On-Line Processing of Grammatical Structure Using Reservoir Computing

Xavier Hinaut, Peter Ford Dominey

On-Line Processing of Grammatical Structure Using Reservoir Computing[*]

Xavier Hinaut[1,2], Peter F. Dominey[1,2]

[1] Stem Cell and Brain Research Institute, INSERM U846, 18 Avenue Doyen Lepine, 69500 Bron, France
[2] Université de Lyon, Université Lyon I, 69003, Lyon, France
{Xavier.Hinaut, Peter.Dominey}@inserm.fr

Abstract. Previous words in the sentence can influence the processing of the current word in the timescale of hundreds of milliseconds. The current research provides a possible explanation of how certain aspects of this on-line language processing can occur, based on the dynamics of recurrent cortical networks. We simulate prefrontal area BA47 as a recurrent network that receives on-line input of "grammatical" words during sentence processing, with plastic connections between cortex and striatum (homology with Reservoir Computing). The system is trained on sentence-meaning pairs, where meaning is coded as activation in the striatum corresponding to the roles that different "semantic words" play in the sentences. The model learns an extended set of grammatical constructions, and demonstrates the ability to generalize to novel constructions. This demonstrates that a RNN can decode grammatical structure from sentences in an on-line manner in order to generate a predictive representation of the meaning of the sentences.

Keywords: On-line Language Processing, Grammatical Structure, Language, Echo State Networks, Recurrent Neural Networks, Reservoir Computing, Prefrontal Cortex, Striatum.

1 Introduction

One of the most remarkable aspects of language processing is the rapidity with which it takes place. This is revealed perhaps most clearly in studies of evoked brain potentials (ERP) where a word that violates predictions about the developing meaning or grammatical structure can yield brain responses as rapidly as 200-600 ms [1,2,15]. This suggests that the brain is accumulating evidence on-line, and predicting or generating expectations about the subsequent structure of the incoming sentence.

Friederici and colleagues investigated language processing in patients with lesions of the basal ganglia and observed that these subjects failed to produce a normal P600

---

[*] This is a corrected version with minor revisions of the paper: *Hinaut X., Dominey P.F.: On-Line Processing of Grammatical Structure Using Reservoir Computing. In: A.E.P. Villa et al. (Eds.), ICANN 2012, Part I, LNCS 7552, pp. 596–603, 2012.* The revision concerns reservoir language processing in the discussion section.

in response to syntactic anomalies [3]. While there is accumulating evidence for the role of the cortico-striatal system in processing some aspects of the grammatical structure of language in real-time, the underlying mechanisms and their implementation in neural structures of the cortico-striatal system remains an important open research topic.

We have previously developed neural network and more computational models of thematic role assignment in sentence processing [4-8]. Thematic role assignment involves determining who did what to whom – or extracting the thematic roles (*agent*, *object*, *recipient*) for the verbs in a sentence. See Figure 1. Thus for the sentence "*The boy who took the ball from the man was bitten by the dog*" *boy* is the agent of *took*, and the object of *bitten*.
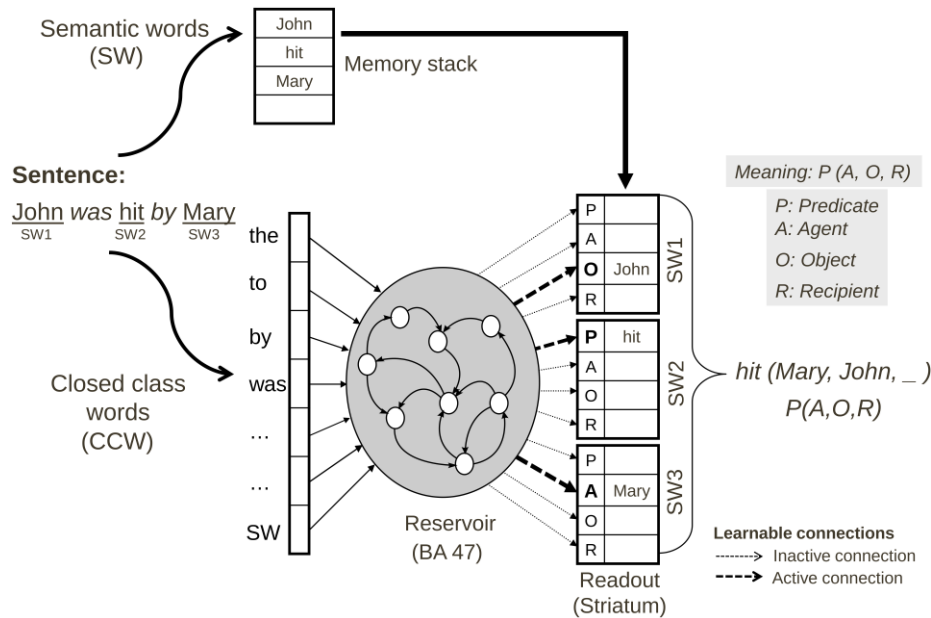

Fig 1. Thematic role assignment task. In the current experiments we focus on the bottom part of the figure with the CCW inputs, reservoir and the readout units. (Top) Each semantic word is kept in a memory stack and reassigned at the end to the thematic role (*predicate*, *agent*, *object*, *recipient*) identified by the model. (Bottom) Inputs are the closed class words plus a single input for all SW.

Our models are based on the principle that the information necessary to perform this thematic role assignment is encoded in the sentence by the configuration of grammatical function words (e.g. determiners, auxiliary verbs, prepositions) within the sentence [9]. The sequence of closed class words forms a pattern of activity within the RNN, and this pattern can be associated with the corresponding thematic role specification. In the reservoir computing (RC) approach (see [10] for a review) a RNN with fixed connections is used to encode the spatiotemporal structure of an

input sequence, and connections to a readout layer are trained using fast methods to produce a desired output in response to input sequences.

We test the hypothesis that the model illustrated in Figure 1 can be used to learn a set of grammatical constructions, and the following predictions: (1) On-line processing: Striatal (readout) activity will reflect an on-line estimation of the grammatical structure of the sentence. The final parse may be predicted before the end of the sentence, and this may change as new words invalidate the current parse. The changes in neural activity in the striatal readout may reflect language related ERPs recorded during human sentence processing. We will investigate this link with ERPs in more detail in future research. (2) Generalization: To a limited extent, the system should be capable of generalizing grammatical processing to new constructions if constituents of those constructions are represented in the initial training data.

2 Methods

2.1 Model

Our model makes a parallel between brain anatomy and reservoir computing framework. Prefrontal cortex (Brodmann area 47) is modeled as a fixed recurrent network and striatum as a separate population connected to cortex via modifiable synapses, corresponding respectively to the reservoir and readout. The reservoir is composed of leaky neurons with sigmoid activation. Equation (1) describes the internal update of activity in the reservoir:

$$x(t+1) = (1 - \alpha)\,x(t) + \alpha\,f(W_{in}\,u(t) + W_{res}\,x(t)) \ . \tag{1}$$

where $x(t)$ represents the reservoir state; $u(t)$ denotes the input at time t; α is the leak rate; and $f(\cdot)$ is the hyperbolic tangent (tanh) activation function. W_{in} is the connection weight matrix from inputs to the reservoir and W_{res} represents the recurrent connections between internal units of the reservoir. The initial state of the internal state $x(0)$ is zero. The linear readout layer is defined as:

$$y(t) = W_{out}[1\,;x(t)] \ . \tag{2}$$

where $y(t)$ is the output (striatal) activity (thematic roles) and W_{out} the output weights matrix. To learn the connection weights W_{out} between the reservoir (BA 47) and the readout (striatum), we used ridge regression.

The number of units used in the reservoir is 500 and 1000 (we tested two reservoir sizes). By definition, the matrices W_{in} and W_{res} are fixed and randomly generated. Internal weights (W_{res}) are drawn from a normal distribution with mean 0 and standard deviation 1, with a connectivity of 10%. Then we rescale the spectral radius – SR (largest absolute eigenvalue of the generated matrix W_{res}) – to 2. Input weight matrix W_{in} have values chosen randomly between 0.75 and -0.75 with a 50% chance.

The density of the input connections is also 10%. A leak rate of 0.1 was used. The regression parameter for which we found correct performances was 10^{-8}.

2.3 Corpus

Grammatical constructions were generated from a context-free-grammar for English with verbs taking 1 to 3 arguments (e.g. *walk*, *cut* or *give* could have respectively 1, 2 or 3 arguments). Each construction could have 0 or 1 relative clauses; the verbs of the relative clause could take 1 or 2 arguments. All possible word orders were used, e.g. there are 6 possible word orders for a verb with 2 arguments - APO, AOP, PAO, POA, OAP, OPA, *Agent Predicate Object* (APO) is the most common in English - and 24 possible word orders for a verb with 3 arguments - APOR, APRO, AOPR, *etc.*. All possible word orders were generated both for main and relative clause. All possible insertion of the relative clause was performed - after the 1st, 2nd or 3rd noun in the main clause. In total, 462 constructions were generated.

2.2 Input and Output Coding

Given an input sentence, the model should assign appropriate thematic roles to each noun. Sentences are represented in the input as grammatical constructions, where specific instances of semantic words (SW) - noun and verb - are replaced by *SW* markers. Thus, a given grammatical construction can code for multiple sentences, simply by filling in the *SW* markers with specific words. Constructions are presented to the reservoir in a sequential order, one word at a time. Learning consists in binding the appropriate role assignments to each SW in the readout neurons, which are specified during learning.

   Each dimension of the input codes one word or marker. Semantic words (also called open class or content words) are all coded with a single input neuron labeled SW. The input dimension is 13. This corresponds to: '-ed', '-ing', '-s', 'by', 'is', 'it', 'that', 'the', 'to', 'was', ',', '.', 'SW', with 11 slots for closed class words (CCW), 1 for SW, 1 for the comma and 1 for the period.

   The number of readout units is 42 (=6*4*2-6): 6 semantic words, each of which could have one of the 4 possible thematic role (*predicate, agent object, recipient*) that could be related to both the main and relative clause; as relative clause never admit verb with more than 2 arguments - *i.e.* there is no *recipient* - we have to subtract 6 output units from the total. Inputs have value 1 when the corresponding word is presented, 0 otherwise. Teacher outputs has value 1 if the readout unit should be activated, -1 otherwise.

2.4 Learning Conditions and Error Measures

We define a *meaning* as the role of a SW for either the main or relative clause (SW/clause). A *meaning* is obtained from the concerned readout units in 2 steps: threshold the activity at 0 and perform a winner-takes-all between the 4 possible roles;

the winning role is considered as the *meaning* of the model for this SW/clause. If there is no activity above the threshold, then no *meaning* is considered for this SW/clause. During learning, input sentences are presented, and the corresponding readouts coding the meaning are activated, and the system should learn the reservoir to readout weights.

Two error measures were evaluated: the *meaning error* (*m.err.*) and the *sentence error* (*s.err.*); the former is the percentage of *meanings* whose activity was erroneous; the latter is the percentage of sentences that were not fully understood (i.e. sentences in which there is at least one *meaning* erroneous). If some *meanings* are irrelevant they are ignored in the error measure; the *meaning* error is thus obtained averaging only on relevant *meanings*. For instance if there are only 2 SW in a construction, if one of them is erroneous, the *meaning* error will be 0.5 and the *sentence* error will be 1.

We considered conditions in which learning occurred starting at the presentation of the first word until the end of the sentence (*continuous learning*), or only at the end of the sentence (*sentence final learning*). In the first case, the system will begin to predict the meaning of the sentence from the outset of the sentence, and can thus display anticipatory and on-line activity.

In order to test the hypothesis that learning and generalization would be influenced by the grammatical structure of the corpus, we test the model in the same conditions but with *scrambled* data: for each construction the word order was randomized. We predict that in these conditions the model will learn the constructions, but will fail to generalize since there is no inherent structure to generalize over.

3 Results

3.1 On-line Grammatical Comprehension

In the *continuous learning* condition, the readout should give the correct role for each SW as soon as possible. In Figure 2, we can see that the neural activity represents the expected probabilities for each of the possible thematic roles. As successive words arrive, these probabilities are updated, illustrating a form of on-line reanalysis of the sentences. The neural activity reflects the statistics of the training corpus. Thus, in Figure 2 Left, early in the sentence, the model estimates that semantic word 1 is the agent of both verb 1 and the agent or the object of verb 2, reflecting their equal probability in the corpus. Only with the arrival of the 3rd semantic word does the system change this probability estimation for the final, correct one.

This example for semantic words 1 and 4 in this sentence is fully characteristic of the activity in the readout neurons for all of the sentences in the corpus. In all cases, neural activity evolves over the presentation of successive words, reflecting the current probability of the possible meaning at that point in the sentence.
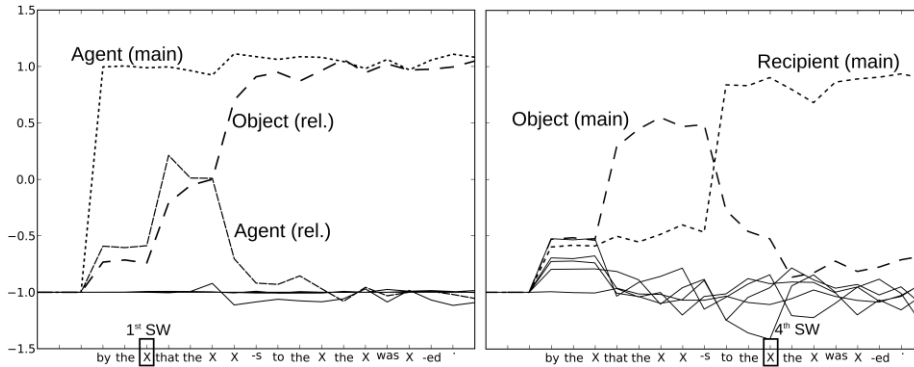
Fig. 2. Activity of readout units corresponding to the meaning of 1st (left) and 4th (right) Semantic Words for the *continuous learning* condition. The construction given in input is "*by the SW that the SW SW-s to the SW was SW-ed.*"; for instance the following sentence is an instantiation of this construction: "*by the cat that the dog bites to the mouse the cheese was given*". *main* and *rel.* indicate roles for main clause and relative clause respectively. 500 internal units were used for this experiment.

3.2 Generalization Capability

In order to evaluate the ability of the system to generalize to constructions not in the training set, we performed cross validation: we used 90% of corpus for training and 10% for testing, and averaged over 10 different reservoir instances. With 500 internal units, we obtained 9.0% ($\pm$0.5) *meaning error* (*m.err.*) and 27.6%($\pm$1.6) *sentence error* (*s.err.*) for *sentence final* learning (*SFL*), and 12.2% ($\pm$0.8) *m.err.* and 61.0% ($\pm$3.5) *s.err.* for *continuous* learning (*CL*). With 1000 internal units, we obtained 7.4% ($\pm$0.5) *m.err.* and 23.9% ($\pm$1.2) *s.err.* for *SFL*, and 8.4% ($\pm$0.5) *m.err.* and 37.9% ($\pm$2.6) *s.err.* for *CL*. Even better performances could be reached with more internal units and finer tuning of the (SR, τ) parameters, especially for the *CL* condition as it can be seen in Figure 3 b. (*right*).

Determining whether generalization depends on the grammatical structure within the corpus, with 500 internal units we observed that while the *scrambled* set (as the standard corpus) can be learned perfectly in the *SFL* condition, we obtained 75.1% ($\pm$0.8) *meaning* error and 99.8% ($\pm$0.2) *sentence* error rates for *final sentence* learning, and 67.4% ($\pm$0.9) and 99.6% ($\pm$0.2) for *continuous* learning respectively. Similar results are obtained with 1000 internal units. This means, that no construction can be generalized in the absence of underlying grammatical structure.

3.3 Network Learning and Parameter Sensitivity Analysis

In order to examine the evolution of learning performance, we performed an experiment where the model was successively exposed to an increasing sample of constructions. As illustrated in Figure 3 a., error is reduced as training set size

increases. To determine the robustness of the model to variations in parameters regulating network dynamics, we systematically varied SR and τ ($1/\alpha$ from Eqn. (1)). We observed good robust generalization in a large subset of this parameter space, illustrated in dark central diagonal bands in Figure 3 b.
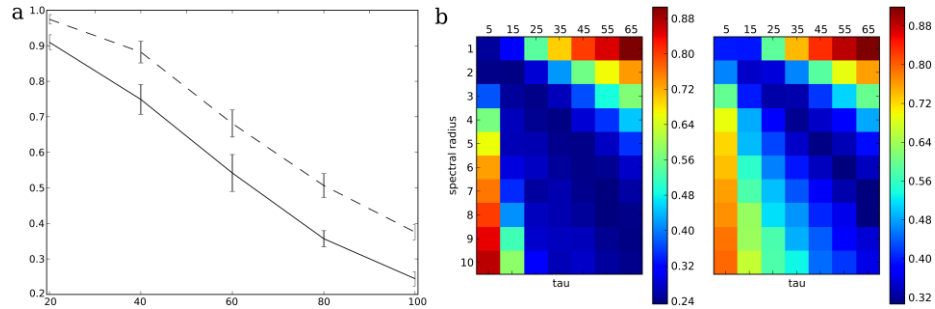

Fig. 3. a. Sentence errors as a function of the proportion of corpus used. Increasing corpus size (20%, 40%, 60%, 80% and 100%) was tested with 1000 internal units for *sentence final* (solid line) and *continuous* (dashed line) learning. **b**. Error rate as a function of spectral radius (1-10) and τ (5-65). Color scale indicates level of *sentence* error for *sentence final* learning (left) and *continuous* learning (right) for 1000 internal units. For all figures, errors were averaged over 10 instances.

4 Discussion

The study of language processing in recurrent networks has a rich history. Elman [11] showed how the simple recurrent network (SRN) with plastic recurrent connections formed a representation of the underlying syntactic structure of the learned language. Interestingly, Tong [12] showed that without explicitly learning internal representations echo state networks perform equally well in the task of predicting the next word during sentence processing. Few studies have examined thematic role assignment [13], and none have examined thematic role assignment for complex sentences in an on-line fashion.

The current research advances the state of the art in neural network processing of natural language in several respects. First, it is a reservoir-based neural network model of language processing whose task is to extract the meaning of the sentence, in terms of labeling the semantic roles of the open class words, as opposed to predicting the next word in the sentence [11]. Secondly, it is a reservoir-based neural network model of sentence processing whose neural dynamics reflect the ongoing assignment of meaning to semantic words. This "probabilistic" estimation of meaning in real-time may correspond to neural activity including the P600 observed in human sentence processing when such probabilities are violated.

Our performance may potentially be increased by adding *working memory* units: in [14] the authors use a reservoir to process a flow of characters. The task is to predict the next character, which is highly depending on current level of bracket nesting. By using additional units to save the current level of bracket nesting (up to 6 levels), they

show that the performances are highly increased with the use of these units. One way of enhancing our model could be to use this kind of units in order to memorize the nesting of relative clauses. Allowing such extensions to the hierarchical coding capability is a rich topic of future research.

References

1. Friederici A.D., Mecklinger A., Spencer K.M., Steinhauer K., Donchin E.: Syntactic parsing preferences and their on-line revisions: a spatio-temporal analysis of event-related brain potentials. Brain Res. Cogn. Brain Res. 11, 305-323 (2001)
2. Hagoort P.: The fractionation of spoken language understanding by measuring electrical and magnetic brain signals. Philos. Trans. R Soc. Lond. B Biol. Sci. 363, 1055-1069 (2008)
3. Frisch S., Kotz S.A., von Cramon D.Y., Friederici A.D.: Why the P600 is not just a P300: the role of the basal ganglia. Clin. Neurophysiol. 114, 336-340 (2003)
4. Dominey P.F., Hoen M., Blanc J.M., Lelekov-Boissard T.: Neurological basis of language and sequential cognition: evidence from simulation, aphasia, and ERP studies. Brain Lang. 86, 207-225 (2003)
5. Dominey P.F., Hoen M.: Structure mapping and semantic integration in a construction-based neurolinguistic model of sentence processing. Cortex 42, 476-479 (2006)
6. Dominey P.F., Hoen M., Inui T.: A neurolinguistic model of grammatical construction processing. J. Cogn. Neurosci. 18, 2088-2107 (2006)
7. Dominey P.F., Inui T.: Cortico-striatal function in sentence comprehension: insights from neurophysiology and modeling. Cortex 45, 1012-1018 (2009)
8. Dominey P.F., Inui T., Hoen M.: Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing. Brain Lang. 109, 80-92 (2009)
9. Bates E., Wulfeck B., MacWhinney B.: Cross-linguistic research in aphasia: an overview. Brain. Lang. 41, 123-148 (1991)
10. Lukoševičius M., Jaeger H.: Reservoir computing approaches to recurrent neural network training. Computer Science Review 3(3) 127-149 (2009)
11. Elman J.L.: Learning and development in neural networks: the importance of starting small. Cognition 48, 71-99 (1993)
12. Tong M.H., Bicket A.D., Christiansen E.M., Cottrell G.W.: Learning grammatical structure with Echo State Networks. Neural Networks 20(3), 424-432 (2007)
13. van der Velde F., de Kamps M.: Neural blackboard architectures of combinatorial structures in cognition. Behavioral and Brain Sciences 29(1), 31-70; discussion 70-108 (2006)
14. Pascanu R. and Jaeger H.: A neurodynamical model for working memory. Neural networks 24(2), 199-207 (2011)
15. Friederici A.D., Bahlmann J., Heim S., Schubotz R.I., Anwander A.: The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proc. Natl. Acad. Sci. USA 103, 2458-2463 (2006)
16. Oger toolbox, http://organic.elis.ugent.be/oger