Which Neural Network Architecture matches Human Behavior in Artificial Grammar Learning?

Andrea Alamia (andrea.alamia@cnrs.fr); Victor Gauducheau (gauducheau.victor@hotmail.fr); Dimitri Paisios (dimitripaisios@gmail.com); Rufin VanRullen (rufin.vanrullen@cnrs.fr)

CerCo, CNRS, Université de Toulouse - Toulouse, 31055 (France)

Abstract:

In recent years artificial neural networks achieved performance close to or better than humans in several domains: tasks that were previously human prerogatives, such as language processing, have witnessed remarkable improvements in state of the art models. One advantage of this technological boost is to facilitate comparison between different neural networks and human performance, in order to deepen our understanding of human cognition. Here, we investigate which neural network architecture (feed-forward vs. recurrent) matches human behavior in artificial grammar learning, a crucial aspect of language acquisition. Prior experimental studies proved that artificial grammars can be learnt by human subjects after little exposure and often without explicit knowledge of the underlying rules. We tested four grammars with different complexity levels both in humans and in feedforward and recurrent networks. Our results show that both architectures can "learn" (via error back-propagation) the grammars after the same number of training sequences as humans do, but recurrent networks perform closer to humans than feedforward ones, irrespective of the grammar complexity level. Moreover, similar to visual processing, in which feedforward and recurrent architectures have been related to unconscious and conscious processes, our results suggest that explicit learning is best modeled by recurrent architectures, whereas feedforward networks better capture the dynamics involved in implicit learning.An extended version of this work is available as preprint at: https://arxiv.org/abs/1902.04861

Keywords: Artificial Grammar Learning; Chomsky hierarchy; feedforward and recurrent neural networks; implicit learning.

Introduction

In recent years the field of neural networks has undergone a substantial revolution boosted by deep learning approaches (Lecun et al., 2015). Different architectures have reached human-like performance in domains that were previously considered as sole prerogative of the human brain, such as perception or language. Part of this success originates from insights provided by cognitive sciences, in which brain-inspired solutions are implemented in functional models (Hassabis et al., 2017). Conversely, it is possible to investigate the computational processes that take place in the human brain by comparing them with artificial functional models (Yamins and DiCarlo, 2016). For this purpose, Artificial Grammar Learning represents an ideal venue, given its well-established roots in both the cognitive and computer science literature. On the one hand, a formal definition of grammar complexity (i.e. Chomsky's hierarchy, Chomsky, 1956) provides a theoretical framework to study grammar learning; on the other hand, previous studies in humans set a well-defined experimental framework to compare human behavior with the performance of different neural network architectures. Previous studies have shown that participants can learn artificial grammars equally well irrespective of their level in the Chomsky hierarchy. However, it is still debated what determines participants' behavior, and different theoretical accounts have been proposed (Pothos, 2007): one theory suggested that participants learn the grammar rules implicitly; a revised account of the same hypothesis suggested that participants do not learn the full spectrum of grammatically correct transitions, but only a subset of it, such as bigrams or trigrams (Reber and Lewis, 1977). A further computational perspective posits that humans learn sequences as recurrent models, that is, by decoding relevant features from former adjacent items in the string(Cleeremans et al., 1989). However, contrary to human experiments in which subjects typically see a few dozen examples, all considered computational models have been trained with large datasets, and sometimes with significant overlap between training and test sets, making it difficult to draw any substantial comparisons with human cognition.

In this study we tested 4 grammars spanning over 3 Chomsky's hierarchy levels. Both human participants and artificial neural networks were trained and tested on datasets generated from those grammars. Importantly, we aimed to use comparable amounts of training for humans and artificial neural networks. Our purpose was to investigate which architecture—feed-forward vs. recurrent networks—better captures human behavior as a function of grammar complexity. Moreover, as AGL is an established framework to contrast implicit and explicit learning(Dienes and Perner, 1999), we aimed at testing whether these modes could be related respectively to feedforward and recurrent architectures, similarly to findings in visual perception (Lamme and Roelfsema, 2000).

Materials and Methods

Artificial grammar dataset

We performed 4 experiments with different artificial grammars (fig.1A), each composed of the same amount of correct and incorrect sequences. According to the Chomsky hierarchy, two grammars were *regular*, referred in the following as grammar A and grammar B, one was *context-free* and one was *context-specific*. Figure 1A shows in details how sequences were generated in each grammar, and provide few examples of correct and incorrect sequences.

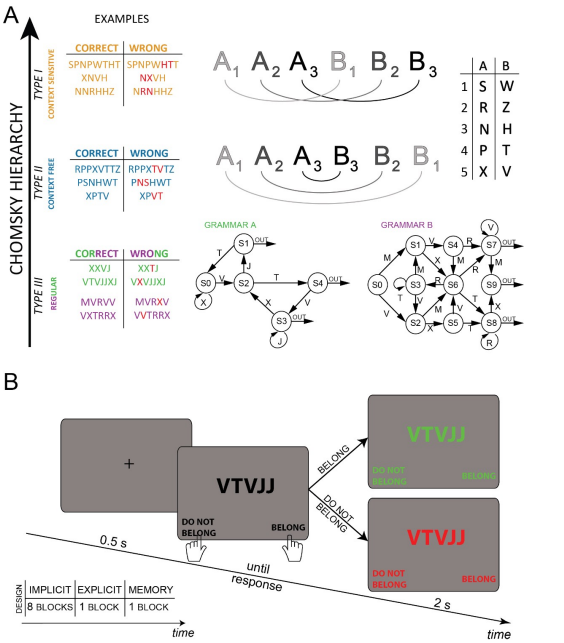

Figure 1. A) A schematic representation of the 4 grammars employed in this study, arranged according to the Chomsky hierarchy, and few examples of correct and wrong sequences. B) Time-course of each trial. All the human experiments employed the same design. The sub-table reports the number of blocks for each session: implicit, explicit and memory.

Humans

Experimental design. The same experimental design was applied for each grammar. Each trial started with a fixation cross, followed by a string of letters displayed in the center of the screen (fig.1B). Participants (N=15 for Grammar A, B and CF; n=11 for CS) were informed that there were two groups of respectively correct and wrong sequences, and they were asked to classify each presented sequence. Visual feedback was provided at the end of each trial. Each participant performed one session lasting approximately 1-hour and composed of 10 blocks. During the first 8 blocks, labeled as *implicit* in the sub-table in figure 1B, participants were not explicitly informed about the existence of the rules generating the sequences. Each block of the *implicit* part counted 60 trials, for 480 trials in total. A questionnaire was provided between the 8th and the 9th block to assess participants' explicit knowledge of the rules. The questionnaire was different for each grammar, asking specific questions about the rules. The last 2 blocks (labeled respectively as *explicit* and *memory* in figure 1B) served as control conditions. The results of thelast 2 blocks and the questionnaire are not shown here (see link to the full preprint in the abstract).

Artificial neural networks

Experimental design. The neural network design was composed of two parts: a first parameter search, and a subsequent comparison with human behavior. Both were implemented using the Keras library (Chollet, 2015), back-ended in Tensorflow (GoogleResearch, 2015). Altogether, we trained feedforward and recurrent architectures, each composed of a series of fully connected layers. All networks were trained to classify the sequences as correct or wrong, employing the same dataset (i.e. 4 grammars) and the same amount of trials as in the human experiments.

Regarding the parameter search, we aimed at determining the parameters whereby each architecture scored closest to human performance. We tested a range of networks, varying the *number of layers* and the *learning rate*, defining a 2-dimensional space. The training set was composed of 500 sequences (roughly similar to humans, who viewed 480 training examples), whereas the validation and testing set were composed respectively of 100 and 200 sequences. All the layers of a given network counted the same number of neurons except the output layer, which had only one neuron. The number of neurons was chosen such that all networks within the same 2-dimensional space had (roughly) the same number of free parameters. Each parameters space counted 6x20=120 networks, each one trained 20 times with random weights initialization. At first, we determined which networks provided the closest-to-human performance, defined as the smallest difference between the average performance of each network and the between-subjects mean performance on the last block. We selected the network with the smallest absolute difference as the one closest to human behavior. Once we determined the closest-to-human networks, we obtained their respective learning curves by varying the training set size progressively from 100 to 500 sequences, with a stride of 100.

Feedforward architectures. Feedforward neural networks were composed of fully connected dense layers. The input layer counted 12xK neurons, representing the one-hot encoding of the 12-letters longest possible string (K represents the total number

of letters, equal to 4, 5 and 10 for grammar A, B and CFG/CSG respectively). We employed zero-right padding when shorter sequences were fed to the network. All activation functions were defined as rectified linear units, i.e. 'ReLu', except the output neuron, which was implemented with a sigmoid function. The loss function was defined as 'mean-square-error', and optimized by means of stochastic gradient descent with Nesterov momentum set to 0.9, and decay equals to 1e-06. In all grammars, both in the parameter search and in the learning curve estimation, we considered 1 epoch, batch size of 15.

Recurrent architectures. Recurrent neural networks were composed of fully recurrent connected layers, in which each neuron was connected to itself and all other neurons in its layer. Starting from each sequence's first letter, at each time step the following letter was provided to the network as a one-hot encoded vector. The input layer thus counted as many neurons as letters in the grammars alphabet. A sigmoid activation function armed the output neuron, whose activation determined the classification decision after the last letter of the string was fed to the network. Learning occurred via back-propagation through time. As in the feedforward architecture, we considered 'mean-square-error' as loss function, optimized with the keras function rms-prop. We set rho to 0.9, epsilon to 1e-8 and decay to 0. As for the feedforward networks, we employed only 1 epoch (500 samples) and batch size of 15 in both the parameter search and the learning dynamic part.

Results

As shown in figure 2, at every level of the Chomsky's hierarchy, both participants and neural networks learned the rules above chance within the limited amount of trials. Regarding the NN, we first performed a parameter search in order to identify the networks whose performance was closest to the human one. Due to space limitation, we do not report this analysis here (see the abstract for the link to the preprint). However, regarding the feedforward networks, we reported that the best results were obtained with the lowest number of layers (corresponding to the highest number of neurons per layer), at each level of the Chomsky's hierarchy. Concerning the recurrent networks, we observed that 1) the highest accuracies corresponded to the lowest learning rates, and 2) in regular grammars, the closest-to-human network did not correspond to the one with the best performance. Finally, we obtained the learning curves for the 8 best-performing models (4 grammars * 2 network types) in order to compare them with the human ones. The results are shown in figure 2. Our analysis revealed that learning occurred over trials but differently for FF, RR and humans (all BF>>3e+15, error<0.01%). Specifically, a post-hoc analysis for each grammar separately suggests that recurrent architectures are closer to human behavior at every level of the Chomsky hierarchy, with the exception of grammar B, for which recurrent and feed-forward models cannot be distinguished, as further discussed below.

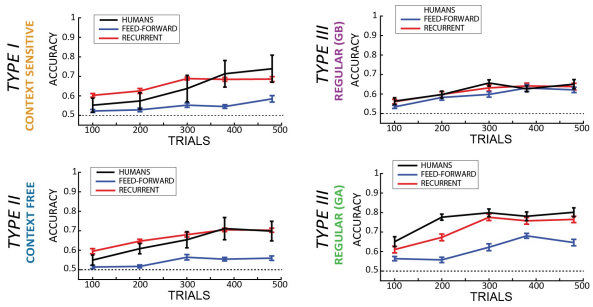

Figure 2. Results over trials for humans (in black) feedforward (in blue) and recurrent (in red) networks. For humans in (A) each bin is an average over 40 trials (20 trials before and after respectively, except the last bin which includes the last 40 trials of the experiment). Each panel represents a grammar type.

In order to shed some light on the difference we observed in the regular grammar B, we collected 8 additional artificial regular grammars from a recent review (Katan and Schiff, 2014) and tested both our recurrent and feedforward architectures on each grammar. Given our previous results, we hypothesized that RR networks would perform better (i.e., closer to humans) than FF networks in simpler grammars. Consequently, we defined 5 simple metrics to characterize the complexity of each grammar (see fig.3 for details).

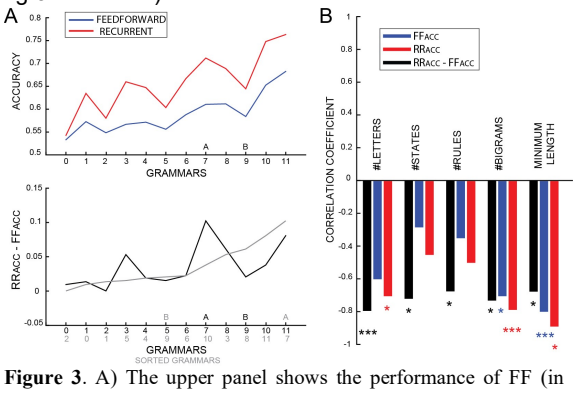

Figure 3. A) The upper panel shows the performance of FF (in blue) and RR (in red) architectures for each grammar (from 1 to 10 of the x-axis), averaging over 20 initialization over a training set of 500 trials. Grammars A and B are respectively 7 and 9. The difference in the performance between the two architectures is shown in the lower panel (the same sorted difference in light grey). B) Pearson indexes obtained correlating the performance of FF and RR networks and their difference (respectively in blue, red and black) with the 5 complexity metrics across the 10 grammars

Overall, the recurrent network always performed better than the feedforward one for all 10 grammars (fig.3A, upper panel). However, the difference between the two architectures was not constant across grammars. As shown in figure 3B, such difference correlated significantly (and negatively) with each of the complexity metrics we defined (see fig 3B; all BF>5, error<0.01%), suggesting that the difference between the two architectures is inversely correlated with the complexity of the grammars. Recurrent networks outperform feedforward ones in simpler, but less so in more complicated grammars.

Discussion

In this study we demonstrated that recurrent neural networks mimic human artificial grammar learning more closely than feedforward architectures, irrespective of the grammar's level within the Chomsky hierarchy. This result supports the hypothesis that recursion is key in cognitive processes such as language (Jackendoff, 2011). As already mentioned, previous studies showed that humans can learn to classify sequences as correct or not according to grammatical rules, and such knowledge appears to be to some extent implicit (Rohrmeier et al., 2012). However, in a usual AGL experiment, participants are first asked to memorize a set of sequences (training phase) and then to classify a new set as correct or not (testing phase). Here, we combined the two phases such that training and testing occur at the same time, allowing us to track the learning dynamics as it progresses. This design let us compare the participants' learning with the artificial networks' one. Importantly, we showed that both feedforward and recurrent neural networks can learn artificial grammars within the same limited number of trials as for human participants. However, the overall behavior of the recurrent networks, and in particular their learning dynamics, was closer to human behavior.

Previous studies have investigated mostly recurrent rather than feedforward networks in AGL, even though implicit learning has been frequently related to feedforward processes(Lamme and Roelfsema, 2000; Boly et al., 2011). Furthermore, most of the studies employed thousands (or tens of thousands) of sequences to train the models' parameters. In our study, we implemented a fairer comparison with human performance, as both artificial networks and human participants were trained on the same limited number of examples (~500). Moreover, differently than previous studies on artificial grammar learning in humans, we adopted an experimental design in which training and testing occurred at the same time, allowing a direct comparison of learning dynamics (i.e., learning curves) between humans and artificial neural networks. All in all, this comparison reveals that recurrent models perform closer to humans than feedforward ones, except in more complicated –and supposedly implicit- grammars. For those grammars (e.g. Grammar B), human performance remains poor, and can equally well be accounted for by recurrent or feedforward models. In the two regular grammars we observed a significant difference in participants' awareness of the rules. In grammar A participants performed better at the questionnaire than in grammar B, coherently with the hypothesis that simpler grammars are more likely to be learnt consciously (Sun and Peterson, 1994; Halford et al., 1998). The distinction between implicit and explicit processes has been expressly designed in an integrated model tested also on artificial grammar tasks (Sun, 2006; Sun et al., 2007), providing evidence in favor of the hypothesis that implicit processes precede explicit ones (Windey and Cleeremans, 2015) and are prominently involved in complex grammars (Reber, 1976; Halford et al., 1998). Interestingly, our results reveal that in the more implicit grammar A, both ANN architectures reliably tracked human behavior, whereas only recurrent networks achieved this goal in more explicit grammars. This result draws a compelling parallel between feedforward/recurrent and implicit/explicit processes, consistently with results in visual perception (Lamme and Roelfsema, 2000; VanRullen and Thorpe, 2001) and neuroscience (Koch et al., 2016).

References

Boly M, Garrido MI, Gosseries O, Bruno MA, Boveroux P, Schnakers C, Massimini M, Litvak V, Laureys S, Friston K (2011) Preserved feedforward but impaired top-down processes in the vegetative state. Science (80-) 332:858–862.

Chollet F (2015) Keras: Deep Learning library for Theano and TensorFlow. GitHub Repos:1–21.

Chomsky N (1956) Three models for the description of language. IRE Trans Inf Theory 2:113–124.

Cleeremans A, Servan-Schreiber D, McClelland JL (1989) Finite State Automata and Simple Recurrent Networks. Neural Comput 1:372–381

Dienes Z, Perner J (1999) A theory of implicit and explicit knowledge. Behav Brain Sci 22:735-755-808.

GoogleResearch (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. Google Res

Hassabis D, Kumaran D, Summerfield C, Botvinick M (2017) Neuroscience-Inspired Artificial Intelligence. Neuron 95:245–258.

Jackendoff R (2011) What is the human language faculty?: Two views. Language (Baltim) 87:586–624

Katan P, Schiff R (2014) Does complexity matter? Meta-Analysis of learner performance in artificial grammar tasks. Front Psychol 5.

Koch C, Massimini M, Boly M, Tononi G (2016) Neural correlates of consciousness: progress and problems. Nat Rev Neurosci 17:395

Lamme VAF, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23:571–579.

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521:436–444.

Pothos EM (2007) Theories of artificial grammar learning. Psychol Bull 133:227–244.

Reber AS (1976) Implicit learning of synthetic languages: The role of instructional set. J Exp Psychol Hum Learn Mem 2:88–94.

Rohrmeier M, Fu Q, Dienes Z (2012) Implicit Learning of Recursive Context-Free Grammars. PLoS One 7.

Sun R (2006) The CLARION cognitive architecture: Extending cognitive modeling to social simulation. Cogn MultiAgent Interact:79–99 Available at: http://ebooks.cambridge.org/ref/id/CBO9780511610721.

Sun R, Peterson T (1994) Learning in reactive sequential decision tasks: the CLARION model. In: Proceedings of International Conference on Neural Networks (ICNN'96), pp 1073–1078

VanRullen R, Thorpe SJ (2001) The time course of visual processing: From early perception to decision-making. J Cogn Neurosci 13:454–461.

Windey B, Cleeremans A (2015) Consciousness as a graded and an all-or-none phenomenon: A conceptual analysis. Conscious Cogn

Yamins DLK, DiCarlo JJ (2016) Using goal-driven deep learning models to understand sensory cortex. Nat Neurosci 19:356–365.