Rules versus Statistics in Biconditional Grammar Learning: A Simulation based on Shanks et al. (1997)

Bert Timmermans (Bert.Timmermans@vub.ac.be)
Dienst Persoonlijkheids- en Sociale Psychologie
Vrije Universiteit Brussel
Pleinlaan 2
1050 Brussels – Belgium

Axel Cleeremans (Axel.Cleeremans@ulb.ac.be)
Séminaire de Recherche en Sciences Cognitives
Université Libre de Bruxelles,
Avenue F.-D. Roosevelt, 50 – CP 122
1050 Brussels – Belgium

Abstract

A significant part of everyday learning occurs incidentally — a process typically described as implicit learning. A central issue in this and germane domains such as language acquisition is the extent to which performance depends on the acquisition and deployment of abstract rules. In an attempt to address this question, we show that the apparent use of such rules in a simple categorisation task of artificial grammar strings, as reported by Shanks, Johnstone, and Staggs (1997), can be simulated by means of a simple recurrent network, and may thus turn out not be incompatible with the acquisition of statistical regularities rooted in the processing of exemplars of the presented material.

Introduction

Over development and learning, we acquire a considerable amount of information incidentally. Natural language offers perhaps the most striking example of such incidental learning: Infants do not need to be explained grammar rules in order to be able to communicate effectively and are presumably unaware of the fact that they are learning something at all. Adult speakers likewise “know” whether expressions of their native language are grammatically correct but can seldom explain why.

Implicit Learning

The notion of *"implicit learning"* (IL) usually designates cases in which a person learns about the structure of a fairly complex stimulus environment, without necessarily intending to do so, and in such a way that the resulting knowledge is difficult to express (Berry and Dienes, 1993). In short, IL is the ability to learn without awareness (Cleeremans, Destrebecqz, and Boyer, 1998), as opposed to explicit learning, which is strategy- and/or hypothesis-driven, and of which one tends to be consciously aware.

IL can produce *implicit knowledge*. According to Cleeremans (1997), "at a given time, knowledge is implicit when it can influence processing without possessing in and of itself the properties that would enable it to be an object of representation, and implicit learning is the process by which we acquire such knowledge." (p.199) As for the notion of "representation", we agree with Perruchet and Vinter (submitted), who state that a representation has to *represent* an entity in the real world and has to be in and of itself manipulable *as* that entity (Perruchet and Vinter talk about its "function within a causal system"). Therefore, an entity that is an *object of representation* has to exist independently from the "hardware" of the system by which it is represented, making it available for information-processing operations in a variety of contexts (Cleeremans, 1997) — such as a rule that is applicable to different instances of a certain problem.

Inherent to this issue is the question of whether the mechanisms through which implicit and explicit knowledge are acquired are best viewed as being subtended by separate processing systems. This is exactly what has been suggested by Shanks and colleagues (Shanks and St John, 1994; Shanks, Johnstone, and Staggs, 1997; St John and Shanks, 1997), who proposed to abandon the distinction between Implicit and Explicit Learning in terms of conscious awareness being present or not, and instead suggested that the distinction is one of rule-based versus memory-based learning processes. Before going any deeper into this matter, let us consider two different ways of looking at learning in general, to illustrate how they can possibly account for Implicit Learning.

Computational Modelling of IL

Two views come forth when considering the mind in general, and implicit learning in particular: the symbolic and the connectionist approach. Each has its own view on how knowledge is represented and how it might be manipulated. The symbolic metaphor is usually associated with rule-based learning, while the connectionist metaphor is associated with memory-based learning based on the statistical characteristics of the stimuli.

The Symbolic Metaphor. Cleeremans & Jiménez (submitted) point out that a symbol system leaves no room for a concept like IL. In a symbol system, expressions that are formed are static representations of (real-world) entities or relations, stored in the system's memory. These symbols, be it of objects or of rules, have to be interpreted by something — a processor — when they are to be used by the system to augment its knowledge base (memory), that is, to learn. From this perspective, IL can only exist if one assumes the existence of a *cognitive unconscious*, i.e. a subset of the mind that can basically process all the information that the conscious system can process, only minus consciousness. Consequently, consciousness is purely epiphenomenal in this framework. It is exactly the fact that all symbols have in and of themselves the property of being an accessible representation, independent of the processor, which makes them unsuitable as a metaphor for implicit knowledge. For it is impossible to conceive of any knowledge that could influence processing while remaining unavailable to outside inspection. Importantly, this perspective also makes it possible to assume the existence of *abstract* knowledge that remains inaccessible to conscious inspection.

The Connectionist Metaphor. By constrast, in a connectionist network, there is no external processor engaged in learning, that is, learning does not consist of augmenting a distinct knowledge base. Instead, learning in a connectionist network is the *result* of changes that occur in the network (weight-change between units). These changes are themselves *caused by* information processing, i.e. the coupling of a certain input with a desired output Thus, this processing also changes the process of learning (through for example back-propagation of the error between the actual and the desired output). Furthermore, as transient knowledge in a connectionist network consists of activation patterns, instead of symbols, a piece of knowledge does not have to be "interpreted" by the central processor before it can influence processing. These properties make it possible for a connectionist network to possess knowledge that can influence behaviour despite failing to be represented as such. It makes it possible to consider implicitness as something more than simply a property of the database or a property of the processor.

From the connectionist point of view, subjects are said to base their judgements on the basis of exemplar information, without explicitly extracting abstract generalities, or rules – the abstract processing is performed online during the test, when necessary. The episodic account provides a refined version of mere instance-based processing (e.g. Neal & Hesketh, 1997). One of the most popular instances of traiditional connectionist networks is the Simple Recurrent Network (SRN), as proposed by Elman (1990). Here, judgements are no longer based on instances, but on instances *within their context*. Learning is nothing more than a byproduct of the processing itself (weight-change), while retrieval results from the overlap between processes operating during study- and test-phases. Several variations on this basic principle have been proposed, but the main point remains as stated: no abstract rules in implicit learning. Instead, more fragmentary knowledge is used to gradually and dynamically build up representations op the stimulus environment. This leaves room for implicitness, not in the way of equating the existence of representations with accesibility to consciousness (as do for example Perruchet and Vinter, submitted; O'Brien and Opie, 1999), but in virtue of the dynamical aspects of representation building. For example, it might be possible to conceive of conscious representations as being structured differently than unconscious ones, or as being of lower quality.

Experimental Research on IL

Recently, some of the processes involved in word segmentation have been described as rooted in the same mechanisms as implicit learning and frequency estimation. For instance, Saffran et al. (1997) conducted an experiment on word segmentation in artificial speech. They exposed children (6-7 years old) and adult subjects to a continuous speech flow such as *bupadapatubitutibudutabapidabu*. Subjects were told that the experiment was about the influence of auditory stimuli on creativity (to make sure learning was incidental and not intentional). The only cues to word boundaries were the transitional probabilities between pairs of syllables (e.g., *bu-pa*), which were higher within words than between words. Afterwards, subjects heard two sets of sounds, each consisting of three syllable pairs, and were told to decide which one sounded more like the tape they had heard. Both adult and child subjects managed to perform well above chance, suggesting that learning might proceed in the absence of attention and the intention to do so, even despite the brevity of the exposure (one ore two times a 21' tape). The fact that children did as well as adults suggests a robust phenomenon that might play a role in natural language acquisition.

In another interesting artificial language experiment, Marcus et al. (1999) claim to have showed that 7-month-old infants can "represent, extract, and generalise abstract algebraic rules." In short, the infants were exposed to artificial "sentences" during a training phase, and subsequently were presented with a few test items, some of them belonging to the same language, while others introduced some structural novelty. For example, when an infant had been habituated to *gatiti* or *linana* (both having an ABB structure), it was subsequently presented with test sentences such as *wofefe* or *wofewo* (the last one being of ABA structure). The basic set-up is similar to the Saffran et al. (1997) experiment, with the important difference that there where the Saffran et al. test items were composed of the same material as the training items, Marcus et al. introduced a change in the sensory content of the material. That is, prior to hearing the above illustrated test items, the infants had never heard /*wo*/, or /*fe*/. Still, infants tended to listen more to the sentences containing a structural novelty. As a result, since this task could not be performed on the basis of mere transitional probabilities, Marcus et al. concluded that infants had the capacity to represent algebraic rules. However, Marcus et al.'s claim that an SRN could not model the observed effect was disputed by Elman (Seidenberg & Elman, 1999; Elman, 1999) and McClelland and Plaut (1999), basically on the account that an overlap need not be present in the "raw input" itself. Instead "the relevant overlap of representations required for generalisation […] can arise over internal representations that are subject to learning." (McClelland & Plaut, 1999, p.2) Transfer and generalization remain precarious issues, however, when it comes to computational modelling in a connectionist network. An experiment by Shanks et al. (1997) clearly illustrates this point.

Biconditional AGL: Shanks et al. (1997)

As mentioned before, Shanks and St John (1994) proposed to abandon the idea of the conscious/unconscious dichotomy in favour of a rule-based/instance-based dichotomy. The basic idea is that humans possess two learning systems capable of creating distinct forms of mental representation, one system consisting of symbolic rule-abstraction mechanisms and the other involving subsymbolic, memory-based, connectionist mechanisms (see Shanks, 1998, for a discussion). In this context, Shanks et al. considered transfer in AGL tasks to be at least to some extent mediated by abstract (rule-) knowledge and claimed that people systematically become aware of the relevant regularities in AGL tasks where only rule learning is possible. To demonstrate, Shanks et al. exposed subjects to artificial grammar strings generated by a biconditional grammar (see also Mathews et al., 1989). Biconditional grammars involve cross-dependency recursion (see Christiansen & Chater, 1999) such that letters that appear at each position before and after a central dot depend on each other. An example is given in Figure 1, where letter D is paired with F, G with L, and so on.

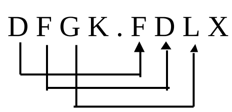

Figure 1. A biconditional grammar string as used by Shanks et al. (1997). Possible letters in each position before the dot are linked biconditionally with the letters that may appear after the dot.

Shanks et al. constructed biconditional grammar training strings as well as a set of grammatical and nongrammatical and test strings, in such a way that grammatical and nongrammatical test items could not be distinguished on the basis of their overlap with the training strings in terms of bigrams or trigrams (or any other *n*-gram). During training, two groups of subjects were shown strings one a time on a computer screen and had to perform one of two tasks on each trial. One group (the match group) had been told that the task was about memory, and had to select the correct string among five strings presented on screen. The other group (the edit group) was told that the strings had been constructed according to rules and that their task was to find them. On each trial, edit subjects had to indicate which letters they thought violated or confirmed to the rules, and were subsequently given feedback. All subjects then performed a classification test in which they were asked to decide which strings were grammatical or not. Shanks et al. showed a dissociation between the two groups: While the edit group performed well and most subjects extracted the rules, the match group performed at a chance level, thus suggesting that "instance-memorisation and hypothesis-testing instructions recruit partially separate learning processes." (Shanks et al., 1997, p.243)

The basic claim is thus that, in order to perform the biconditional grammar task, it is necessary to conceive of some abstract (symbolic) rule-like knowledge of the grammatical structure, and that, subsequently, the distinction made between grammatical and nongrammatical strings cannot be simulated by a connectionist network making use of simple frequency statistics. The goal of this paper is to demonstrate that in fact no such abstract rules are necessary and that, at least under some conditions, biconditional grammar learning can be accomplished by a network developing representations based on frequency statistics.

A Simulation of Shanks et al.

Simulation Parameters and Procedure

The Simple Recurrent Network (SRN) is a connectionist network especially designed to predict the next step in a sequence. Its design allows it to "keep in memory" the earlier steps in that sequence, by using what preceded as a *context*. This context is a copy of the learning-state at time t-1, which is fed back into the network at time t, together with the new input. In this way, the network is able to integrate the new input with what it has already learned in earlier stages, and will predict on this basis the sequence step at t+1. A typical example of an SRN is given in Figure 2.

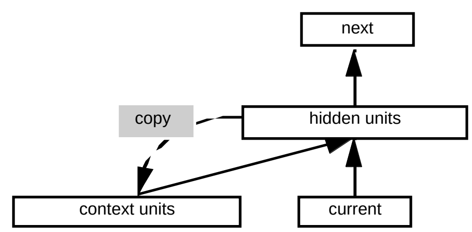

Figure 2. The Simple Recurrent Network as conceptualised by Elman (1990).

Importantly, on each time step the context units contain a copy of the *patterns of activation* that existed over the *hidden units* at t-1. As described in Servan-Schreiber, Cleeremans and McClelland (1988, 1989; see Cleeremans, 1993), learning progresses in a continuous fashion through three stages, during which more and more temporal contingency information is incorporated in the context, and hence in the hidden unit representations. The statistical regularities the SRN uses to predict the next letter are thus gradually "stored" in the hidden unit representations of the network. As a consequence, the network becomes able to behave in a rule-like manner and to predict the next element in the sequence *as if* it knew the grammar rules.

Network Architecture and Parameters. The SRN had 9 input and output units, necessary for representing the information that was available to the subjects in the Shanks et al. experiment. (The six letters of which the strings were composed, D, F, G, L, K and X, as well as the beginning and end of each string, together with the dot in between the first and the last four letters of a string.) The number of hidden units (and hence context units) was 100, which made use of logistic adjustment. The learning algorithm was error backpropagation, with a learning rate of .15 and the context being reset to zero after each complete string presentation. Weight adjustment was not applied on the connections from context to hidden units (1 on 1 relation). Momentum was set at .9.

Training Material. The basic training material consisted of a set of 18 strings as designed by Shanks et al. (List 1). Based on these strings, they created 18 grammatical and 18 nongrammatical strings.

The items were to meet four objectives: (1) Grammatical strings had to conform to the biconditional grammar: Letter position 1 is linked to 5, 2 to 6 and so on, with the linked letters being D–F, G–L, and K–X. (2) The use of the 6 letters was balanced, so that each letter appeared 3 times in each of the 8 letter locations. (3) Each training string differed from all other training strings by at least 4 letter locations. (4) Each training item had a grammatical similar item and an ungrammatical similar item that each differed from the training item by only 2 letter positions. Each training item was different from all other test items by at least 3 letter locations. The basic simulation was carried out on exactly these strings. A training epoch consisted of all 18 strings being presented once to the network, in a random fashion.

Measurement of Accuracy. Different measurements of accuracy exist, of which we used the *Luce ratio* (Luce, 1969) — a simple measure of relative strength in which the activation of the target output unit is divided by the sum of the activations of all output units. To assess network performance, we considered the average Luce ratios for all strings. In addition, we also considered the Luce ratio on a letter-by-letter basis for more detailed analyses.

Simulation Results

Learning. In order to assess learning, the network was tested before and during training on seven occasions. On each test, the network was tested on the 18 grammatical training strings, the 18 new grammaticals, and 18 nongrammaticals. Results were obtained over 9 simulations and averaged. As described before, the Luce ratio of the output was computed for each element of each string. Subsequently, the ratios were compared over the two conditions of interest (grammatical test/nongrammatical test) by means of an ANOVA, for each learning step.

As can be seen in Figure 3, the SRN was indeed able to discriminate between grammatical and nongrammatical strings. Original training strings were learned almost perfectly from 100 epochs onwards. Further, the network clearly discriminates between novel grammatical and nongrammatical strings (i.e., better predictions for grammatical strings), even *before* it is completely successful in mastering the training strings. ANOVA measures are, at 50 epochs, $F(1,161)=24.1$, $p<.001$; at 100 epochs, $F(1,161)=36.3$, $p<.001$; at 300 epochs, $F(1,161)=33.5$, $p<.001$. From 1000 epochs onwards, the network gets a little 'overtrained' on the original strings, causing it to do somewhat less well on the unseen strings; at 1000 epochs $F(1,161)=13.3$, $p<.001$; at 3000 epochs $F(1,161)=8.34$, $p<.005$. The figure also makes it clear that the main effect is not due to some initial biasing since initial performance is identical for the three types of strings (prior to training, $F(1,161)=1.13$, ns; at 10 epochs, $F(1,161)=.048$, ns).

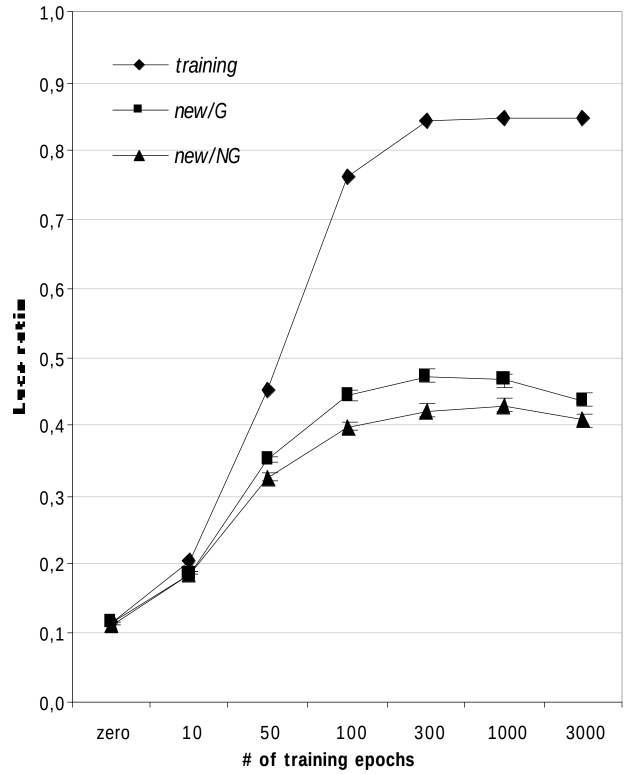

Figure 3. Network learning, measured with the Luce ratio. Error bars are shown for novel G and nonG strings.

Based on these findings we can therefore conclude that contrary to what Shanks et al. claimed, the SRN can in fact distinguish between grammatical and nongrammatical strings generated by a biconditional grammar without making use of explicit rules. In order to rule out the possibility of the SRN merely having learned to predict the dot and/or the end of a string, we computed the mean Luce ratios on a letter-by-letter basis, as presented below in Table 1. Shown are the important ratios, belonging to the letters after the dot (ratios for training strings had value 1). When the difference exceeds .05, the highest ratio is in bold.

Table 1 clearly shows the mean Luce ratios on a letter-by-letter basis to be higher in grammatical than in nongrammatical strings. This indicates that the network has learned something other than merely the dot or the end of the string.

Table 1. Mean Luce ratios on a letter-by-letter basis, in each position, after 3000 epochs, for grammatical and nongrammatical test strings (included is the frequency of occurrence of the letter in each position).

GRAM

	5th	#	6th	#	7th	#	8th	#
D	**.25**	(2)	**.89**	(2)	.01	(4)	**.69**	(4)
F	**.70**	(4)	.50	(4)	**.55**	(1)	.18	(3)
G	**.71**	(2)	**.49**	(2)	**.26**	(5)	**.38**	(3)
K	**.41**	(2)	**.72**	(5)	.09	(3)	**.99**	(2)
L	.58	(4)	.99	(3)	**.11**	(3)	**.68**	(3)
X	.43	(4)	**.56**	(2)	.33	(2)	.01	(3)

NGRAM

	5th	#	6th	#	7th	#	8th	#
D	.19	(2)	.70	(2)	**.72**	(4)	.13	(4)
F	.39	(4)	**.87**	(4)	.20	(1)	**.75**	(3)
G	.37	(2)	.37	(2)	.17	(5)	.01	(3)
K	.46	(2)	.32	(5)	**.19**	(3)	.33	(2)
L	**.74**	(4)	1.00	(3)	.00	(3)	.50	(3)
X	**.50**	(4)	.38	(2)	**.50**	(2)	.00	(3)

When the network fails to learn. In order to illustrate exactly when a network can learn, we include a simulation of a situation in which it *fails to learn*. We created two (grammatical) strings with a high degree of similarity, FGGG.DLLL and KGGG.XLLL, and presented them to the network in the same way as was done in the main simulation. Figure 4 shows the activation values of the 9 output nodes for one string (the other showed the same evolution in activation values), as well as the evolution of the Luce ratios.

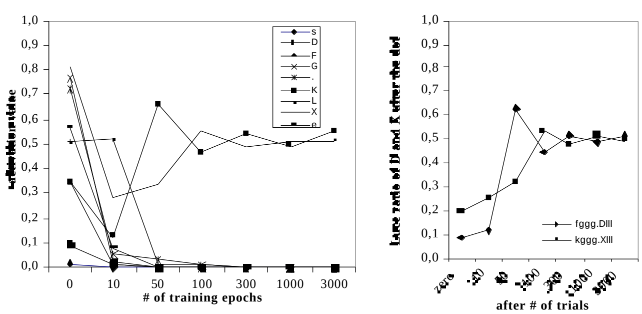

Figure 4. Left Panel: Evolution of output unit activations after presentation of the dot in the FGGG.DLLL string. Right Panel: Evolution of the Luce ratio after presentation of the dot for the two strings FGGG.DLLL / KGGG.XLLL.

Here , the network fails to reach a decision: it gets stuck at a "post-dot" activation value of 0.5 for both D and X (exactly the same plot is produced for the other string). The reason why learning fails in this case is addressed in the discussion.

Discussion

What had to be shown *was* shown, namely that a connectionist network, more precisely a Simple Recurrent Network, is able to make a distinction between grammatical and nongrammatical letter strings, generated from a biconditional grammar as used by Shanks et al. (1997). These strings were designed so that, according to them, subjects *had* to make use of abstract rules in order to accomplish the categorisation task. This paper clearly demonstrates that this is not the case, and that judgements of grammaticality using biconditional grammars can be made by extracting statistical features out of the material.

One of the major challenges in working with connectionist networks is how to probe the hidden units in order to "unfold" the complex representation of the stimulus material. Cluster analysis or principal component analysis performed on the hidden unit activations are standard ways of doing so, but may not always provide insight into how the representations enable the network to solve the task. The fact that cluster analysis does not reveal a clear structure does not necessarily imply that there *is* no structure. It may simply mean that the representational aspect needed to accomplish the most important aspect of the task, is *not* the most important aspect. Thus, clustering will not be carried out on that aspect — which, importantly, does not necessarily entail that the network is unable to use the relevant information successfully (see Cleeremans, 1993).

Biconditional grammars are difficult to master because they require maintaining information accross intervening irrelevant items. Servan-Schreiber et al. (1991) explored the conditions under which the network can carry information about distant sequential contingencies (e.g. 1–5) across intervening elements, to distant, to-be-predicted elements. It appeared that this information is retained as long as it is in some way relevant to predicting each intervening item (the *prediction-relevance* criterion). When it is not, the relevant information tends to be lost as training progresses, as a consequence of the way of the way in which representations of the temporal context are only gradually built up. Indeed, for different predictions to be achieved at any point in a sequence, the network needs to have developed different internal representations of the sequence so far. When two sequences are identical for a number of time steps so that the relevant information for making different predictions has to be retained over these intervening elements, each training trial actually induces the development of increasingly similar internal representations of the two sequences (because they require similar predictions)— exactly the opposite of what would be required for the network to master the material. Hence, the network fails to predict the fifth letter in the example above because the first letter of each string fails to be prediction-relevant when processing the intermediate Gs and ends up, as a result, with internal representations that fail to be sufficiently distinctive of each string to enable it to make different predictions about the fifth letter when presented with the dot.

Shanks et al. however, could not present the extremely simple (and for the network, extremely difficult) material to their subjects, for everyone would have discovered the rule in that case. Importantly however, the way in which their material is constructed results, for instance, in all the training strings to be determined by their *first two* elements — something that enables the network to learn the construction paths of each training string very quickly. In addition, in most cases, sequential information was in fact prediction-relevant on each step, which makes it easy for the network to distinguish between grammatical and nongrammatical strings. These findings suggest that the

Shanks et al. material was in fact inadequate to test for the rule based versus memory based distinction. As mentioned before however, it is clearly impossible to conceive of easy strings like KGGG.XLLL for which the rules are not discovered by subjects.

Insofar as simulations are concerned, while the SRN fails on such degenerate cases (unlike human subjects), the issue of whether this failure reflects a principled limitation of connectionist networks in general remains an open issue. Servan–Schreiber et al. showed that even very slight adjustments to the statistical structure of otherwise identical sequences could greatly enhance the prediction accuracy of the SRN. Thus, embedded information, as in recursive structures, need only be prediction-relevant in terms of the statistical distribution of the embedded elements for such structures to be successfully mastered by an SRN. There is also accumulating evidence that the pattern of failures observed with models like the SRN closely mimic that observed with human subjects (e.g., Christiansen & Chater, 1999) in the domain of natural language learning.

Empirically, we would like to suggest that experiments be carried out on a slightly different basis than used in Shanks et al., since their 'match' group showed no sign at all of having learned the material. One possibility would consist of changing the instructions of the match group so that attention is not *drawn away* from certain properties that might allow subjects to become sensitive to the structural properties of the material.

To conclude, we have demonstrated that a simple connectionist network can in fact master material previously considered to *require* the acquisition of rule-based knowledge for mastery of novel instances to occur. This outcome does not entail that rule-based learning never occurs (as it obviously does for some subjects in Shanks et al.'s experiments), but simply that biconditional grammars might not address all the issues involved in efforts to dissociate rule-based vs. memory-based learning processes in the implicit learning literature. Further simulation work will attempt to explore these issues in greather depth.

Acknowledgments

Axel Cleeremans is a Research Associate of the National Fund for Scientific Research (Belgium). This work was supported by a grant from the Université Libre de Bruxelles in support of IUAP program #P/4-19.

References

Berry, D.C. and Dienes, Z. (1993). *Implicit learning: Theoretical and empirical issues.* Lawrence Erlbaum Associates, Hove.

Christiansen, M., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance, *Cognitive Science*, *23*, 157-205.

Cleeremans, A. (1993). *Mechanisms of implicit learning.* MIT Press, Cambridge, MA.

Cleeremans, A. (1997). Principles for implicit learning. In D.C. Berry (Ed.), *How implicit is implicit learning?*, pp. 195–234. Oxford University Press.

Cleeremans, A., Destrebecqz, A., and Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*, 406–416.

Cleeremans, A. and Jiménez, L. (submitted). Implicit cognition with the symbolic metaphor of mind: Theories and methodological issues.

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

Elman, J.L. (1999). Generalization, rules and neural networks: A simulation of Marcus et al. (1999).

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol.1). New York: Wiley.

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

Mathews, R.C., Buss, R.R., Stanley, W.B., Blanchard-Fields, F., Cho, J.R., & Druhan, B. (1989). Role of implicit and explicit procesess in learning from examples. *Journal of Experimental Psychology: Learning, Memory, and Cognition, , 15,* 1083-1100.

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

Neal, A. and Hesketh, B. (1997). Episodic knowledge and implicit learning. *Psychonomic Bulletin and Review, 4*, 24–37.

O'Brien, G. and Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences, 22*, 175-196.

Perruchet, P. and Vinter, A. (submitted). The self-organizing consciousness.

Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A., and Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

Seidenberg, M.S. & Elman, J.L. (1999). Do infants learn grammar with algebra or statistics? Letter in *Science, 284*.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J.L. (1991). Graded State Machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7*, 161–193.

Shanks, D.R. (1998). Distributed representations and implicit knowledge. In K. Lamberts and D.R. Shanks (Eds.), *Knowledge, concepts & categories*, pp.197–214. Psychology Press, Hove.

Shanks, D.R., Johnstone, T., and Staggs, L. (1997). Abstraction processes in artificial grammar learning. *The quarterly Journal of Experimantal Psychology, 50A*, 216–252.

Shanks, D.R. and St John, M.F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367–447.

St John, M.F. and Shanks, D.R. (1997). Implicit learning from an information processing standpoint. In D.C. Berry (Ed.), *How implicit is implicit learning?*, pp. 124–161. Oxford University Press.