to satisfy its needs, but the caregiver presence in not permanent. The first convergence is reached on iteration 59. At iteration 70, the caregiver left the robot, and the robot began to explore new strategies that could help it obtain the desired object. The robot learns to rely on its action "Take" to obtain autonomously the object that can satisfy its need, hence the second convergence observed in the iteration 173.

Fig. 5. Evolution of the moving average reward in the three tested experimental scenarios.

Table II illustrates the association between the robot's vocabulary and actions with different contexts. Each of these contexts is the merging of the robot's internal state and the perceptual context, which provides information about the environment, such as the presence or absence of the caregiver, nearby objects, etc. For this scenario, the first three of the six contexts created correspond to the robot's internal states ("Thirst", "Hunger" and "Boredom") plus the presence of the caregiver, while the last three correspond to the same internal states plus the absence of the caregiver. After learning, each context has a single output with a convergent Q-value.

Histogram 7 illustrates the number of words acquired by the robot in both conditions: with and without motor actions. These words were learned during interaction and are grounded either in the robot's sensory inputs or its motor joint positions.

After the learning phase, the LSTM recurrent neural network successfully generated the motor joint angles for the

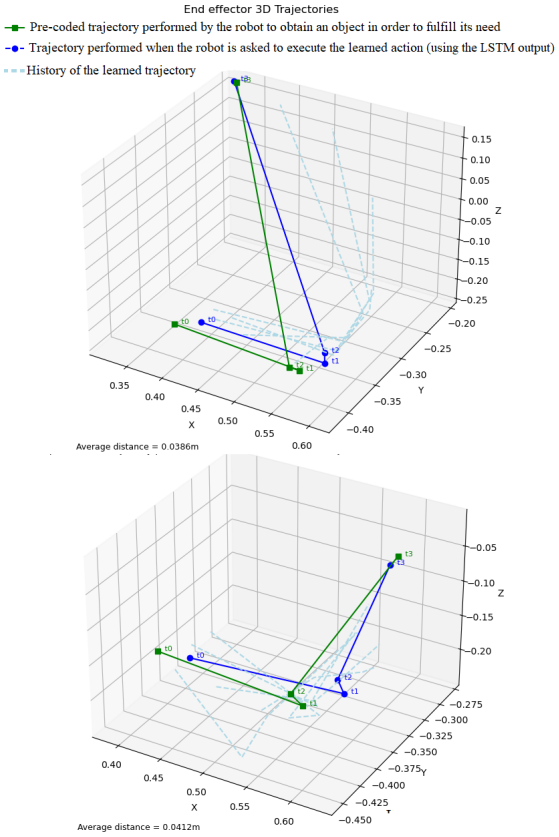


Fig. 6. Comparison between the trajectory of the robot arm's end effector when performing an action aimed at obtaining an object to meet a need, and the learned trajectory executed by the robot when asked to perform the action (Euclidean distance). Top, the action "give" an object. Bottom, the action "take" an object.

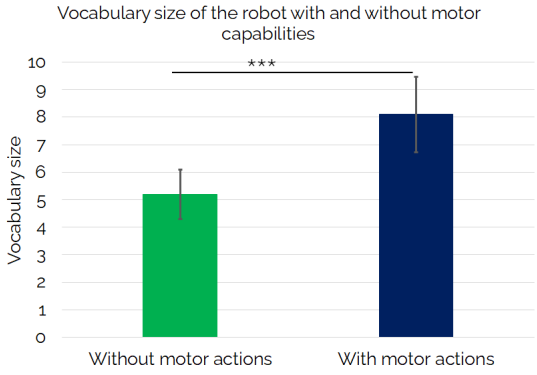


Fig. 7. Comparison of the robot's acquired vocabulary size with and without motor capabilities.
*** very significant statistical difference from t-test: p-value<0.001.

trajectories corresponding to phrases describing the robot's manipulation of objects (Figure 6).

Figure 8 shows the PCA-based dimensionality reduction of trajectories generated by the robot's recurrent neural network in response to single-word phrases from the two lexical categories of verbs and nouns. The PCA trajectories of the nouns reflect their position in the experimental setup (Fig.3).

	”wada”	”naba”	”maba”	”daba”	”paba”	”bada”	”bama”	”babe”	”waba”	”wama”	”cup”	”car”	”apple”
”Thirst”	0.2	0.4	0.1	0.2	0.02	-0.2	0.3	0.3	0.02	0.3	4	0	0
”Boredom”	0.4	0.8	0.2	0.6	0.6	0.1	0.5	0.7	0.1	0.6	0.6	4	0
”Hunger”	0.02	0.1	0.1	0.4	0.5	0.4	0.1	0.8	0.5	0.3	0.0	0.4	4

TABLE I

Q-TABLE OF ASSOCIATION BETWEEN THE ROBOT'S VOCABULARY AND INTERNAL STATES IN THE CASE WHERE THE ROBOT RELIES ONLY ON ITS VERBAL OUTPUT TO COMMUNICATE ITS NEEDS. IN BLUE, THE NEW WORDS ACQUIRED FROM THE CAREGIVER FEEDBACK.

	”wada”	”naba”	”maba”	”daba”	”paba”	”bada”	”bama”	”babe”	”waba”	”wama”	”take”	”give”	”hold”	”point”	”cup”	”car”	”apple”
Context 1	0.2	0.6	0.6	0.4	0.0	0.5	0.5	0.4	0.1	0.4	0.1	0.6	0.6	0.8	4	0	0
Context 2	0.6	0.2	-0.4	0.5	0.1	0.0	0.3	0.1	-0.3	-0.3	-0.4	0.3	0.6	0.2	0.7	4	0
Context 3	0.8	0.2	-0.3	-0.3	0.5	0.1	0.4	-0.4	0.2	0.0	0.7	0.6	0.0	-0.3	0.1	0.1	4
Context 4	-0.6	-0.6	-0.4	-0.3	0.1	-0.7	0.2	-0.7	0.1	0.1	4	-0.3	-0.3	-0.4	0.8	0.6	0.7
Context 5	-1.7	-1.3	-1.7	-1.7	-1.7	-1.7	-1.4	-1.4	-1.5	-1.3	4	-1.7	-1.7	-1.5	0.4	0.4	0.0
Context 6	-0.7	-0.8	-0.4	-0.8	-0.4	-1.2	-0.6	-0.7	-0.9	-0.8	4	-0.8	-0.5	-0.6	-0.4	-0.6	-0.7

TABLE II

Q-TABLE OF THE ASSOCIATIONS BETWEEN THE ROBOT'S VOCABULARY AND ACTIONS WITH ITS INTERNAL STATES AND CONTEXTS, IN SCENARIO WHERE IT RELIES ON VERBAL AND BEHAVIORAL OUTPUTS. THE CAREGIVER IS NOT ALWAYS PRESENT, CAUSING THE ROBOT TO LEARN TO TAKE OBJECTS TO SATISFY ITS NEEDS. THE FIRST THREE OF THE SIX CONTEXTS CREATED CORRESPOND TO THE ROBOT'S INTERNAL STATES (“THIRST”, “HUNGER” AND “BOREDOM”) PLUS THE PRESENCE OF THE CAREGIVER, WHILE THE LAST THREE CORRESPOND TO THE SAME INTERNAL STATES PLUS THE ABSENCE OF THE CAREGIVER. IN YELLOW, THE ACTIONS THE ROBOT CAN PERFORM. IN BLUE, THE NEW ACQUIRED WORDS.

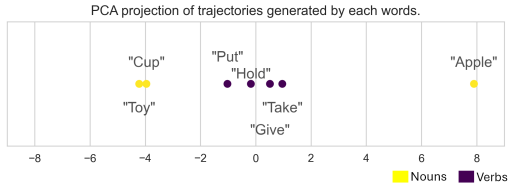


Fig. 8. Projection of the trajectories produced by the robot's recurrent neural network following input from two lexical categories: nouns and verbs.

VII. DISCUSSION

In the control scenario, the robot starts to receive positive rewards after the learning phase, which means that the caregiver starts to understand the words used by the robot and gives it the correct object, with each object having a stable name. The drawback of the first architecture is that the robot depend always on the caregiver to give it an object and is not expected to reach complete autonomy. In the second architecture, where the robot is endowed with motor capabilities, we observe convergence at approximately the same speed as in the first, since the robot also relies on its verbal output and on the caregiver to satisfy its needs. At iteration 70, the caregiver leaves the robot, and the robot began to explore new strategies that could help it obtain the desired object. After 103 iterations, the robot learns to rely on its action “Take” to obtain the object that can satisfy its need, hence the second convergence observed after the iteration 173.

The difference in vocabulary size visible in histogram 7 can be explained by the labels the robot obtains when manipulating objects. Each time the robot holds, points at or takes an object, it establishes a joint attention between the two parties on the given object and facilitates its naming. This result supports prior research suggesting that early gestural production in children can predict their vocabulary development [31]. The size difference in the vocabulary is also due to the nature of the words acquired by the robot, when endowed with motor capabilities, the robot also learns the names of the actions it can perform. Unlike noun learning, verb learning in our architecture was only possible when the robot had motor skills. This aligns with findings in infants which show that motor actions provide a powerful basis for the learning of verbs [11]. The dimensionality reduction of the trajectories generated by the robot after hearing verbs and nouns shows the appearance of a clustering of the two lexical categories (Fig. 8). During the learning phase, the robot heard verb-noun phrases labeling its action performed on an object (e.g.: hold apple, give toy...). In contrast, the trajectory generated here corresponds solely to a phrase of one word, which can be a verb or a noun, as shown in figure 8. In the inference phase, when the robot hears nouns, it directs its action towards the corresponding object. As for verbs, since they always need a noun (object) as an argument, giving the robot a verb-only sentence will generate a neutral trajectory equivalent to the mean of all learned trajectories, as shown in figure 8. As a result, and for this learned vocabulary, the architecture is able to differentiate between nouns and verbs in the trajectory space. In humans, the ability to discriminate between word classes is considered to be the first step in learning syntactic constraints and how to combine words in order to generate new sequences and phrases [42].

VIII. CONCLUSION

We presented an active learning robot architecture to study how motor development affects language acquisition. As in humans, the results show that motor skills can be seen as a catalyst for verbs learning. With motor skills, the first difference being the size of the vocabulary and the grammatical category that the robot develops, the robot has also acquired the semantic and pragmatic aspects of language, each learned word is associated either with perceptual information (nouns) or with motor joints (verbs). The robot's actions enabled it to learn not only verb words, but also nouns. When the robot pointed at the object or manipulated it, it received the name of the object from the caregiver. The results also support the findings suggesting that verb learning is embodied and requires motor

experiences and cues to form the meaning of the verb, unlike noun learning which is supported only by observation, which may explain the early noun advantage. The results also show the beginning of the categorisation of the two fundamental grammatical forms: nouns and verbs. Regarding the limitations that can be addressed in future work, our architecture is not yet fully developmental, the robot uses a pre-coded repertoire of words and actions, we assume that the robot masters some specific actions and the goal here is to learn to name them and study how they affect language, this can be modified in the future using for example reinforcement learning algorithms dedicated to actions acquisition. We will further improve the experimental design by including several participants in the study. We also intend to expand on the motivation system which we expect to help with the acquisition of high-level language and the learning of sequence of words.

ACKNOWLEDGMENT

This research is funded by a EUTOPIA PhD Co-tutelle grant.

REFERENCES

[1] K. Howard, G. Roberts, J. Lim, K. J. Lee, N. Barre, K. Treyvaud, J. Cheong, R. W. Hunt, T. E. Inder, L. W. Doyle, *et al.*, "Biological and environmental factors as predictors of language skills in very preterm children at 5 years of age," *Journal of Developmental & Behavioral Pediatrics*, vol. 32, no. 3, pp. 239–249, 2011.

[2] C. Suarez-Rivera, E. Linn, and C. S. Tamis-LeMonda, "From play to language: Infants' actions on objects cascade to word learning," *Language Learning*, vol. 72, no. 4, pp. 1092–1127, 2022.

[3] E. A. Walle and J. J. Campos, "Infant language development is related to the acquisition of walking.," *dev. psychology*, vol. 50, p. 336, 14.

[4] J. M. Iverson, "Developing language in a developing body: The relationship between motor development and language development," *Journal of child language*, vol. 37, no. 2, pp. 229–261, 2010.

[5] E. Orr, "Object play as a mediator of the role of exploration in communication skills development," *Infant Behavior and Development*, vol. 60, p. 101467, 2020.

[6] I. Brooker and D. Poulin-Dubois, "Is a bird an apple? the effect of speaker labeling accuracy on infants' word learning, imitation, and helping behaviors," *Infancy*, vol. 18, pp. E46–E68, 2013.

[7] K. L. West and J. M. Iverson, "Language learning is hands-on: Exploring links between infants' object manipulation and verbal input," *Cognitive development*, vol. 43, pp. 190–200, 2017.

[8] S. Waxman, X. Fu, S. Arunachalam, E. Leddon, K. Geraghty, and H.-j. Song, "Are nouns learned before verbs? infants provide insight into a long-standing debate," *Child dev. perspectives*, vol. 7, p. 155, 2013.

[9] D. Gentner, "Why verbs are hard to learn," *Action meets word: How children learn verbs*, pp. 544–564, 2006.

[10] T. B. Piccin and S. R. Waxman, "Why nouns trump verbs in word learning: New evidence from children and adults in the human simulation paradigm," *Language Learning and devel.*, vol. 3, p. 295, 2007.

[11] K. L. West, K. K. Fletcher, K. E. Adolph, and C. S. Tamis-LeMonda, "Mothers talk about infants' actions: How verbs correspond to infants' real-time behavior.," *devel. psychology*, vol. 58, no. 3, p. 405, 2022.

[12] K. L. Frewin, S. A. Gerson, R. E. Vanderwert, and C. Gambi, "Parent-reported relations between vocabulary and motor development in infancy: Differences between verbs and nouns," *Infancy*, vol. 30, p. e12638, 2025.

[13] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection science*, vol. 15, no. 4, p. 151, 2003.

[14] J. Weng, "Developmental robotics: Theory and experiments," *International Journal of Humanoid Robotics*, vol. 1, no. 02, pp. 199–236, 2004.

[15] L. She and J. Chai, "Interactive learning of grounded verb semantics towards human-robot communication," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1634–1644, 2017.

[16] J. Y. Chai, R. Fang, C. Liu, and L. She, "Collaborative language grounding toward situated human-robot dialogue," *AI Magazine*, vol. 37, no. 4, p. 32, 2016.

[17] J. M. Siskind, "Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic," *Journal of artificial intelligence research*, vol. 15, pp. 31–90, 2001.

[18] R. Cantrell, P. Schermerhorn, and M. Scheutz, "Learning actions from human-robot dialogues," in *2011 RO-MAN*, pp. 125–130, IEEE, 2011.

[19] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Grounding verbs of motion in natural language commands to robots," in *Experimental robotics: 12th Int. symposium on experimental robotics*, p. 31, Springer, 2014.

[20] K. Sugiura and N. Iwahashi, "Learning object-manipulation verbs for human-robot communication," in *Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*, pp. 32–38, 2007.

[21] T. Nakamura, T. Nagai, and N. Iwahashi, "Grounding of word meanings in multimodal concepts using lda," in *2009 IEEE/RSJ Int. Conference on Intelligent Robots and Systems*, p. 3943, IEEE, 2009.

[22] D. Marocco, A. Cangelosi, and S. Nolfi, "The emergence of communication in evolutionary robots," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 361, no. 1811, pp. 2397–2421, 2003.

[23] A. F. Morse, V. L. Benitez, T. Belpaeme, A. Cangelosi, and L. B. Smith, "Posture affects how robots and infants map words to objects," *PloS one*, vol. 10, no. 3, p. e0116012, 2015.

[24] A. F. Morse, J. De Greeff, T. Belpeame, and A. Cangelosi, "Epigenetic robotics architecture (era)," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 4, pp. 325–339, 2010.

[25] L. K. Samuelson, L. B. Smith, L. K. Perry, and J. P. Spencer, "Grounding word learning in space," *PloS one*, vol. 6, no. 12, p. e28095, 2011.

[26] M. A. K. Halliday, *Language of early childhood*. A&C Black, 2006.

[27] Z. Lemhaouri, L. Cohen, and L. Cañamero, "The role of the caregiver's responsiveness in affect-grounded language learning by a robot: Architecture and first experiments," in *2022 IEEE International Conference on Development and Learning (ICDL)*, pp. 349–354, IEEE, 2022.

[28] A. Markelius, S. Sjöberg, Z. Lemhauori, L. Cohen, M. Bergström, R. Lowe, and L. Cañamero, "A human-robot mutual learning system with affect-grounded language acquisition and differential outcomes training," in *Int. Conference on Social Robotics*, p. 108, Springer, 2023.

[29] L. Cohen and A. Billard, "Social babbling: The emergence of symbolic gestures and words," *Neural Networks*, vol. 106, pp. 194–204, 2018.

[30] E. Heikkinen, E. Silvennoinen, I. Khan, Z. Lemhaouri, L. Cohen, L. Cañamero, and R. Lowe, "Human-robot mutual learning through affective-linguistic interaction and differential outcomes training [preprint]," *arXiv:2407.01280*, 2024.

[31] J. Olson and E. F. Masur, "Infants' gestures influence mothers' provision of object, action and internal state labels," *Journal of Child Language*, vol. 38, no. 5, pp. 1028–1054, 2011.

[32] D. Cañamero, "Modeling motivations and emotions as a basis for intelligent behavior," in *Proceedings of the first international conference on Autonomous agents*, pp. 148–155, 1997.

[33] M. Lewis and L. Cañamero, "Hedonic quality or reward? a study of basic pleasure in homeostasis and decision making of a motivated autonomous robot," *Adaptive Behavior*, vol. 24, no. 5, pp. 267–291, 2016.

[34] I. Cos, L. Cañamero, and G. M. Hayes, "Learning affordances of consummatory behaviors: Motivation-driven adaptive perception," *Adaptive Behavior*, vol. 18, no. 3-4, pp. 285–314, 2010.

[35] R. A. Hamzah and H. Ibrahim, "Literature survey on stereo vision disparity map algorithms," *Journal of Sensors*, vol. 2016, no. 1, 2016.

[36] A. G. Levitt and J. G. A. Utman, "From babbling towards the sound systems of english and french: A longitudinal two-case study," *Journal of child language*, vol. 19, no. 1, pp. 19–49, 1992.

[37] D. Caligiore and G. Baldassarre, "Development of reaching and grasping: towards an integrated framework based on a critical review of computational and robotic models," *Reach-to-Grasp Behav.*, p. 319, 18.

[38] I. S. Howard and P. Messum, "Modeling the development of pronunciation in infant speech acquisition," *Motor Control*, vol. 15, p. 85, 2011.

[39] J. Saunders, C. Lyon, F. Forster, C. L. Nehaniv, and K. Dautenhahn, "A constructivist approach to robot language learning via simulated babbling and holophrase extraction," in *2009 IEEE Symposium on Artificial Life*, pp. 13–20, IEEE, 2009.

[40] R. S. Sutton, A. G. Barto, *et al.*, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.

[41] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Adv. in neural info. processing sys*, vol. 27, 2014.

[42] M. D. Braine, "The ontogeny of english phrase structure: The first phase," *Language*, pp. 1–13, 1963.