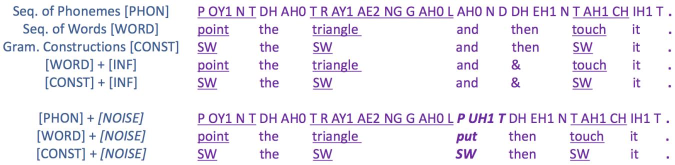
Neuro-Inspired Model for Robots Learning Language from Phonemes, Words or Grammatical Constructions

Xavier Hinaut
Inria Bordeaux Sud-Ouest, Talence, France.
LaBRI, UMR 5800, CNRS, Bordeaux INP, Université de Bordeaux, France.
Institut des Maladies Neurodégénératives, UMR 5293, CNRS, Université de Bordeaux, France.
xavier.hinaut@inria.fr

***Abstract*—There has been a considerable progress these last years in speech recognition systems. The word recognition error rate went down with the arrival of deep learning methods. However, if one uses cloud-based speech API and integrates it inside a robotic architecture, one still encounters considerable cases of wrong sentences recognition. Thus speech recognition can not be considered as solved especially when an utterance is considered in isolation of its context. Particular solutions, that can be adapted to different Human-Robot Interaction applications and contexts, have to be found. In this perspective, the way children learn language and how our brains process utterances may help us improve how robot process language. Getting inspiration from language acquisition theories and how the brain processes sentences we previously developed a neuro-inspired model of sentence processing. In this study, we investigate how this model can process different levels of abstractions as input: sequences of phonemes, sequences of words or grammatical constructions. We see that even if the model was only tested on grammatical constructions before, it has better performances with words and phonemes inputs.**



Seq. of Phonemes [PHON]	P OY1 N T	DH AH0	T R AY1 AE2 NG G AH0 L	AH0 N D	DH EH1 N	T AH1 CH	IH1 T	.
Seq. of Words [WORD]	point	the	triangle	and	then	touch	it	.
Gram. Constructions [CONST]	SW	the	SW	and	then	SW	it	.
[WORD] + [INF]	point	the	triangle	and	&	touch	it	.
[CONST] + [INF]	SW	the	SW	and	&	SW	it	.
[PHON] + *[NOISE]*	P OY1 N T	DH AH0	T R AY1 AE2 NG G AH0 L	***P UH1 T***	DH EH1 N	T AH1 CH	IH1 T	.
[WORD] + *[NOISE]*	point	the	triangle	***put***	then	touch	it	.
[CONST] + *[NOISE]*	SW	the	SW	***SW***	then	SW	it	.

Fig. 1. Symbol sequences given as input to the neural network depending on the conditions. The same sentence (see 2nd line, WORD condition) is given as input in order to see the effect of the different conditions. In PHON cond., a sequence of phonemes is inputed into the network using CMU's dictionary representation: e.g. *point* is replaced by the sequence of symbols "*P, OY1, N, T*". In CONST cond., semantic words are replaced by a SW symbol. In INF cond., infrequent words are replaced by "*&*" symbol: here, *then* is replaced by "*&*". In NOISE cond., 5% of the words are randomly replaced by another one: here, *and* is replaced by *put*.

I. Introduction

A. Robots as models to study language acquisition

Robots are interesting for studying language in many perspectives. Some of the long lasting questions are, for instance how languages evolve or emerge [11], [12], how language or symbols in general could be grounded [5], [10] or how the linguistic or non-linguistic symbols may emerge from grounding [15]. In particular, one may be interested to have a robot able to mix vision and dialog interaction in order to vocally command the robot to grasp some objects in complex environments [1], [6], [16]. However, even if some of these systems provide some transparency on how they work, they rarely help to understand how our brain processes languages or how children could acquire one.

Developmental architectures [4], [8] are inspired from children development and do not require to have all (vocabulary or syntactic) abilities prefixed since the beginning of the learning period. Some studies have used different cognitively inspired frameworks with robotics, such as Embodied Construction Grammar [3] and construction grammar [4], [9]. Our brains process utterances in a robust fashion in a variety of contexts: we believe that the lack of brain-inspiration in these systems results in a gap of robustness with human performance. In our approach, we try to build an architecture that is able to tackle several of these points and get a step closer to the understanding of brain processes, language developmental strategies and symbol grounding.

B. Our question and hypothesis

Considering a system that learns to parse a sentence given a stream of inputs, one question is **what is the optimal level of abstraction of the inputsphonemes, words or grammatical constructions?**

Here, we only compare purely symbolic input representation and do not consider raw acoustic signals or distributed representation of coding, such as word embedding. In particular we want also to see the **robustness to noise of these different representations**. This is particularly important when such a system is used with real speech signals, and have to deal with the misrecognition of words. We previously started a step in that direction by enabling the model to generalize on sentences with unknown (or unrecognized) words [7].

II. Discussion

Despite the small corpus used, the current performances are already interesting and useful for small corpus applications in Human-Robot Interaction experiments. Because the core part of the model is a generic neural architecture, it

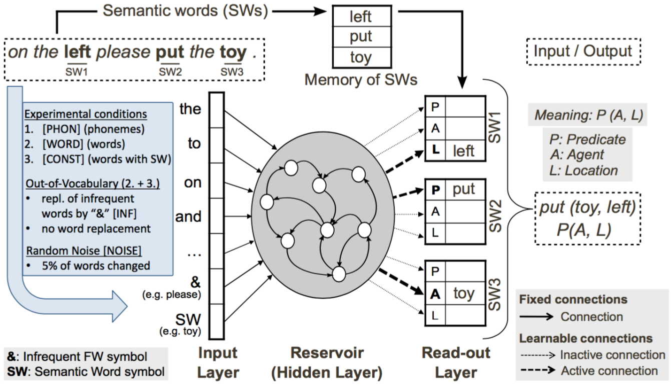


Fig. 2. **Sentence parsing model with different input conditions.** The system processes inputs as follows: (top left) from a sentence as input, the model outputs (middle right) an action command that can be performed by a robot. The processing of the sentence is sequential: each symbol of the sequence (phoneme, word, ...) is given one at a time as a one-hot encoding. The final thematic roles for each SW is read-out at the end of the sentence (but partial predictions can be read-out when the parsing is on-going). Before entering the neural network, the sentence is preprocessed depending on the main condition (PHON, WORD or CONST) and on the optional condition (INF and/or NOISE). Semantic Words (nouns, verbs, ...) are replaced by a SW symbol. Infrequent function words (IWs) are replaced by the *&* symbol. Here, the input layer only represents word symbols, but in the PHON condition these are replaced by phonemes. Example of input sequences for different conditions can be seen in Figure 1. Figure modified from [7].

TABLE I
SOME SENTENCE EXAMPLES FROM THE NOISY ENGLISH CORPUS. DIFFERENT TYPE OF SENTENCES ARE GIVEN: 1. SEQUENCE OF ACTIONS 2. IMPLICIT REFERENCE TO VERB 3. IMPLICIT REFERENCE TO VERB AND OBJECT 4. CROSSED REFERENCE 5. REPEATED ACTION 6. UNLIKELY ACTION 7. PARTICULAR FW

TYPE	SENTENCE EXAMPLE
1	touch the circle **after** having pushed the cross to the left
1	put the cross on the left side and **after** grasp the circle
2	**move** the circle to the left **then the cross to the middle**
3	**put** first the triangle on the middle **and after on the left**
4	***push* the triangle** and *the circle* ***on the middle***
5	hit **twice** the blue circle
5	grasp the circle **two times**
6	put the cross to the right and **do a u-turn**
7	put **both** the circle and the cross to the right

could be easily reused or adapted for other computational or robotic experiments in language acquisition. In particular, we would like to extend this work by integrating our neural parser with multi-modal (e.g. vision, sensori-motor, ...) and behavioral robotic experiments [2]. For instance, the semantic and syntactic information of such complex sentences could be integrated into robotic experiments grounding linguistic symbols to robot behavior and to the visual modality [13], [14], [17]. Syntactic richness of natural language sentences are often simplified in such experiments (for the benefit of motor or visual modalities), and rather rely on stereotypical sequence of few semantic words without function words (e.g. "hit left blue"). Our model could help in such architectures by increase the syntactic variability a robotic architecture could deal with.

Conditions	*Default*	*INF*	*NOISE*
PHON	**18.49 (1.76)**	$\xrightarrow{=}$ 18.49 (1.76)	33.11 (0.77)
WORD	**18.12 (1.38)**	**16.51 (1.26)**	**29.73 (0.48)**
CONST	21.46 (1.41)	17.71 (1.49)	40.53 (0.77)

TABLE II
MEAN ERROR IN PERCENT (AND STANDARD DEVIATION) FOR FULL SENTENCE COMPREHENSION FOR DIFFERENT CONDITIONS.

Supplementary material and source code are available at https://github.com/neuronalX/Hinaut2018_icdl-epirob

REFERENCES

[1] E. Bastianelli, G. Castellucci, D. Croce, R. Basili, and D. Nardi, "Effective and robust natural language understanding for human-robot interaction." in *ECAI*, 2014, pp. 57–62.

[2] P. F. Dominey and J.-D. Boucher, "Developmental stages of perception and language acquisition in a perceptually grounded robot," *Cognitive Systems Research*, vol. 6, no. 3, pp. 243–259, 2005.

[3] M. Eppe, S. Trott, and J. Feldman, "Exploiting deep semantics and compositionality of natural language for human-robot-interaction," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*. IEEE, 2016, pp. 731–738.

[4] J. Gaspers, P. Cimiano, K. Rohlfing, and B. Wrede, "Constructing a language from scratch: Combining bottom–up and top–down learning processes in a computational model of language acquisition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 9, no. 2, pp. 183–196, 2017.

[5] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[6] J. Hatori *et al.*, "Interactively picking real-world objects with unconstrained spoken language instructions," *arXiv preprint arXiv:1710.06280*, 2017.

[7] X. Hinaut, J. Twiefel, M. Petit, P. F. Dominey, and S. Wermter, "A recurrent neural network for multiple language acquisition: Starting with english and french," in *NIPS 2015 Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, 2015.

[8] N. Iwahashi, "Robots that learn language: Developmental approach to human-machine conversations," in *Symbol Grounding and beyond*. Springer, 2006, pp. 143–167.

[9] M. Panzner, J. Gaspers, and P. Cimiano, "Learning linguistic constructions grounded in qualitative action models," in *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE, 2015, pp. 121–127.

[10] D. K. Roy, "Learning visually grounded words and syntax for a scene description task," *Computer speech & language*, vol. 16, no. 3-4, pp. 353–385, 2002.

[11] M. Spranger and L. Steels, "Emergent functional grammar for space," *Experiments in Cultural Language Evolution*, vol. 3, pp. 207–232, 2012.

[12] L. Steels, "The synthetic modeling of language origins," *Evolution of communication*, vol. 1, no. 1, pp. 1–34, 1997.

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

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

[15] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol emergence in robotics: a survey," *Advanced Robotics*, vol. 30, no. 11-12, pp. 706–728, 2016.

[16] M. Tenorth and M. Beetz, "Knowrob: A knowledge processing infrastructure for cognition-enabled robots," *The International Journal of Robotics Research*, vol. 32, no. 5, pp. 566–590, 2013.

[17] T. Yamada, S. Murata, H. Arie, and T. Ogata, "Dynamical integration of language and behavior in a recurrent neural network for human–robot interaction," *Frontiers in neurorobotics*, vol. 10, p. 5, 2016.