



Discussion

Language acquisition in the absence of explicit negative evidence: can simple recurrent networks obviate the need for domain-specific learning devices?

Gary F. Marcus*

Department of Psychology, New York University, New York, NY 10012, USA

Rohde & Plaut, 1999 argue that their work with Elman's simple recurrent network (henceforth, SRN) "suggests that learning the structure of natural language may be possible despite a lack of explicit negative feedback ... in the absence of detailed innate language-acquisition mechanisms". They further argue that "a key factor in overcoming the 'logical problem' of language acquisition (Baker&McCarthy, 1981) is the use of implicit negative evidence." (Implicit negative evidence is information about something that does not appear when it was predicted to appear.)

R&P are surely correct that some versions of the simple recurrent network do not rely on negative evidence and that such networks are able in some cases to utilize implicit negative evidence.¹ But R&P do not show that these models avoid the kinds of errors that children make, do not show that these models derive the same generalizations as children do, and do not show that these models use indirect negative evidence in ways that would obviate the need for innate, domain-specific learning devices. All that they offer is a simulation of a tiny fragment of a simplified version of English; they do not fit the model's data against any data derived from children. Their system does not provide any sort of syntactic or semantic representation of the sentences that it is exposed to, and it does not make a principled distinction between infrequent and ungrammatical sentences. This is not enough to establish the adequacy of the model, and more careful inspection reveals a serious, principled limitation that stems directly from its treatment of implicit negative evidence.

* Tel.: +1-212-998-7874; fax: +1-212-995-4018.

E-mail address: gary.marcus@nyu.edu (G.F. Marcus)

¹ Implicit in R&P's argument is the idea that nativism rests on the lack of negative evidence. But the question about whether negative evidence is available is in fact independent of nativism. As I put it in Marcus (1993), "Even if perfect negative evidence were available, innate constraints on the generalizations which children make would be necessary because many plausible errors simply never occur."

Implicit negative evidence is only useful if a learner knows which non-appearances matter, and which do not. Simple recurrent networks are unable to make this distinction, instead often taking evidence about non-occurrence to imply the “ungrammaticality” of grammatical utterances.² For example, imagine watching a documentary about a new sport called *daxing*, in which the object was to *fleedle* your opponent. If you heard over and over that *Smith fleedled Jones*, you would infer that it was also grammatical to say that *Smith fleedled Belanger* or *Smith fleedled Ripken*.

But the simple recurrent network does the opposite, because it takes each occurrence of *fleedle* without Ripken to indicate that *Smith fleedled Ripken* is ungrammatical. Every time the model predicts ‘Ripken’ as a possible continuation to *fleedle* and instead some other item (e.g. Jones) appears as a continuation, the model tends to change the weights in such a way that if same sentence fragment were again presented, that the model would activate ‘Ripken’ less strongly. Thus the more often the model is exposed to *Smith fleedled Jones* the more it takes implicit negative evidence to rule out the grammaticality of *Smith fleedled Ripken*.

To take another example, my colleagues and I (Marcus, Vijayan, Rao, & Vishton, 1999) recently discovered that infants who were habituated to a series of sentences in artificial language with an “ABA” structure (such as *ga ti ga* and *li na li*) would attend longer to novel sentences that had a different grammar (e.g. *wo fe fe*) than novel sentences that had the same grammar (e.g. *wo fe wo*). To capture this in the prediction framework adopted by Rohde and Plaut, the SRN would have to predict (say) *wo* as the continuation to the sentence fragment *wo fe* _____. But in a series of simulations that I describe at <http://psych.nyu.edu/gary/science/es.html>, I found that a model like Plaut & Rohde’s could not capture the behavior of the infants. As in the *fleedle* case, the model was undermined by its treatment of indirect negative evidence. Whenever *wo* did not appear, the network adjusted the connection weights in a way that made it less likely that the model would predict *wo* as a continuation – each time, it takes the absence of *wo* as a bit of (misleading) implicit negative evidence. Hearing a sentence like *ga ti ga* should make a learner more sure that the right continuation to *wo fe* _____ is *wo*, not less.

One can modify the simple recurrent network such that it capture our data, but only by changing it in ways that cause it to embody a different hypothesis about language learning. For example, Seidenberg and Elman (1999a) were able to get an adaptation of the SRN to learn our ABA grammar – but their model depended on being part of a larger system that was provided with negative evidence. (It was given a target of 1 if a habituation sentence was an ABA sentence, and a target of zero

² Strictly speaking, simple recurrent networks do not make a distinction between grammatical and ungrammatical at all. The only output that they produce is a set of probabilities which estimate the likelihoods of particular words appearing as continuations to a given sentence fragment that it is exposed to. Following customary practice, I will suppose for the sake of argument that the extent to which model takes a given input sentence to be grammatical is a function of how strongly it activates a given output node given a particular continuation. If a given node is strongly activated, the model takes that continuation to be grammatical; if the output node in question is weakly activated, the model takes that continuation to be ungrammatical.

otherwise.) This model shows how an infant who was endowed with a simple recurrent network and supplied with reliable negative evidence could have succeeded in our task. Such model is inappropriate, however, because infants in our experiment, like children in general, were not in fact given negative evidence. (For further discussion of our infant results and how they might be implemented, see Altmann & Dienes, 1999; Christiansen & Curtin, 1999a,b; Dominey & Franck, 1999; Marcus, 1999a,b,c,d,e; McClelland & Plaut, 1999; Negishi, 1999; Seidenberg & Elman, 1999a,b; Shastri, 1999; Shultz, 1999).

Trying to learn everything with one simple “general purpose” learning model is asking too much. It is mistake in the first place, of course, to think that there is any such thing as a single general purpose learning device. As Rohde and Plaut put it, all learning systems ‘have constraints built into them’. The empirical evidence reviewed in this brief reply suggests that the constraints built into the simple recurrent network simply are not the right ones. This does not guarantee that the right constraints are language-specific, but nor does the SRN provide any convincing evidence to the contrary. Physics has the first law of thermodynamics (you can’t win) and the second law (you can’t break even); learning has Wolpert’s, 1996 law of induction: there’s no free lunch. As authors such as Quine (1960) and Goodman (1955) have stressed, problems of induction are always underdetermined. Implicit negative evidence can be helpful, but only for a learner that builds in the right sorts of constraints; systems like the simple recurrent network build in the wrong sorts of constraints and wind up ruling out many grammatical sentences simply because they have not been heard. What makes child language learners so special is that they are not so easily led astray.

Acknowledgements

I thank Steve Pinker for helpful comments.

References

- Altmann, G.T.M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science*, 284, 875a. Available only on-line, at <http://www.sciencemag.org/cgi/content/full/284/5416/875a>.
- Baker, C. L., & McCarthy, J. J. (1981). *The logical problem of language acquisition*, MIT Press: Cambridge, MA.
- Christiansen, M.H., & Curtin, S.L. (1999). The power of statistical learning: No need for algebraic rules, *Proceedings of the Twenty first Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum. Available on-line at <http://www.siu.edu/~psycho/faculty/morten/statlearn.html>.
- Christiansen, M. H., & Curtin, S. L. (1999). Transfer of Learning: Rule acquisition or statistical learning? *Trends in Cognitive Sciences*, 3, 289–290.
- Dominey, P.F., & Franck, R. (1999). Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure in the infant. Submitted.
- Goodman, N. (1955). *Fact, fiction, and forecast*, Indianapolis: Bobbs-Merrill.
- Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning in 7-month-old infants. *Science*, 283, 77–80.

- Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, 46, 53–85.
- Marcus, G. F. (1999a). Connectionism: with or without rules? Response to J.L. McClelland and D.C. Plaut. *Trends in Cognitive Sciences*, 3, 168–170.
- Marcus, G. F. (1999b). Do infants learn grammar with algebra or statistics? Response to Seidenberg & Elman, Eimas, and Negishi. *Science*, 284, 436–437.
- Marcus, G. F. (1999c). Reply to Christiansen & Curtin. *Trends in Cognitive Sciences*, 3, 290–291.
- Marcus, G. F. (1999d). Reply to Seidenberg & Elman. *Trends in Cognitive Sciences*, 3, 289.
- Marcus G.F. (1999e). Rule learning by seven-month-old infants and neural networks: response to Altmann and Dienes. *Science*, 284, 875a. Available only on-line, at <http://www.sciencemag.org/cgi/content/full/284/5416/875a>.
- McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences*, 3, 166–168.
- Negishi, M. (1999). Do infants learn grammar with algebra or statistics? *Science*, 284, 435.
- Quine, W. V. O. (1960). *Word and object*, Cambridge, MA: MIT Press.
- Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence! How important is starting small? *Cognition* 72, 67–109.
- Seidenberg, M. S., & Elman, J. L. (1999a). Do infants learn grammar with algebra or statistics? *Science*, 284, 435–436.
- Seidenberg, M. S., & Elman, J. L. (1999b). Networks are not ‘hidden rules’. *Trends in Cognitive Sciences*, 3, 288–289.
- Shastri, L. (1999). Infants learning algebraic rules. *Science* 285, 673.
- Shultz, T. R. (1999). *Rule learning by habituation can be simulated in neural networks. Proceedings of the Twenty first Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.
- Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8, 1341–1390.