A Developmental Robot Model of Early Language Acquisition

Zakaria Lemhaouri, Laura Cohen, Ann Nowé, Lola Cañamero

A Developmental Robot Model of Early Language Acquisition

Zakaria Lemhaouri[1,2], Laura Cohen[1], Ann Nowé[2], and Lola Cañamero[1]

[1]ETIS Lab, CY Cergy Paris University - ENSEA - CNRS UMR8051, France
[2]Artificial Intelligence Lab, Vrije Universiteit Brussel (VUB), Belgium

***Abstract*—Recent NLP techniques have enabled a considerable advance in the generation and understanding of natural language. But given the way these neural NLP systems learn, and the astronomical amounts of data required, they cannot provide answers about how human infants learn and acquire language, as they do not follow the same language development trajectory. We propose a robot cognitive model of early human language acquisition inspired by the way human babies learn language. The robot relies on social interaction, making it an active learner, with a caregiver to acquire motivation-grounded language. The robot's modular architecture enables it to be situated in the same conditions as a human child acquiring language. The aim of this model is to provide a tool for testing hypotheses related to questions about the process of language development in humans.**

I. Introduction

Affect and motivation are central in the development of the necessary capacities for language [1]. Usage and functional-based theorists argue that communication emerges due to its use as a means by infants to convey functional meanings, even before they have mastered adult language [2]. For instance, communication can be a means of obtaining a desired object by asking an adult for it, or to reinforce a social bond. To give the robot the ability to learn language in this functional way, we endow it with a modular architecture capable of learning multiple associations based on motivations.

II. Proposed Approach and Method

The overall architecture is shown on Fig.1. The formalism is related to the sensory-motor PerAc neural architecture [3] and consists of three modules: the motivation, visual perception and phonological modules. The **Motivation module** (fig.1.B) modulates the robot's internal motivation as a function of time and visual perception (fig.3a) [4]. The robot uses a winner-take-all strategy to decide which motivation to prioritize. Each internal need is modeled by a homeostatic variable that decreases over time and increases when the need is fulfilled. The robot drive $d_i(t)$ is defined as the deviation between the current homeostatic variable and its optimal value. The robot's motivation to satisfy a need depends on the related drive (internal factor) and the intensity of the stimulus (external factor) that can satisfy it [5]:

$$m_i(t) = d_i(t) + d_i(t).s_i(t) \quad (1)$$

This research is funded by a EUTOPIA PhD Co-tutelle grant.
This abstract is submitted to WACAI 2024 workshop (Affects, Compagnons Artificiels & Interaction) in the category short descriptions of realizations, demonstrations and experiments in progress.

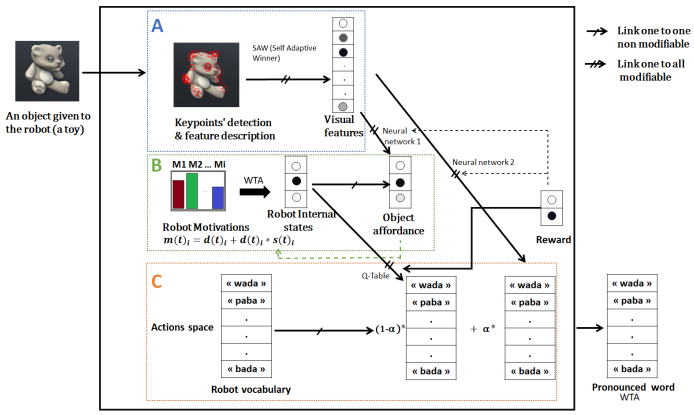


Fig. 1: The overall architecture consists of three modules: The visual perception module A, the motivation module B and the phonological module C.

Figure 3a shows an example of the evolution of one of the robot's motivations: the motivation to eat. At time t_1, the caregiver gave the robot an edible object, which decreased the robot's motivation to "eat" ($d(t)$ goes to zero). At time t_2, an edible object was presented in the robot's environment ($s(t)$ becomes different to zero), thus increasing the motivation to eat. The stimulus s is estimated by the visual perception module. We used the estimated activation of each class, given by the first neural network (fig.1.B), as the intensity of the stimulus corresponding to each object. The **visual perception module** (fig.1.A) enables the robot to perceive its environment and learn the name and affordance of each object. The robot detects the key points of perceived objects using FAST algorithm [6] and clusters the key points of each object using an agglomerative hierarchical clustering algorithm [7]. Key point descriptors are calculated with the SIFT algorithm [8] and stored in a visual features matrix V using an online incremental learning method based on Kohonen map [9]. Each new descriptor is compared to those already stored in V, if the similarity is below a fixed threshold, the most similar descriptor is replaced by the mean of the two, otherwise the new descriptor is recruited directly to V. The V matrix is used as input to two neural networks to create associative learning (fig.1.A). The **Phonological module** (fig.1.C) of the robot is composed of a vocabulary of two-syllable words corresponding to 10 of the most frequent syllables of an 8-month-old infant [10]. The phonological module also contains a text-to-speech unit that allows the robot to vocalize its words.

Learning the associations between modules

The robot learns the associations between each pair of these modules. The goal is for the robot to be able to say a word

when it is in a given internal state, to learn to name the objects in its visual field, and to know which internal need each object is capable of satisfying. The **association visual perception-motivation** is achieved by training a neural network - which has the V matrix as input - to predict the name of the detected object and which internal need can be satisfied by it. The synaptic weights of this neural network are updated according to the Widrow-Hoff rule [11]:

$$\Delta\omega_{ij} = \epsilon V_i(y_j - \hat{y}_j) \tag{2}$$

with : ϵ: the learning rate, y_j: The internal state satisfied by the object, and $\hat{y}_j$: The predicted object affordance.
The same update rule is used to create the **association between visual perception and phonological modules**. To achieve the last **association motivation-phonological modules**, we extend the RL framework proposed by [12]: in this approach, each of the robot's internal needs can be satisfied by a specific object. The robot begins by randomly producing a word when one need outweighs the others. The caregiver - who doesn't have knowledge of the robot's internal need - reacts to the robot's vocalization by choosing an object and handing it to the robot. If the given object satisfies the robot's need, the motivation related to this need decreases, a reward of +1 is given to the robot, which expresses its satisfaction with a happy gesture. Otherwise, the word receives a reward of -1, which decreases the probability of reusing the same word in this context, and the robot expresses its dissatisfaction with a sad gesture. In RL, this problem can be formulated as a contextual multi-armed bandit problem. In each state, the value Q of action a (word) is calculated using the equation:

$$Q_{n+1}(a) = \frac{h-1}{h}Q_n(a) + R_n \tag{3}$$

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 greedy policy to select a word according to its internal needs. The **robot pronounced word** is a weighted sum between the Q-table and the neural network word prediction of the visual module, using a winner-takes-all strategy (fig.1.C).

III. Experimental setup and Results

To test our model, we used the humanoid robot Reachy with the Unity simulation environment (fig.2), the robot has three internal states: hunger, thirst and curiosity, which can be satisfied by the five objects present. When the robot express its need by a word (from its vocabulary of 10 words) a human caregiver gives it one of the objects. The robot can express its satisfaction or frustration by putting its antennae up or down. The average of the rewards is used as an evaluation metric (at each time step n, it is computed on the previous 50 values). The convergence time is defined as the number of iterations needed to reach 90% convergence. The results were calculated on the average of 5 repetitions the experiment.

The results show the convergence of the moving average reward (fig.3b). Convergence is reached after 76 iterations. Table I shows the association between the robot's vocabulary and the internal needs after the learning.

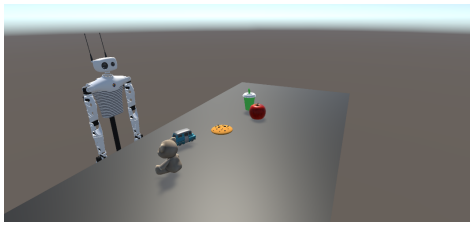
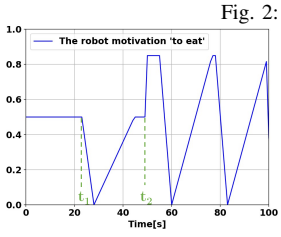
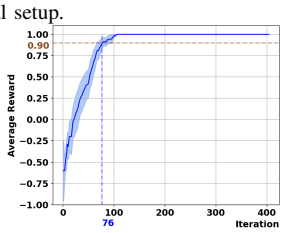


Fig. 2: Experimental setup.



(a) Example of the evolution of one of the robot's motivations.

Fig. 3



(b) Evolution of the moving average reward.

	"wada"	"naba"	"maba"	"daba"	"paba"	"bada"	"bama"	"babe"	"waba"	"wama"
"Thirst"	-0.13	-0.01	-0.59	-0.41	-0.40	4	-0.053	-0.41	-0.33	-0.493
"Hunger"	-0.39	-0.29	0.040	-0.34	4	0.47	-0.40	4	0.14	0.65
"Curiosity"	0.069	0.35	-0.39	0.42	0.59	0.72	4	-0.42	0.01	4

TABLE I: Q-table of association between the robot's vocabulary and internal states.

The visual module was tested by showing the robot a set of images of the learned objects, one per image. We evaluated whether the robot could associate a consistent name with each object and could predict the internal state satisfied by the object. The prediction accuracy was 100%.

IV. Discussion and Conclusion

The robot was able to associate words from its vocabulary with its internal states as demonstrated by the convergence of the moving average reward. Reaching convergence after 76 interactions (fig.3b) means that the robot has learned to choose consistent words that depend on its internal needs, and that the robot is understood by the caregiver, which allows it to obtain the desired objects. In the Q-Table the number of convergent words and their affordance correspond to the number of objects present in the chosen experimental setup and the motivation they can satisfy (fig 2). The high accuracy of the object recognition can be explained by the limited number of objects in the robot's environment and the optimal experimental conditions of the simulation.
The presented architecture enabled the robot to learn language in a functional way by learning the names of objects, their affordances, and the word to say to request an object to satisfy a need. These multiple associations resulted in a language grounded in the physical world and in the robot's internal needs, giving the acquired language a "meaning" instead of just non-grounded symbols.
Since we consider that language acquisition in this functional way is consistent with certain theories of human language learning, our ongoing experiments aim to test hypotheses related to the process of learning in humans, as in [13], and related to language development questions like how parental responsiveness can facilitate language learning in infants [14] and how the extra-linguistic context (such as sensory perception, environment, motivations and interactions) impacts language development.

REFERENCES

[1] S. G. Shanker and S. I. Greenspan, "The role of affect in language development," *THEORIA. An International Journal for Theory, History and Foundations of Science*, vol. 20, no. 3, pp. 329–343, 2005.

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

[3] P. Gaussier and S. Zrehen, "Perac: A neural architecture to control artificial animals," *Robotics and Auton. sys.*, 1995.

[4] 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.

[5] O. Avila-Garcıa and L. Cañamero, "Using hormonal feedback to modulate action selection in a competitive scenario," in *From Animals to Animats 8: Proceedings of the 8th Intl. Conf. on Simulation of Adaptive Behaviour, MIT Press, Cambridge, MA*, pp. 243–52, Citeseer, 2004.

[6] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*, pp. 430–443, Springer, 2006.

[7] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[8] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[9] T. Kohonen, *Self-organization and associative memory*, vol. 8. Springer Science & Business Media, 2012.

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

[11] B. Widrow and M. Hoff, "Ire wescon convention record," *IRE, New York*, pp. 96–104, 1960.

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

[13] 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 *International Conference on Social Robotics*, pp. 108–122, Springer, 2023.

[14] 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.