Pre-Wiring and *Pre-Training*: What does a neural network need to learn truly general identity rules?

Raquel G. Alhama and **Willem Zuidema**
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
{rgalhama,w.h.zuidema}@uva.nl

Abstract

In an influential paper, Marcus et al. [1999] claimed that connectionist models cannot account for human success at learning tasks that involved generalization of abstract knowledge such as grammatical rules. This claim triggered a heated debate, centered mostly around variants of the Simple Recurrent Network model [Elman, 1990]. In our work, we revisit this unresolved debate and analyze the underlying issues from a different perspective. We argue that, in order to simulate human-like learning of grammatical rules, a neural network model should not be used as a *tabula rasa*, but rather, the initial wiring of the neural connections and the experience acquired prior to the actual task should be incorporated into the model. We present two methods that aim to provide such initial state: a manipulation of the initial connections of the network in a cognitively plausible manner (concretely, by implementing a "delay-line" memory), and a pre-training algorithm that incrementally challenges the network with novel stimuli. We implement such techniques in an Echo State Network [Jaeger, 2001], and we show that only when combining both techniques the ESN is able to learn truly general identity rules.

1 Introduction

One of the crucial aspects of language is that it allows humans to produce and understand an unlimited number of utterances. This is possible because language is a rule-governed system; for instance, if we know that the English present participle is formed by appending *-ing*, then we readily *generalize* this pattern to novel verbs. Accounting for how humans learn these abstract patterns, represent them and apply them to novel instances is the central challenge for cognitive science and linguistics. In natural languages there is an abundance of such phenomena, and as a result linguistics has been one of the main battlegrounds for debates between proponents of symbolic and connectionists accounts of cognition. One of the most heated debates was concerned with accounting for the regular and irregular forms of the English past tense. Rumelhart and McClelland [1986] proposed a connectionist model that allegedly accounted for the regular and irregular forms of the past tense. However, this model was fiercely critized by Steven Pinker and colleagues [Pinker and Prince, 1988, Pinker, 2015], who held that rules are essential to account for regular forms, while irregular forms are stored in the lexicon (the 'Words-and-Rules' theory).

A similar debate emerged with the publication of Marcus et al. [1999], this time centered on experimental results in Artificial Grammar Learning. The authors showed that 7 month old infants generalize to novel instances of simple ABA, ABB or AAB patterns after a short familiarization. Crucially, this outcome could not be reproduced by a Simple Recurrent Network (SRN) [Elman, 1990], a result that was interpreted by the authors as evidence in favour of a symbol-manipulating system:

> Such networks can simulate knowledge of grammatical rules only by being trained on all items to which they apply; consequently, such mechanisms cannot account for how humans generalize rules to new items that do not overlap with the items that appear in training. [Marcus et al., 1999, p. 79]

This claim triggered many replies, some of which proposed variations of the original model. However, in this debate the issues of whether neural networks are capable at all of *representing* general rules, of whether backpropagation is capable of *finding* these general rules from an arbitrary initial state or only from an appropriately chosen initial state are sometimes conflated. The latter issue – what initial state does a neural network model need to have success in the experiment – has, in our view, not received enough attention (but see Seidenberg and Elman [1999a], Altmann [2002]). This will be therefore the focus of this paper, in which we explore two directions. First, we ask which initial values of the connection weights could encourage generalization while remaining cognitively plausible (*pre-wiring*); second, we investigate the role of previous experience in creating an initial state in the network that would facilitate generalization (*pre-training*). We employ a prewiring and a pretraining technique in an Echo State Network (ESN) (Jaeger, 2001), and show that only when combining both techniques the ESN is able to accurately generalize to novel items.

2 Background

2.1 Empirical Data

Marcus et al. [1999] investigate the generalization abilities of 7 month old infants by conducting three Artificial Grammar Learning experiments. In their first experiment, the participants are familiarized to syllable triplets that follow a certain grammar: ABA for a randomly assigned group of infants, and ABB for the other. The stimuli contain 16 different triplets, each repeated 3 times. Those triplets are arranged in a 2-min. auditory speech stream, such that syllables are separated by a pause of 250 ms, and triplets of syllables are separated by 1s.

After the familiarization, the infants participate in a test phase, in which their looking times (to the speaker device that plays the stimuli) are recorded. The speaker plays a randomized set of triplets from both grammars, in order to see if infants can discriminate between them. Crucially, the test triplets contain syllables that were not used in the familiarization stimuli.

The results show a statistically significant difference between mean looking times to consistent and inconsistent grammars in both group of infants. The authors then conclude that infants can discriminate among ABA and ABB grammars.

Table 1: Stimuli used in experiment 2 in Marcus et al. [1999].

	Familiarization				Test
ABA	le di le le je le le li le le we le	wi di wi wi je wi wi li wi wi we wi	ji di ji ji je ji ji li ji ji we ji	de di de de je de de li de de we de	ba po ba ko ga ko
ABB	le di di le je je le li li le we we	wi di di wi je je wi li li wi we we	ji di di ji je je ji li li ji we we	de di di de je je de li li de we we	ba po po ko ga ga
3x triplet (random order)					

This experiment was repeated with a more carefully controlled set of syllables, which we report in table 1. Infants also exhibit significantly different behavioural responses in this second experiment. Finally, an additional experiment was performed, in this case using AAB vs. ABB grammars, in order to determine whether the rule learnt before was simply the presence or absence of an immediate repetition. Infants also showed significantly different responses in this experiment.

In the light of these results, the authors concluded that: (i) 7 m.o. infants can extract grammar-like rules, (ii) they can do it not based solely on statistical information (as would be evidenced from the additional controls in experiment 2, and (iii) the extracted rule is not merely the presence or absence of an immediate repetition.

2.2 Generalization and Neural Networks

Marcus [1998] argues that certain types of generalizations are unattainable for certain types of neural networks: concretely, those that lie *outside the training space*. The author defines *training space* as the combination of all feature values that network has witnessed during training. If there exist feature values that have never appeared during training, any item displaying that feature value lies outside the training space. For neural networks that are trained with the backpropagation algorithm, generalization to items outside the training space is, according to the author, extremely unlikely to occur due to what he calls *training independence*, which stands for the fact that the algorithm updates the weights of nodes independently of the activations of other nodes in the same layer.

In Marcus et al. [1999], the authors provide empirical evidence in support of this idea, by simulating the presented experiment in a Simple Recurrent Network (SRN) [Elman, 1990], a neural network architecture that incorporates an additional context layer that maintains an exact copy of the hidden layer and presents it to the network in the subsequent timestep, providing the model with memory in this way. The SRN is trained to predict the next syllable in the familiarization stimuli, and then tested on its ability to predict the final syllable of test items consistent with the familiarization grammar. This model failed to produce the correct predictions, confirming the hypothesis of the researchers.

Some following publications proposed to change the encoding of the input (Christiansen and Curtin [1999], Christiansen et al. [2000], Eimas [1999], Dienes et al. [1999], Altmann and Dienes [1999], [McClelland and Plaut, 1999]), the task (Seidenberg and Elman [1999a], Seidenberg and Elman [1999b], the neural network architecture (Shultz [1999], Sirois et al. [2000], Shultz and Bale [2001]), or – relevant to our work — incorporating some form of pre-training (Seidenberg and Elman [1999a], Altmann [2002]). Many of these models were subject of criticism by Marcus (Marcus [1999a],Marcus [1999b],Marcus [1999c],Marcus [1999d]), who argued that the models either involved some form of symbolic manipulation or did not adequately represent the experiment. About the model of Altmann [2002], which involves pre-training similar to the regime we explore in section 5, Marcus [1999e] points out, without giving any details, that, even if the model distinguishes grammatical from ungrammatical stimuli to some degree, it is unclear whether the model can actually learn the underlying general rule or discovers some heuristic that weakly correlates with it. In our work, we employ a neural network architecture that was not previously explored for this task (an Echo State Network, a type of Reservoir Computing network), and we report additional performance measures that tell us more about how general the learned rules are.

3 Simulations with a Simple Recurrent Network

Before presenting our simulations with the ESN model, we report our replication of the original simulations. We implement a Simple Recurrent Network as described in Elman [1990], and we train it to predict the next syllable in the input. As in Marcus et al., we use distributional encoding of phonetic features (based on Plunkett and Marchman [1993]). But unlike the original simulations, we do not encode the pause between triplets as an additional symbol; instead, we do not update the weights in the network when it predicts the first syllable of the next triplet.

In order to remain close to the test used in the experiments with infants, we test the network on both consistent and inconsistent sequences. We take the predicted vector for the third syllable of each triplet, and we find the closest vector that corresponds to one of the seen syllables (both from training and from test). We then evaluate whether the accuracy for consistent and inconsistent triplets is significantly different (for 16 runs of the model, equivalent to the number of infants in the experiment).

The test set used in the original experiments, as can be seen in Table 1, is based solely in two triplet types of each grammar. For this reason, we also evaluate our model with an extended test set that contains 5 additional random novel syllables of each type (A and B), consisting therefore of 25 test triplets.

We try 160 parameter settings for each familiarization grammar, varying the hyperparameters of the model: the size of the hidden layer, the learning rate and the number of epochs [1] . Figure 1 shows the proportion of these runs that yield a significant difference in the predictions for the two classes of test items (those that are consistent with the grammar used in training and those which are not). For the

[1]We found that the values of the three hyperparameters had a significant effect on the accuracy of the predicted syllables in the test.

responses that are significantly different, we separate those for which the neural network responds better to the consistent grammar (in white) from those in which the inconsistent grammar is favoured (in grey).

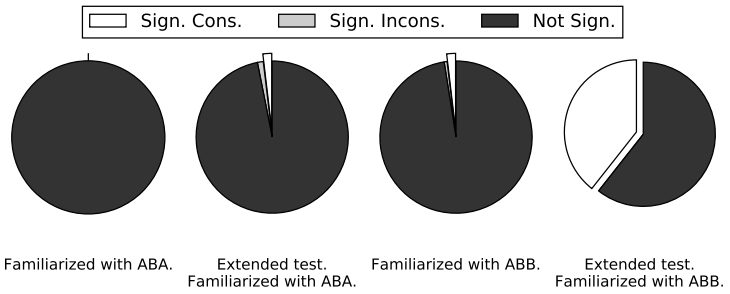

Figure 1: Proportion of parameter settings that yield significant (white), non-significant (dark grey) and inconsistently significant (light grey) responses, for simulations with an SRN.

As shown in the graphic, most of the simulations yield non-significant responses between grammars, in spite of a notable proportion of significant responses in the ABB condition for the extended test, possibly due to the fact that immediate repetitions are easier to learn[2]. We therefore confirm that the Simple Recurrent Network does not reproduce the empirical findings.

4 Simulations with an Echo State Network

Recurrent Neural Networks, such as the SRN, can be seen as implementing memory: through gradual changes in synaptic connections, the network learns to exploit temporal regularities for the function it is trained on. An alternative way to learn time-dependent relations is that offered by Reservoir Computing (RC) approaches, such as the Liquid State Machine [Maass et al., 2002] and the model adopted here, the Echo State Network (ESN) [Jaeger, 2001, Frank and Čerňanskỳ, 2008]. In RC models, the weights in the hidden layer (which is dubbed "reservoir") remain untrained, but – if satisfying certain constraints (the so-called "Echo State Property", which depends on the scaling of the weights in the reservoir based on the spectral radius parameter) – the dynamics exhibited by the reservoir "echo" the input sequence: some memory of the input lingers on for some time in the recurrent connections. In other words, the state of the reservoir depends on the fading history of the input; after a long enough input, the initial state does not determine the final states of the network.

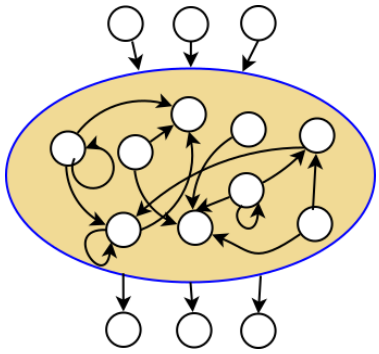

Figure 2: The Echo State Network.

The formalization of the ESN model is as follows. For an input u at time t, the activation x of the nodes in the reservoir is defined as:

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

where W^{in} are the input weights, W^{res} are the internal weights of the reservoir, and f is a non-linear function, generally $tanh$.

The activation of the output is defined as:

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

where W^{out} are the weights that connect the reservoir with the output nodes, and f^{out} is a function, which might be different from the function applied to the reservoir; in fact, it often consists on a simple identity function.

We implement a basic ESN with tanh binary neurons, and we follow the same procedure described in section 3 to train the network with backpropagation[3]. We try 200 parameter settings for each

[2]This was also observed in the SRN model in Altmann [2002].

[3]We have also run simulations with Ridge Regression, with similar results.

familiarization grammar, varying the hyperparameters of the model: the number of nodes in the reservoir, the input scaling, the spectral radius, learning rate and epochs.[4] Figure 3 shows the proportion of these runs that yield a significant difference in the predictions.

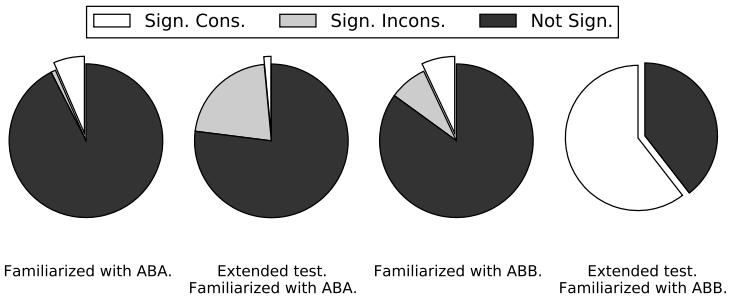

Figure 3: Proportion of parameter settings that yield significant (white), non-significant (dark grey) and inconsistently significant (light grey) responses, for the simulation with the basic ESN.

As can be seen, the results based on the Marcus et al. test set differ greatly from those in our extended test. This confirms our intuition that the amount of test items is crucial for the evaluation. For this reason, we base our analysis of the behaviour of the model in the extended test; however, it is important to notice that the amount of test items could have also played a role in the actual experiments with infants (see also section 7).

The plots of the extended test condition clearly show an assymetry between the grammars: more than half of the parameter settings yield significant responses for the ABB, while in the case of ABA, less than a quarter of the simulations are significant, and most of them are actually favouring the inconsistent grammar, which is precisely ABB. As mentioned before, the reason for this assymetry is probably due to the fact that immediate repetitions are easier to learn, since they are less affected from the decay of the activation function; for now, it suffices to say that the behaviour of the model towards ABA does not suggest that it could be a potential explanation for the experimental results.

5 *Pre-Wiring*: Delay Line Memory

In order to succeed in the prediction task, the model must predict a syllable that is identical to one presented before. In the previous simulations, we relied on the memory that ESNs offer through the recurrent connections and the Echo Property [Jaeger, 2002]. However, there exist several computational alternatives that brains may use to implement memory [Chaudhuri and Fiete, 2016]. We now explore one such model of memory: a delay line.

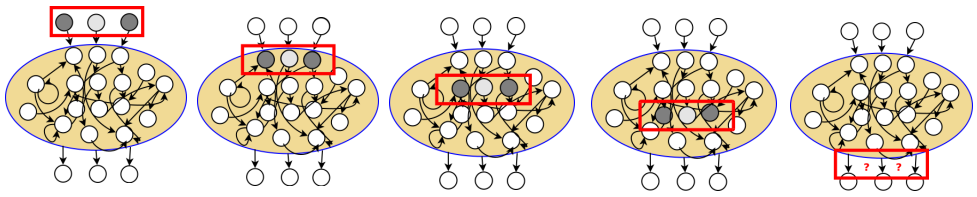

Figure 4: Depiction of five timesteps in the DeLi-ESN. The highlighted nodes show the activations in the delay line (the activation of the rest of the nodes is not illustrated).

Computationally, a delay line is a mechanism that ensures the preservation of the input by propagating it in a path ("line") that implements a delay. In the brain, delay lines have been proposed as part of the sound localization system [Jeffress, 1948], and they have been identified through intracellular

[4] We found that the values of the input scaling and the learning rate had a significant effect on the accuracy of the predicted syllables in the test.

recordings in the barn owl brain [Carr and Konishi, 1988]. In a neural network, a delay line is naturally implemented by organizing a subnetwork of neurons in layers, with "copy connections" (with weights 1 between corresponding nodes and 0 everywhere else) connecting each layer to the next. In this way, the information is kept in the network for as many timesteps as dedicated layers in the network (see figure 4).

We implement a delay line memory in the ESN (creating thus a new model that we call *DeLi-ESN*) by imposing this layer structure in the reservoir. We run 1200 combinations of parameter settings with the DeLi-ESN, including also a parameter that establishes some amount of noise to add to the weights of the reservoir. In this way, some models contain a less strict delay line; the greater the noise, the closer the model is to the original ESN.

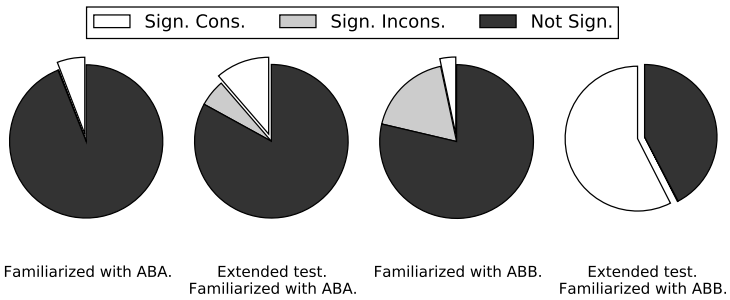

Figure 5: Proportion of parameter settings that yield significant, non-significant and inconsistently significant responses in the tests, for the simulation with DeLi-ESN.

The results, illustrated in Figure 5, show an increased number of significant responses (in favour of the consistent grammar) for the extended test of ABA familiarization: the addition of the delay line memory indeed helps in the detection of the identity relation. But in spite of the positive effect of the delay line, we need to ask ourselves to what extent these results are satisfactory. The pie plots show the likelihood of obtaining the results that Marcus et al. found in their experiments with our model, and in order to do so, we use the same measure of success (i.e. whether the responses for each grammar are significantly different). However, the models hardly ever produce the correct prediction[5]. For this reason, in the next section, we adopt a stricter measure of success. We discuss this issue further in section 7.

6 *Pre-Training*: Incremental-Novelty Exposure

The infants that participated in the original experiment had surely been exposed to human speech before the actual experiment; however, in most computational simulations this fact is obviated. We hypothesize that prior perceptual experience could have triggered a bias for learning abstract solutions: since the environment is variable, infants may have adapted their induction mechanism to account for novel instances. We now propose a method to pre-train a neural network that aims to incorporate this intuition.

In this training regime —which we call Incremental-Novelty Exposure, or INE for short— we iteratively train and test our model; so for a certain number of iterations i, the model is trained and tested i times, with the parameters learnt in one iteration being the initial state of the next iteration. The test remains constant in each of these iterations; however, the training data is slightly modified from one iteration to the next. The first training set is designed according to the Marcus et al. stimuli: 4 syllables of type A and 4 syllables of type B are combined according to the pattern of the familiarization grammar (ABA or ABB). In the second iteration, one syllable of type A and one of type B are deleted (that is, all the triplets involving those syllables are removed from the training set), and a new syllable of type A and one of type B are incorporated, such that new

[5]We find that the values of the input scaling, learning rate, spectral radius, reservoir size, and reservoir noise each have a significant effect on the accuracy of the predicted syllables in the test, although exact prediction accuracy remains low (rarely above 12% for ABA and above 20% for ABB familiarization) even for the best combination of parameters.

triplets of the familiarization pattern are generated with the new syllables (combined as well with the already-present syllables). Therefore, for each training iteration, the model is exposed to a similar training set as the previous iteration, but there is a small amount of novelty.

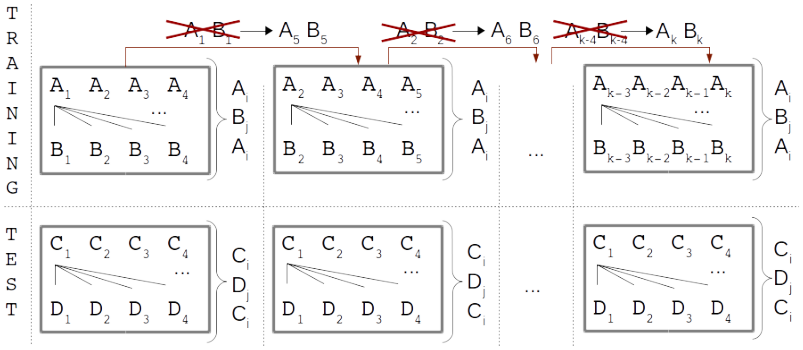

Figure 6: Depiction of the Incremental-Novelty Exposure. Barred syllables are removed from the training set after the training of the corresponding iteration has finished (so they do not remain in the training set of the next iteration).

We simulate 600 different hyperparameter configurations, varying the reservoir size, noise in the delay line, input scaling, spectral radius, learning rate, and epochs. Figure 6 illustrates how the mean accuracy evolves at each stage of the INE procedure of one representative run. As it can be seen in the graphs, the accuracy is really low in the beginning (corresponding to a simulation without pre-training) but, with more iterations –and thus with more novel items incorporated in the training set–, the model becomes better, presumably by finding a more general solution.

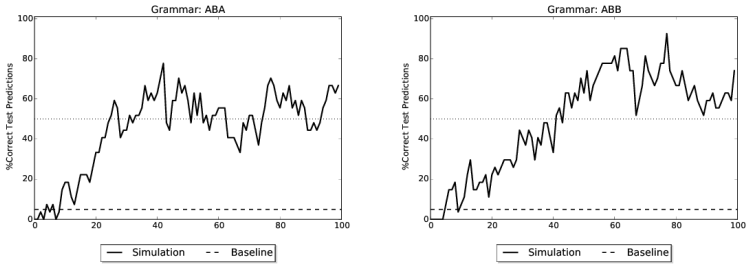

Figure 7: Performance over 100 iterations with Incremental Novelty Exposure, for one representative run in the ABA familiarization condition (left) and one in the ABB condition (right).

In order to test that these results are robust, we compute the mean over the accuracy for the last quarter of the tests (in our case, the last 25 tests, corresponding to the rightmost curve in the graph), for a few runs. The results are fairly similar in each run, as can be seen in figure 8.

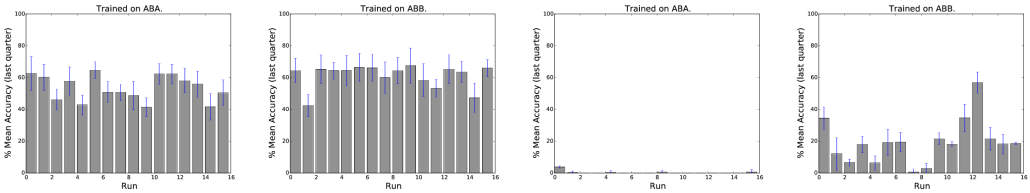

Figure 8: Mean accuracy for the tests in the last quarter simulation in the INE simulation, for DeLi-ESN (left) and basic ESN (right).

Both graphs show that the combination of the DeLi-ESN and the INE drastically boosts the generalization capabilities of the ESN. However, we should identify what is the contribution of the DeLi-ESN. Figure 8 shows the mean accuracy (again, for the last quarter of the regime), for 16 runs

of the basic ESN in the INE regime. The effect of the delay line memory is blatantly clear: when removed, the accuracy is close to 0 for ABA, and mostly around 20% for ABB.

7 Discussion

The Marcus et al. publication conveyed two bold statements: first, that infants spontaneously extract identity rules from a short familiarization, and second, that neural networks are doomed to fail in such simple task. Our work suggests that both the initial optimism for the generalization abilities of infants and the pessimism towards neural networks were overstated.

Our work investigates the initial conditions that a neural network model should implement in order to account for the experimental results. First, we have proposed that networks should be *pre-wired* to incorporate a bias towards memorization of the input, which we have implemented as a delay line. With such *pre-wiring*, the model yields a notable proportion of significantly different responses between grammars. But despite such apparent success, the accuracy of the model in syllable prediction is very low.

Therefore, even though the successful discrimination between grammars is generally understood as abstraction of the underlying rule, our results show that significantly different responses can easily be found in a model that has not perfectly learnt such a rule. The corollary is that the generalization abilities of infants may have been overestimated; as a matter of fact, null results in similar experiments also point in that direction (see for instance footnote 1 in Gerken [2006]; also Geambasu&Levelt, p.c.).

But can neural networks go beyond grammar discrimination and accurately predict the next syllable according to a generalized rule? Our work shows that this can be achieved when prior experience is incorporated in the model. We have hypothesized that, from all the information available in the environment, it is the gradual exposure to novel items what enhances generalization. This particular hypothesis deviates from related studies in which (i) an SRN was pre-trained to learn the relation of *sameness* between syllables [Seidenberg and Elman, 1999a], and (ii) an SRN was pre-trained with a set of sentences generated from a uniformly sampled vocabulary. Although the data used in our pre-training is less realistic than that used by Altmann [2002], our evaluation method is more strict (since we aim to test for accuracy rather than discrimination); for this reason, we first need to evaluate a model with a more constrained input. The next step in future work should be to explore whether the same results can be obtained when input data involving gradual novelty is generated from a grammar unrelated to the actual experiment.

Finally, from the perspective of the symbol vs. associations debate, at some abstract level of description, the delay line may be interpreted as providing the model with variables (that is, the dedicated group of nodes that encode the input at a certain time may be seen as a register) and the symbolic operation of "copy". It should be noted though that these groups of nodes are not isolated, and therefore, the learning algorithm needs to discover the structure in order to make use of it. Furthermore, it is uncontroversial that items are kept in memory for a certain lapse of time, so this structure is unlikely to constitute the core of the symbolic enterprise. If nevertheless our model is seen as compatible with the theory of rules-over-variables, our approach would then constitute a fundamental advancement for the field in providing a unifying model in which both theories can see their proposals reflected.

Acknowledgments

This research was funded by a grant from the Netherlands Organisation for Scientific Research (NWO), Division of Humanities, to Levelt, ten Cate and Zuidema (360-70-450).

References

Gerry T.M. Altmann. Learning and development in neural networks–the importance of prior experience. *Cognition*, 85(2):B43–B50, 2002.

Gerry T.M. Altmann and Zoltán Dienes. Technical comment on rule learning by seven-month-old infants and neural networks. *Science*, 284(5416):875–875, 1999.

Catherine E. Carr and Masakazu Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Sciences*, 85(21):8311–8315, 1988.

R. Chaudhuri and I. Fiete. Computational principles of memory. *Nature neuroscience*, 19(3):394–403, 2016.

Morten H. Christiansen and Suzanne Curtin. Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Sciences*, 3(8):289 – 290, 1999.

Morten H. Christiansen, C.M. Conway, and Susan Curtin. A connectionist single mechanism account of rule-like behavior in infancy. In *Proceedings of the 22nd annual conference of the cognitive science society*, pages 83–88, 2000.

Zoltán Dienes, Gerry Altmann, and Shi-Ji Gao. Mapping across domains without feedback: A neural network model of transfer of implicit knowledge. *Cognitive Science*, 23(1):53–82, 1999.

P. Eimas. Do infants learn grammar with algebra or statistics? *Science*, 284(5413):435, 1999.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Stefan L. Frank and M Čerňanskỳ. Generalization and systematicity in echo state networks. In & V.M. Sloutsky B.C. Love, K. McRae, editor, *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 733–738. Cognitive Science Society, 2008.

LouAnn Gerken. Decisions, decisions: infant language learning when multiple generalizations are possible. *Cognition*, 98(3):B67 – B74, 2006.

Herbert Jaeger. The ''echo state'' approach to analysing and training recurrent neural networks. Technical report, German National Research Center for Information Technology, 2001.

Herbert Jaeger. Short term memory in echo state networks. Technical report, German National Research Center for Information Technology, 2002.

Lloyd A. Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.

G. F. Marcus, S. Vijayan, S.B. Rao, and P.M. Vishton. Rule learning by seven-month-old infants. *Science*, 283 (5398):77–80, 1999.

Gary F. Marcus. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3):243–282, 1998.

Gary F. Marcus. Connectionism: with or without rules?: Response to J.L. McClelland and D.C. Plaut (1999). *Trends in Cognitive Sciences*, 3(5):168 – 170, 1999a.

Gary F. Marcus. Do infants learn grammar with algebra or statistics? Response to Seidenberg and Elman, Negishi and Eimas. *Science*, 284:436–37, 1999b.

Gary F. Marcus. Reply to Christiansen and Curtin. *Trends in Cognitive Sciences*, 3(8):290 – 291, 1999c.

Gary F. Marcus. Reply to Seidenberg and Elman. *Trends in Cognitive Sciences*, 3(8):288, 1999d.

Gary F. Marcus. Response to Technical Comment on Rule learning by seven-month-old infants and neural networks by gerry t.m. altmann and Zoltán Dienes. *Science*, 284(5416):875–876, 1999e.

J. L. McClelland and David C. Plaut. Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3(5):166–168, 1999.

Steven Pinker. *Words and rules: The ingredients of language*. Basic Books, 2015.

Steven Pinker and Alan Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193, 1988.

Kim Plunkett and Virginia Marchman. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48(1):21–69, 1993.

D.E. Rumelhart and J.L. McClelland. On learning past tenses of English verbs. In D.E. Rumelhart and J.L. McClelland, editors, *Parallel Distributed Processing, Vol. 2*, pages 318–362. MIT Press, Cambridge, MA, 1986.

Mark S. Seidenberg and Jeffrey L. Elman. Do infants learn grammar with algebra or statistics? *Science*, 284 (5413):433, 1999a.

Mark S. Seidenberg and Jeffrey L. Elman. Networks are not 'hidden rules'. *Trends in Cognitive Sciences*, 3(8): 288–289, 1999b.

Thomas R. Shultz. Rule learning by habituation can be simulated in neural networks. In *Proceedings of the twenty first annual conference of the Cognitive Science Society*, pages 665–670, 1999.

Thomas R. Shultz and Alan C. Bale. Neural network simulation of infant familiarization to artificial sentences: Rule-like behavior without explicit rules and variables. *Infancy*, 2(4):501–536, 2001.

Sylvian Sirois, David Buckingham, and Thomas R. Shultz. Artificial grammer learning by infants: an auto-associator perspective. *Developmental Science*, 3(4):442–456, 2000.