From Phonemes to Sentence Comprehension: A Neurocomputational Model of Sentence Processing for Robots

Xavier Hinaut

From Phonemes to Sentence Comprehension: A Neurocomputational Model of Sentence Processing for Robots

X. Hinaut[1,2,3]

1. Inria Bordeaux Sud-Ouest, Talence, France.
2. LaBRI, UMR 5800, CNRS, Bordeaux INP, Université de Bordeaux, Talence, France.
3. Institut des Maladies Neurodégénératives, UMR 5293, CNRS, Université de Bordeaux, Bordeaux, France.

MNEMOSYNE Team
xavier.hinaut@inria.fr

Abstract

There has been an important 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 speech API and integrate it inside a robotic architecture [3][11], one faces a non negligible number of wrong sentence recognition. Thus speech recognition can not be considered as solved (because many sentences out of their contexts are ambiguous). We believe that contextual solutions (i.e. adaptable and trainable on different HRI applications) have to be found. In this perspective, the way children learn language and how our brains process utterances may help us improve how robots process language. Getting inspiration from language acquisition theories and how the brain processes sentences we previously developed a neuro-inspired model of sentence processing [2][4]. In this study, we investigate how this model can process different levels of abstractions as input: sequence of phonemes, seq. 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.

Materials & Methods

Echo State Networks [7]

Update equation of the reservoir (recurrent layer) and the readout (output layer):

$$\mathbf{x}(t+1) = (1-\alpha)\mathbf{x}(t) + \alpha f(\mathbf{W}^{\mathbf{in}}\mathbf{u}(t+1) + \mathbf{W}\mathbf{x}(t)) \quad (1)$$

$$\mathbf{y}(t) = \mathbf{W}^{\mathbf{out}}\mathbf{x}(t) \quad (2)$$

Matrices $\mathbf{W}^{\mathrm{in}}$ and **W** are randomly generated.

Training of the output weights with ridge regression

$$\mathbf{W}^{\mathbf{out}} = \mathbf{Y}^{\mathbf{d}}\mathbf{X}^{\mathbf{T}}(\mathbf{X}\mathbf{X}^{\mathbf{T}} + \beta\mathbf{I})^{-1} \quad (3)$$

Model details and parameters

Number of reservoir units: 500.
Spectral radius: 1. Input scaling: 0.6. Reservoir weight std: 0.1. Leak-rate: 0.06. Regularization param.: 2.5×10^{-4} for PHON and WORD, and 5×10^{-6} for CONST. Infrequent word threshold for INF: 5.

Sentence examples produced by users

touch the circle **after** having pushed the cross to the left
put the cross on the left side and **after** grasp the circle
move the circle to the left **then the cross to the middle**
put first the triangle on the middle **and after on the left**
push* the triangle** and *the circle **on the middle
hit **twice** the blue circle
grasp the circle **two times**
put the cross to the right and **do a u-turn**
put **both** the circle and the cross to the right

Corpus is composed of 190 English sentences. [3][5]
Word to phoneme conversion is done with CMU dict.

Discussion

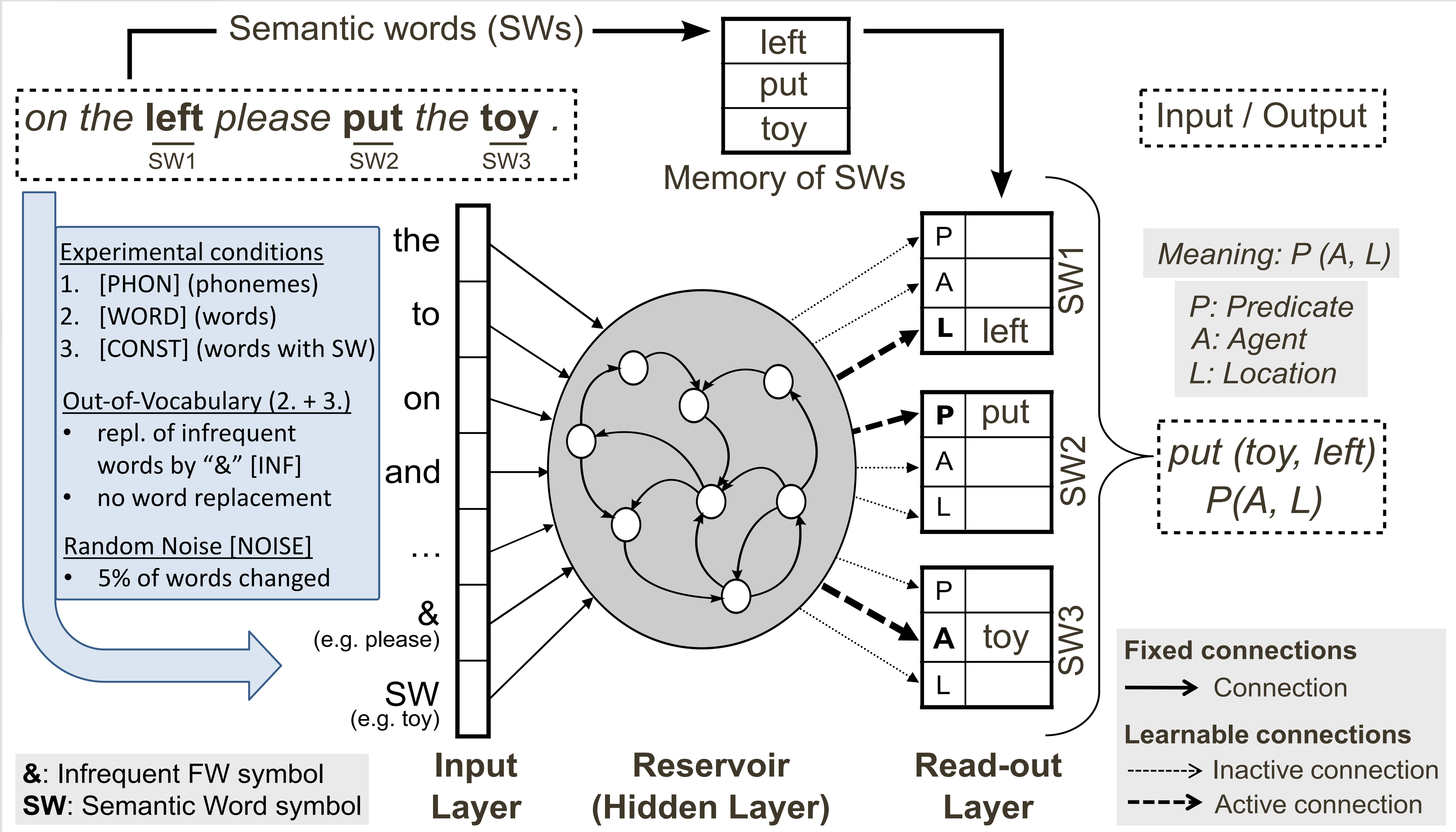
This study tried to understand what kind of input information is most relevant for learning to parse sentences with simple neurocognitive mechanisms (unstructured recurrent networks and Hebbian-like learning). Results showed that WORD condition performing best in normal conditions, but only from a short increase in performance.
We also explored noisy conditions, where 5% of the words were randomly replaced by other words. WORD and PHON conditions resisted better to noise than CONST condition.

Given these results, we speculate that the PHON condition would give better results than WORD cond. when dealing with real speech inputs.

In future work, we will process real speech data in order to: (1) test different speech recognizers that will provide sequences of phonemes or seq. of words; (2) use the recognized phonemes/words to train and test the current model, and see which condition PHON/WORD/CONST gives the best generalization.

Materials & Methods

Sentence parsing model with different input conditions



Example of one input sentence in different conditions

- Input: *"Point the triangle and then touch it"*
- Output: *point (triangle) ; touch (triangle)*

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	.
Grammatical 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	.

Results

Mean error in percent (with std) for full sentence comprehension

Conditions	*Default*	*INF*	*NOISE*
PHON	**18.49 (1.76)**	N/A	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)

Results for 10-fold cross-validation (4-fold for NOISE) averaged over 100 instances. Full sentence comprehension imply that all output roles are correctly recognized.

References

[1] M. Tomasello, Constructing a language: A usage based approach to language acquisition. Cambridge, MA: Harvard University Press, 2003.
[2] P. Dominey, M. Hoen, and T. Inui, "A neurolinguistic model of grammatical construction processing," Journal of Cognitive Neuroscience, vol. 18, no. 12, pp. 2088–2107, 2006.
[3] X. Hinaut, M. Petit, G. Pointeau, and P. Dominey, "Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks," Front in Neurorob, vol. 8, 2014.
[4] X. Hinaut and P. Dominey, "Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing" PloS one, 8, 2, p. e52946, 2013.
[5] 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 CoCo NIPS 2015 Workshop.
[6] A. Goldberg, Constructions: A construction grammar approach to argument structure. University of Chicago Press, 1995.
[7] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks" Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, vol. 148, p. 34, 2001.
[8] Marcus, G. F. et al. (1999). "Rule learning by seven-month-old infants ". Science 283, 77–80.
[9] 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.
[10] J. Bergstra, D. Yamins, and D. D. Cox, "Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms," in Proceedings of the 12th Python in Science Conference, pp. 13–20. 2013.
[11] J. Twiefel, X. Hinaut, M. Borghetti, E. Strahl, and S. Wermter, "Using Natural Language Feedback in a Neuro-inspired Integrated Multimodal Robotic Architecture," in Proc. of RO-MAN, New York City, USA, 2016.
[12] X. Hinaut and J. Twiefel, "Teach your robot your language! trainable neural parser for modelling human sentence processing: Examples for 15 languages," (Submitted).

Links

Video of Human-Robot Interaction:
youtu.be/FpYDco3ZgkU

Corpus and code:
github.com/neuronalX/EchoRob

Acknowledgments

This work was partly supported by the PHC PROCOPE (Campus France - DAAD) LingoRob project 37857TF. We thank Johannes Twiefel for very interesting discussions.