The Statistical Brain: Reply to Marcus' *The Algebraic Mind*

Francisco Calvo (fjcalvo@um.es)
Cognitive Science Programme, East Third Street
Bloomington, IN 47405 USA, and
Departamento de Filosofía, Campus de Espinardo
Murcia, 30100 SPAIN

Eliana Colunga (colunga@psych.colorado.edu)
Department of Psychology, 345 UCB
Boulder, CO 80309-0345 USA

Abstract

Marcus (2001) argues that only those connectionist models that incorporate (classical) rules can account for the phenomenon of transfer of learning in infants. Seidenberg and Elman (1999) have tried to counter to Marcus by means of a simple recurrent network (SRN) trained on a *categorization* task. In this paper we show how a *prediction*-SRN, trained on a simple but structured pre-training set, can preserve its computational equivalency with respect to classical counterparts while eschewing the need to posit rule-governed underlying mechanisms; a criticism that has been raised against Seidenberg and Elman's categorization-based reply.

Introduction

Marcus (2001) distinguishes two separate ontologies in the connectionist realm: implementational connectionism and eliminative connectionism. The former accounts for cognitive phenomena by positing sets of explicit rules that serve the purpose of symbolic manipulation. The latter, in terms of computational abilities which are the result of an associative memory. Marcus argues that the connectionist models which preserve their computational equivalency with respect to classical ones are those that *implement* classical rules.

Transfer of learning in infants

Marcus (2001) assesses the relationship between connectionist theory and rule-governed behaviour by challenging the connectionist to account for data collected in a number of experiments with infants. In one well-known experiment, Marcus et al. (1999) habituated 7-month-old infants to strings of artificial syllables that belonged to an ABA or an ABB abstract grammar—e.g., "le di le" or "wi je je". As Marcus et al. report, infants listen longer to novel sequences that do not conform to the pattern they've been exposed to during habituation (e.g., "ba po po" for infants habituated to ABA sequences, and "ba po ba" for those habituated on ABB ones). The data has been interpreted as showing that infants exploit (rule-governed) abstract knowledge in order to induce the implicit grammar common to different sequences of syllables: "infants extract abstract-like rules that represent relationships between placeholders (variables) such as 'the first item X is the same as the third item Y' or more generally that 'item I is the same as item J'." (Marcus et al., 1999).

Seidenberg and Elman's simulation

Seidenberg and Elman (1999a) tried to account for Marcus et al.'s data in purely statistical terms, while avoiding implementing a classical architecture in doing so. Their strategy was twofold: They first *pre*-trained an SRN (Elman, 1990; fig. 1) on a categorization task. The network had to output 1 or 0, depending on whether the syllable being fed at a given point was a token of the same type as the syllable fed at the previous time step. Once the weights were frozen, they encoded information about the same/different relationship that holds for a large set of syllables that infants surely have already encountered prior to the experiment. The pre-trained SRN can then exploit that knowledge in a subsequent habituation phase where it is exposed to strings of syllables *similar* to those infants had become familiar with in Marcus et al.'s experiment. The network's task in this second phase is to categorize strings of syllables in the ABA or ABB grammatical subsets by outputting a 0 or a 1, respectively.[1]

[1] No training took place after presentation of the first two syllables during habituation, nor on the pre-training output unit (see Seidenberg and Elman, 1999a, for the details).

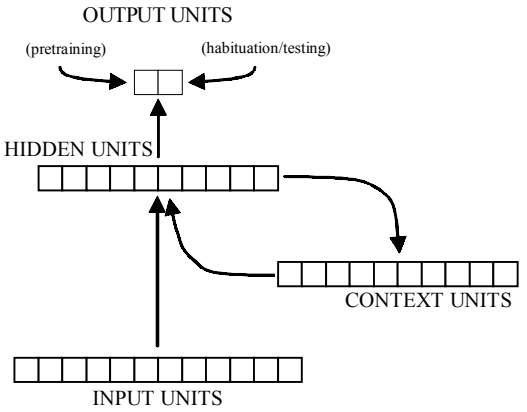

Figure 1: Seidenberg and Elman's SRN (output units for pre-training, and habituation and testing, are different.

The network was then tested on the same ABA and ABB strings that infants had been tested on in the Marcus et al. experiment. Seidenberg and Elman's results show that the network can generalize the knowledge acquired in the habituation phase to novel stimuli—although see Vilcu & Hadley (2001) for a skeptical appraisal of their data. According to Seidenberg and Elman, this supports the view that the infants' behaviour can be accounted for without the positing of an abstract grammar that is somehow being tacitly followed.

Marcus, however, is not moved by Seidenberg and Elman's simulation. Although their model neither implements a classical architecture, nor makes use of symbolic structures directly, the external teaching signal that it requires to backpropagate error so as to adjust the weight matrix in *the categorization task* of the pre-training phase implements a rule: "It incorporates a universally open-ended rule of the sort, *for all syllables x, y, if x=y then output 1 else output 0*." (Marcus, 1999)

Generalization in *prediction*-based SRNs

It is not clear that Marcus' comments are a real threat to connectionism (see Seidenberg and Elman, 1999b, for a rebuttal), but granting, for simplicity's sake, Marcus' charge of implementation, in what follows we report the results of a simulation that is empirically adequate in the face of the infants' data *and* does not implement classicism.

Stimuli

Pretraining corpus

Following Seidenberg and Elman (1999a), we simulated infants' linguistic environment, prior to taking part in the experiment, by exposing a SRN to a pool of syllables (table 1). The corpus consisted of CV syllables formed by concatenating all possible combinations of consonants and vowels in the corpus (resulting in 85 syllable types). CV syllables were encoded in a semi-localist way by concatenating consonants and vowels represented locally (table 3).[2]

Table 1: Stimuli (pre-training phase).

va	ve	vi	vo	vu
pa	pe	pi	po	pu
da	de	di	do	du
ya	ye	yi	yo	yu
ga	ge	gi	go	gu
ka	ke	ki	ko	ku
ba	be	bi	bo	bu
wa	we	wi	wo	wu
ma	me	mi	mo	mu
na	ne	ni	no	nu
za	ze	zi	zo	zu
sa	se	si	so	su
fa	fe	fi	fo	fu
la	le	li	lo	lu
ra	re	ri	ro	ru
ta	te	ti	to	tu
ha	he	hi	ho	hu

Table 2: Habituation stimuli.

le	di	di
je	le	le
li	le	le
we	le	le
wi	di	di
je	wi	wi
li	we	we
we	wi	wi
di	ji	ji
je	ji	ji
ji	li	li
we	ji	ji
de	di	di
je	de	de
li	de	de
de	we	we

[2] Network performance using a distributed version of Plunkett & Marchman (1993) phonetic coding didn't yield significantly different outcomes.

Habituation corpus

The habituation corpus consisted of the same 12 ABB or ABA strings of syllables used by Marcus et al. (table 2). The same semi-localist input encoding was used as in the pre-training corpus.

Table 3: Habituation stimuli coding.

de	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
yi	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0
li	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
we	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
di	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
ye	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0
le	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
wi	0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0

Network architecture

Based on the architecture of Seidenberg and Elman's model, we trained a SRN to test if it could generalize to novel strings of syllables in the line of Marcus et al.'s experiment. The network had 31 input and output (pre-training and habituation) units, and 41 units in both the hidden and context layers (figure 2).

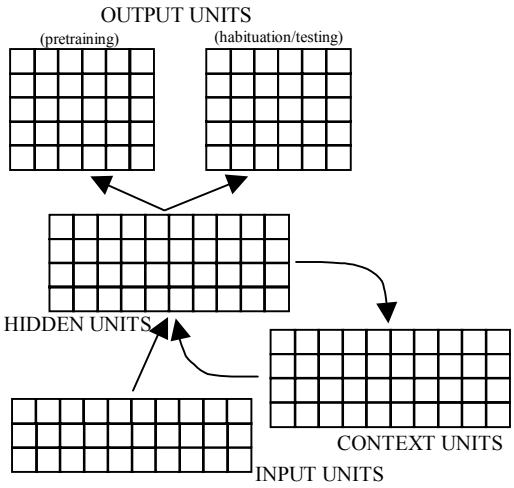

Figure 2: A SRN trained on a prediction task (output units for (i) pre-training, and (ii) habituation and testing, are in different banks).

During pre-training, on the one hand, and during habituation and testing, on the other, different banks of output units were deployed. (No training took place on the habituation bank during pre-training or *vice versa*).

Task

Pre-training

In the pre-training phase we fed the network with 50,000 syllable tokens from the set of 85 syllable types. The task was to predict the next item in the sequence. There were no (complex) "grammatical" constraints such as those in Elman's (1990) classical prediction task. To make the task learnable, we varied the amount of syllables that were duplicated. 5 corpora were created where syllable duplication ranged from 0% to 100%. In this way, a network exposed to a 0%-duplication corpus would hardly reduce its overall error since random dependencies in such a (noisy) data set would cancel each other out. At the other end of the spectrum, a network pre-trained on a 100%-duplication corpus would decrease its prediction error significantly since it would learn that every syllable in the corpus is followed by an identical token. Intermediate corpora (25%, 50%, and 75% repetition) yielded error scores in between (figure 3).

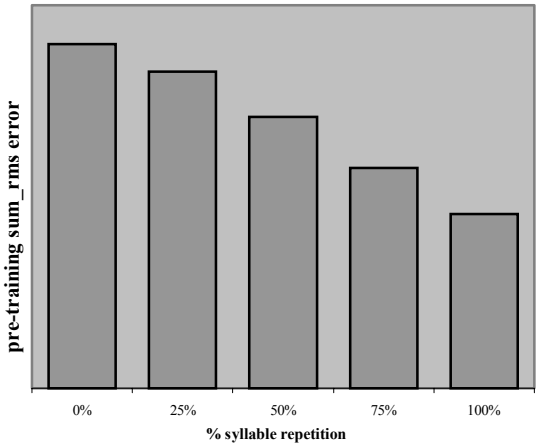

Figure 3: Error scores in trained networks decreased as the structure in the pre-training set (percentage of duplication) increased.

Habituation

With the weights from the pre-training phase frozen, networks were trained on either the ABB or the ABA habituation corpus of table 2. Training was stopped at 4 different points in learning—whenever the sum of the root-mean-

square error (sum_rms) for the habituation output bank of units fell below 1,2; 1,1; 1,0; and 0,9.[3]

Generalization

With the weights frozen from the habituation phase, we tested performance on novel data. To confirm the robustness of our results, we used an extended corpus of 1,000 syllables forming a sequence of ABB and ABA novel strings. The same syllables were used for both grammatical patterns.

20 tests were done for the 20 different weight matrices obtained by habituating the SRN on the 0,9 to 1,2 sum_rms measure stops for all pre-training corpora (0% to 100% duplication). For each trial in the generalization corpus, the prediction error was recorded for next item presented after the presentation of the first two syllables. Differences in prediction error between ABA's and ABB's third syllable were computed. The prediction is that networks habituated to ABB patterns will have an advantage when predicting ABB patterns over when predicting ABA patterns, but networks habituated to ABA patterns will show an advantage predicting ABA patterns over ABB patterns.

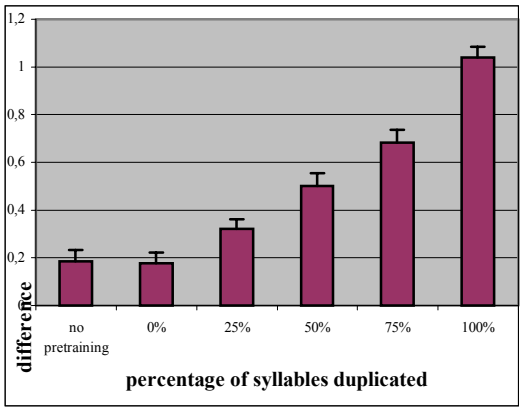

Figure 4: difference between congruent vs. incongruent patterns during generalization after habituation.

Results

To see the role pre-training played in the overall task, we ran 10 networks in each of five Pre-training conditions: no pre-training, pre-training with no duplication (0%), 25% of syllables dupplicated, 50%, 75% and 100%. Half of these networks were trained on ABB patterns during the habituation phase and the other half were trained on ABA patterns during the habituation phase. Figure 4 shows the relative facilitation for predicting congruent over incongruent patterns. To calculate this, we used the average of the incongruent minus the congruent sum_rms for networks pre-trained on 0% to 100%-duplication corpora, and networks that were not pre-trained previously.

If the networks have abstracted the general patterns of the grammars present during habituation, the difference of error in incongruent minus congruent patterns should be positive. Additionally, if the kind of pre-training we are using is in fact necessary for modeling this phenomenon, we expect to see an advantage of congruent patterns in networks trained with some repetition, but not in those trained without repetition or those with no pre-training.

These data were submitted to a 6(Pre-training) x 2(Habituation) ANOVA. The results show a significant main effect of Pre-training ($p<.01$). The congruency effect is bigger for networks that have more duplication in their pre-training. Post-hoc tests revealed no difference between networks that received no pre-training and networks that were pre-trained in an unstructured corpus that had no consistent ($p=.28$). There was also a significant main effect of Habituation ($p<.01$). Post-hoc analysis revealed that the congruency effect was bigger for networks habituated to ABB patterns than for networks habituated to ABA patterns ($p<.01$).

The analysis also revealed a significant interaction between Habituation and Pre-training ($p<.01$) – the effect of pre-training was greater for networks habituated to ABB patterns than for networks habituated for ABA patterns.

Discussion

The results of our simulations show that simple SRN networks, when trained on a simple but structured corpus, will generalize the abstract patterns embodied in their training set and gain an advantage in processing subsequent patterns of the same type. That is, like the infants in Marcus et al's study, the networks that were pre-trained in a corpus in which some syllables were consistently duplicated, learned to distinguish ABB patterns from ABA patterns after a brief period of training akin to infant's habituation.

There is crucial difference between Seidenberg and Elman's model and ours. Their network was trained both in the pre-training and the habituation

[3] The simulations were run with PDP++ (O'Reilly, Dawson, and McClelland), and trained with a learning rate of 0,05 during pre-training, and of 0,01 during habituation. Although calculating error measures of probability-based predictions against likelihood vectors would have been more informative, for current purposes (i.e., assessment of the role of pre-training—see discussion, below) sum_rms values suffice.

phases on a categorization task. Although, that is perfectly legitimate, there is a striking difference between the habituation phase in their simulation and *the way infants were familiarized* with their linguistic environment in the Marcus et al.'s original experiment. Whereas infants were exposed to strings of syllables that belonged to a *single* grammar (e.g., "le di di" and "wi je je"—ABB), Seidenberg and Elman's network was habituated to two sets of strings that conformed to two *different* grammars (e.g., "le di di"—ABB—"wi je wi"—ABA). The task then consisted in the correct categorization of exemplars into the ABB and ABA categories. Our network, on the contrary, is trained in both phases (pre-training and habituation) on a prediction task. Importantly, the model, like the infants of the Marcus et al. experiment, is exclusively exposed to positive examples of ABB strings of syllables (i.e., all generated by a single grammar).

The fundamental component of our simulation is the nature of the pre-training environment. The pre-training environment does not merely consist of a large set of syllables where the input signal is generated by the random concatenation of exemplars in the data set, but, crucially, some syllable tokens are normally encountered followed by other tokens that belong to the same representational type. So, for example, in a (reduced) ecological context, such as the one infants encounter in their first months of life, "ma" and "pa" are very frequently followed by another "ma" and "pa" exemplar, respectively, whereas some other syllables (e.g., "the") may be followed by almost any syllable in the corpus. In these simulations we abstract this distinction to its two extremes and present the network during pre-training with syllables that are either always duplicated or never duplicated. These first-order correlations in the environment amount to sub-regularities that can be exploited by the network in a semi-deterministic prediction task.[4] This is how a SRN can be pre-trained on a prediction task, bypassing some of the problems faced by Seidenberg and Elman's categorization-based solution.

So what do the networks learn from pre-training that helps them abstract the regularities in the habituation task? One possibility is that networks need to develop consistent and coherent representations for the syllables in the corpus (the words in Marcus' habituation) in order to more effectively encode the patterns seen during the short habituation phase. This possibility is weakened, however, by the fact that networks pre-trained in a corpus made out of syllables but containing no additional structure in terms of duplication, do not perform better than the networks that receive no pre-training at all. A second possibility is that during pre-training the networks learn to represent something general about duplication, in other words, *sameness*. This abstraction could be crucial in encoding the patterns during the habituation phase. This could explain the apparent advantage of networks in dealing with ABB patterns compared to ABA patterns – duplication is a part of ABB and could be used to encode it more efficiently. This advantage of ABB patterns, then, may disappear if the networks are trained on a corpus containing longer distance dependencies.

Interestingly, the networks predict differences between ABA and ABB patterns that may fall from the nature of short-term vs. long-term memory. Networks learn ABB patterns faster than ABA patterns and habituate faster to ABB than ABA patterns. We are currently working on an extended model that differentiates short-term (habituation) and long-term (pre-training) memory and concurrently testing these predictions in 7-month-old infants. By combining these simulations with empirical testing of their predictions we hope to shed light on the mechanisms involved in this phenomenon.

Conclusion

The research reported here shows how Marcus' challenge can be met while avoiding the positing of rule-fitting patterns of behavior (allegedly required to constrain novel data). In our view, Marcus correctly points out that "Seidenberg and Elman do not give an account of how the supervisor's rule could itself be implemented in the neural substrate" (2001, p. 65). The teaching signal of a fully supervised categorization task is not ecologically grounded. In our self-supervised prediction task, that's not an issue. Activation-based signals in a prediction task are not to be interpreted in terms of rule-implementation. The teaching signal exploited *is* an activation state of experience. As McClelland emphasizes with regard to the ecological plausibility of this type of signals: "the brain is constantly generating expectations and the discrepancies between these expectations and subsequent outcomes can be

[4] The assumption that syllables are not concatenated in a purely random fashion is empirically justified. Notice that infants start to babble at the age of 6-months. The linguistic stream they encounter as input is thus made up of the combination of their natural environment (with the existing first-order correlations among syllable sets) plus the babbling repetitions uttered by themselves.

used for error-driven learning" (from O'Reilly and Munakata, 2000). The network encodes the existing statistical regularities with no need to process algebra-like information. Our working hypothesis is that infants make use of discrepancies based on expectations to make successful predictions.

Marcus (2001) claimed that connectionist networks would lose their empirical adequacy unless they implemented a classical architecture. Marcus' infant data is accounted for in statistical terms— without the positing of devices that store particular values of variables to perform variable bindings, such as register sets, or other classical resources. The charge of implementation is therefore not applicable to our results, since the ecologically grounded prediction task of the networks does not incorporate universally open-ended rules.

Acknowledgements

This work was supported in part by a *Ramón y Cajal* research contract to the first author (Spanish Ministry of Science and Technology). This material draws out of preliminary work presented at the 6th Conference of the Association for the Scientific Study of Consciousness (ASSC'6-2002) in Barcelona, Spain.

References

Elman, J.L. (1990). Finding structure in time. *Cognition*, *14*, 179-211.

Marcus, G. (1999). Reply to Seidenberg and Elman. *Trends in Cognitive Sciences*, *3*, 289.

Marcus, G.F. (2001). *The Algebraic Mind*. Cambridge, Mass.: MIT Press.

Marcus, G.F., Vijayan, S., Rao, S.B. & Vishton, P.M. (1999). Rule learning in seven-month-old infants. *Science*, *283*, 77-80.

O'Reilly, R. & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, Mass.: MIT Press.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 21-69.

Seidenberg, M.S. and Elman, J.L. (1999a). Generalization, rules, and neural networks: A simulation of Marcus et al. http://crl.ucsd.edu/~elman/Papers/MVRVsimulation.html.

Seidenberg, M.S. and Elman, J.L. (1999b). Networks are not 'hidden rules.' *Trends in Cognitive Science*, *3*, 288-289.

Vilcu, M. and Hadley, R.F. (2001). Generalization In Simple Recurrent Networks. *Proceedings to the 23rd Annual Conference of the Cognitive Science Society* (pp. 1100-1105). Hillsdale, NJ: Lawrence Erlbaum Associates.