Cognitive Critique

NOT QUITE A BOOK REVIEW: CHRISTIANSEN'S *INFINITE LANGUAGES, FINITE MINDS* AS AN INTRODUCTION TO CONNECTIONIST MODELING, RECURSION, AND LANGUAGE EVOLUTION

MIRIAM KRAUSE
Department of Speech-Language-Hearing Sciences
University of Minnesota, Minneapolis, Minnesota

E-MAIL: krau0067@umn.edu

Accepted 20 April 2009

KEYWORDS
connectionist modeling, recursion, language evolution

ABSTRACT

In his 1994 doctoral thesis, Morten Christiansen used connectionist modeling to demonstrate that models could produce recursive language-like output without the use of prescribed recursive rules. The purpose of this paper is to provide a brief overview of Christiansen's arguments and the models he used to support them, as a foundation for understanding their implications regarding the role of recursion in the evolution of language. Based on connectionist principles, Christiansen argued against concepts such as the competence/performance distinction and poverty of stimulus as an objection to learning-based theories of language acquisition. He proposed that recursion, rather than being a prescriptive shaper of language, is an emergent outcome of processing constraints in the human brain.

INTRODUCTION

Morten Christiansen's 1994 doctoral thesis, titled *Infinite Languages, Finite Minds: Connectionism, Learning, and Linguistic Structure*, provides an in-depth exploration of the relationship between connectionist modeling and other views of language that involve an innate Universal Grammar. The thesis represents an important first stage of the author's research trajectory, and continues to be cited consistently despite never having been published. The purpose of this paper is to provide an overview of Christiansen's thesis and its relationship to some more recent work in the area of language evolution, particularly Hauser et al.'s landmark 2002 *Science* article on language evolution. (A brief primer on connectionist modeling is also provided; readers already comfortable with the basic principles will probably want to skip that section.)

Christiansen's 180-page thesis covers a broad range of topics, including several specific connectionist models that particularly focus on the syntactic phenomenon of recursion. Based in part on these models, Christiansen challenges several conventions of linguistic theory as he sees them: first, he argues to discard the competence/performance distinction; next, he challenges the idea that "poverty of stimulus" is an obstacle to learning-based language acquisition; and he also posits that language evolved, not in an exaptationist or adaptationist paradigm (as advocated by Chomsky and Pinker, respectively), but as a "non-obligate symbiant." Each of these topics is discussed in more detail below; the overall purpose of this paper is to explore Christiansen's thesis and its relevance to discussions of language acquisition and evolution.

RECURSION

Recursion is a concept relevant to many areas, including mathematics, music, and language. For example, in mathematics, a recursive function contains itself as part of its definition; a recursive definition is that while an individual's parents are classified as ancestors, the parents of those parents are *also* ancestors. In language, recursive properties include embedding (such as a sentence with another sentence inside it) and iteration (such as a sentence with clauses attached at the end).

Recursion is of particular interest in a discussion of connectionism and language evolution because, as Christiansen says, it

has been "used as an existence proof of the need for increasingly powerful language formalisms to describe natural language" (p. 38). One prominent example of this is the proposal of Hauser et al. (2002), who posit that recursion may be the only characteristic of language that is uniquely human. The authors' primary concerns in their article are to suggest components for the Faculty of Language – Broad and Faculty of Language – Narrow; and to propose empirically testable hypotheses regarding the evolution and current features of these faculties. The assumption that recursion is an inherent feature of language, much less *the* key feature, is emblematic of the conceptualization of language as constrained – or at least scaffolded – by Universal Grammar. Regarding recursive syntax, Hauser et al. state that "[Faculty of Language-Narrow] takes a finite set of elements and yields a potentially infinite array of discrete expressions…[and] there is no non-arbitrary upper bound to sentence length. In these respects, language is directly analogous to the natural numbers" (p. 1571). Hauser et al. acknowledge that external limitations such as memory capacity may impact recursive expression and comprehension, but claim that infinite recursion is nevertheless a feature of language. This distinction is emblematic of the competence/performance contrast that Christiansen challenges: by showing that connectionist models can produce recursion without having it built in to their network structure as rules, he argues that recursion is a feature of language that may emerge due to processing constraints rather than the other way around.

In constructing his models, Christiansen addresses several different types of recursion. These include center-embedded or mirror recursion and cross-dependent or identity recursion. As implied by their names, the first type takes the form ABCCBA, while the second takes the form ABCABC. These types of recursion are both *non-iterative*, meaning that they do not recurse by continually adding features onto the beginning or ending of a sentence (an example in English of a sentence with *iterative* recursion would be "In the mirror, you see yourself looking at yourself looking at yourself looking at yourself…" (Hurford, 2004)). Iterative recursion is subjectively easier to process than non-iterative, and it is non-iterative recursion that Christiansen focuses on in his thesis. These different types of recursion will be further discussed below, in the context of discussing the connectionist models that produce them. Before examining the details of Christiansen's particular models, however, a more general introduction to connectionist modeling is warranted.

CONNECTIONIST MODELING: A PRIMER

Connectionist models are a useful tool for exploring simplified neural networks in studies of how the brain may acquire and use information. McLeod et al. (1998) offer an excellent summary of the principles of connectionist modeling:

> Connectionist modeling is inspired by information processing in the brain. A typical model consists of several layers of processing units. The unit can be thought of as similar to a neuron or a group of neurons. Each unit sums information from units in the previous layer, performs a simple computation on this sum, such as deciding whether it is above a threshold, and passes the result to units in the next layer. The pattern of activity on the first layer reflects the stimulus presented to the model. This pattern gradually becomes transformed to produce the pattern of activity on the final layer – the model's response (p.9)

The influence of one unit on another has a specific weight, which adjusts during a model's training according to its learning algorithm. While no model is sophisticated enough to truly represent how a brain works, individual units in a network are analogous to neurons, and the connection weights are analogous to the excitatory (positive weight) and inhibitory (negative weight) input each neuron receives from various sources. Learning is represented by changes in connection weights. Models often go through separate training and testing phases, simulating learning and mature performance, respectively.

Perhaps the best way to understand how a model works is to describe simple examples. One of the most famous problems that can be solved using a connectionist model is the Boolean XOR function, whose solution demonstrates the necessity of using hidden units in modeling. This function states that, given two inputs, if one but not the other input is activated, the output should be activated; but if neither or both inputs are activated, the output should remain inactive. In symbols, this is represented in Table 1.

Input		Output
0	0	0
0	1	1
1	0	1
1	1	0

Table 1. The XOR problem.

When attempting to solve this function using only an input and output layer, it quickly becomes apparent that no real solution exists (Figure 1), no matter what the connection weights are. If an input of [0,1] or [1,0] passes the output node's activation threshold, then a [1,1] input will erroneously activate the output node as well.

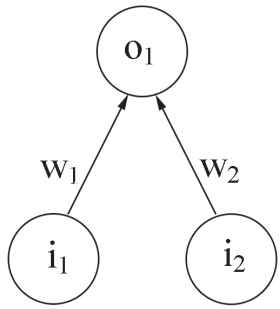

Figure 1. A simple model with two input units and one output unit. This model cannot produce the output required for the XOR function.

However, if the model is expanded to include a hidden layer, it can be trained to produce accurate output for the XOR function. Example connection weights that successfully model this function are shown in Figure 2. Input unit 1 (i1) has an excitatory effect on hidden unit 1 (h1) and an inhibitory effect on hidden unit 2 (h2). If activation is constrained to be between 0 and 1, then (for example) h1 will fire if i1 alone is active; but if i2 is also active, its inhibitory connection weight will counteract the excitatory activity of i1, and h1 will not fire. These counteracting connection weights ensure that h1 and h2 (and, in turn, the output node) will be active if i1 *or* i2, but not both, are active.

In addition to hidden units, arguably the most important feature of a connectionist model is its ability to learn. Truly useful models are much too complex to determine ideal connection weights by

intuition or trial-and-error, as was done with the example in Figure 2. Models are trained based on sets of inputs and desired outputs, using learning algorithms to adjust connection weights slightly with each iteration, or epoch, of training. At the end of each training epoch, actual outputs are compared with the desired outputs, and the differences are used by the learning algorithm to adjust connection weights throughout the network. In experiments such as Christiansen's, the connection weights are set and cannot be adjusted once training is complete; testing involves entering novel inputs and observing the network's outputs.

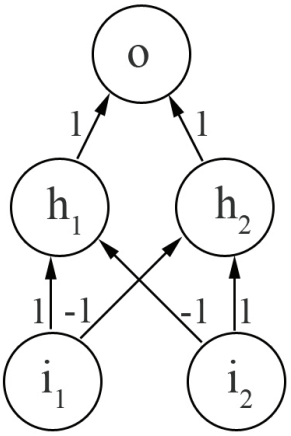

Figure 2. A model with two input units, two hidden units, and one output unit. This model can produce the output required for the XOR function, with example connection weights shown. When both input units are active, their activations cancel each other out at the hidden layer, resulting in zero activation of the output unit.

In the case of Christiansen's models, the desired outputs varied according to the sequence of words in sentences (outputs during testing were essentially predictions of the next word in the sentence). In order to train patterns accounting for time, models may incorporate an additional set of nodes called *context units*. Context units simply store the activation of the hidden units for one time step, adding information from the previous time step to the hidden layer's input for the current step. Figure 3 illustrates this concept, constituting a simple recurrent network (SRN).

With a basic comprehension of the concepts described here, it is possible to understand the models Christiansen describes, and the

arguments he constructs based on his findings. Ultimately, connectionist modeling is useful not only because it can simulate existing behavioral results, but because it can make novel predictions to suggest new directions for behavioral experimentation. The output of a model in a simulation can provide predictions for experimental testing that might not have been thought of based only on traditional theories. Connectionist networks, like any models, are a useful tool to help researchers frame, think about, and explore their questions of interest.

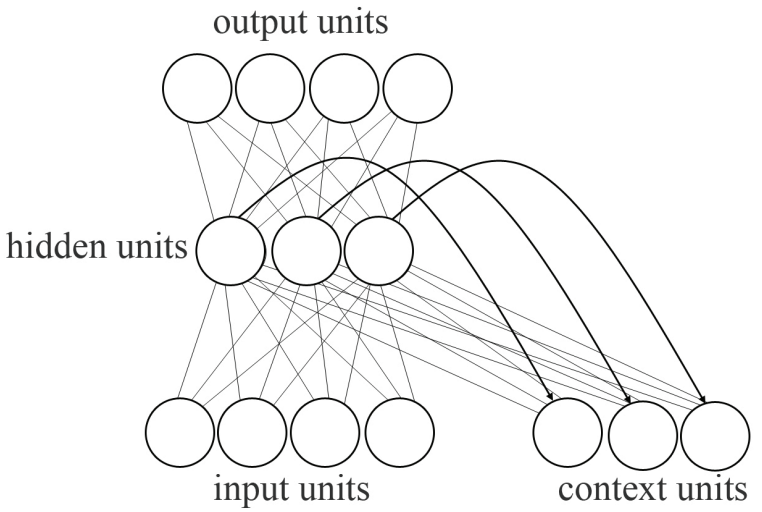

Figure 3. (Based on Plunkett and Elman, 1997) A simple recurrent network (SRN), including context units as well as hidden units. Because the context layer only stores the pattern of hidden unit activation from the previous time step, connection weights are fixed going from the hidden layer to the context layer (bold lines). Weights are adjustable (susceptible to learning during training) in all other connections, including from the context units back to the hidden units.

CHRISTIANSEN'S MODELS

In his thesis, Christiansen describes several of the models that he constructed, all of which were demonstrated to be able to produce recursive output without employing prescribed recursive rules. For the sake of brevity, the models will be described in simplified form

here. The first model had a simple two-"word" vocabulary (A and Z as "verb" and "noun"), with lowercase and capital representing "singular" and "plural." The network was trained on three types of recursive grammar. Counting recursion consisted of sentences that were merely "*n* occurrences of nouns followed by *n* occurrences of verbs… with no agreement between nouns and verbs" (p. 41), for example, aAaZZz. Center-embedded or mirror recursion consisted of "a string *n* of nouns followed by a string *n* of verbs whose agreement features constitute a 'mirror image,'" for example, aAazZz. Finally, cross-dependent or identity recursion comprised sentences with "a string *n* of nouns followed by a string *n* of verbs whose agreement features are ordered identically," for example, aAAzZZ. Each type of recursion included different depths of embedding, or layers of recursion; for example, aAzZ has an embedding depth of two, while aaAZzz has a depth of three. The English sentence, *The cart the horse the man bought pulled broke*, while technically grammatical, has recursion embedded three deep and is rather difficult to process.

Overall, Christiansen's networks (over various trials with 5, 10, and 25 hidden units; varying epochs of training; and other variations) were shown to be capable of strong performance (generally, mean cosine above 0.7 and mean squared error below 0.4) for recursive embedding depths of 0, 1, and 2. Performance decreased significantly at a depth of 3, and results for greater depths are not reported. Christiansen repeated this experiment with an eight-word vocabulary (four "nouns" and four "verbs"), using a network with 20 hidden units and 200-400 epochs of training. Results were very similar to the previous experiment. In addition, there was no significant difference between performance on trained and untrained sentences, demonstrating that the network was able to generalize to novel stimuli.

Christiansen's next set of experiments involved a much more complex model, which used a vocabulary of "two proper nouns, three singular nouns, five plural nouns, eight verbs in both plural and singular form, a singular and a plural genitive marker, three prepositions, and three ('locative') nouns to be used with the prepositions" (p. 78). Two distinct grammars were trained; both included iterative recursive sentences such as *Mary knows that John's boys' cats see dogs*. The center-embedding grammar also included non-iterative sentences such as *girl who men chase loves cats*; the cross-dependency grammar expressed the same idea as *men girl cats*

chase loves (though unfortunately the latter is not grammatical in English).

The results of these more complex modeling experiments were twofold. First, networks trained on both the center-embedded and the cross-dependent grammars were able to perform well. The performance of this type of model is measured by the accuracy of the model's predictions after each word: after the first word of the sentence is input, the relative activation levels for the various output units act as a prediction of what words are likely to come next. At the very beginning of a sentence (before the first word has been input) the output units for singular noun and plural noun are most highly activated, because all the sentences the model trained on begin with nouns. In a specific example, after the initial word *cats*, two of the most strongly activated output units in Christiansen's center-embedded network were the plural transitive verb and the word *who*. Once *who* is chosen as the second input word, the network's strongest activation is again the plural transitive verb. Predictions are made after each word for each sentence and, in this example, the network accurately and strongly predicted the end of the sentence following *cats who John who dogs love chases run*. This is a three-level embedded recursive sentence which is technically grammatical and yet very difficult to comprehend.

Even more interesting than this general success, however, is the finding that the models were only able to successfully learn these grammars using an incremental memory learning strategy, designed to simulate maturational learning by limiting the memory capacity of the model's context units in early stages of training. Christiansen reports that other simulations, not discussed in detail in the thesis, showed that networks not constrained by such maturational learning strategies were unable to satisfactorily learn the two grammar systems under discussion. This result features prominently in the author's later comments on the poverty of stimulus issue.

THEORETICAL IMPLICATIONS OF CONNECTIONISM

Part of the intent of Christiansen's thesis is to explore not only his specific connectionist models but also the repercussions of a connectionist paradigm in general for language processing, acquisition, and evolution. Relative to these three areas, respectively, he discusses the competence/performance distinction; the poverty of stimulus argument; and the concept of language as a non-obligate symbiant.

COMPETENCE/PERFORMANCE DISTINCTION

Even before describing his experiments, Christiansen argues that the distinction between linguistic competence and performance should be discarded on methodological grounds. He then goes on to support that argument on the strength of various experimental results, including those described above. He rejects the competence/performance distinction (C/PD) because of what he calls the "Chomskyan competence paradox" (p.18), in which Universal Grammar rules are "immune to empirical falsification – even (in a pinch) to informant judgments." Christiansen points out that even in a version of C/PD that allows for some processing constraints on a grammar,

> …empirical evidence that appears to falsify a particular grammar can always be rejected as a result of processing constraints – either construed as limitations on working memory (strong C/PD) or as a combination of working memory limitations and false linguistic meta-knowledge (weak C/PD). (p.21)

Having presented his methodological objection to C/PD, he goes on to make the case that, even when improved with training and practice, the capacity for real humans to produce and comprehend multiply embedded recursive sentences is limited, and no evidence exists for an infinite capacity to process them. He also cites research that demonstrated improved performance on ungrammatical patterns with practice, negating the idea that center-embedding has a privileged learnable status as an infinite-capacity component of Universal Grammar. Ultimately, Christiansen introduces connectionism as an alternative "representational vehicle that will allow us to avoid the methodological problems of the C/PD as well as model syntactic processing as being deterministic and unconscious in

compliance with the limitations set by the experimental evidence" (p.29).

POVERTY OF STIMULUS

In addition to challenging the C/PD concept, Christiansen also goes to some lengths to dispute the idea of "poverty of stimulus" as an objection to learning-based theories of language acquisition. The poverty of stimulus argument states that, because the input children receive in the process of language learning is so inconsistent and incomplete, language could not be learned without some built-in linguistic structure such as Universal Grammar. Christiansen addresses five components of the poverty of stimulus argument in turn: arbitrary universals, noisy input, infinite generalization, early emergence, and inadequacy of learning methods.

Regarding arbitrary universals, Christiansen reiterates his view that grammatical principles are not arbitrary, but rather are "natural side-effects of sequential processing and learning in a system with certain memorial and perceptual limitations" (p.150). Thus, the apparent arbitrariness of these universals (he gives the example of subjacency) is not an obstacle to learning, but an artifact of it. His answer to the problem of noisy input is not as thorough; he states that connectionist models are able to learn despite inconsistent training, and that some inconsistencies actually facilitate learning in the models, but he does not speculate about why this is the case. However, he acknowledges that further work is needed to determine whether this principle is truly applicable under the circumstances faced by children learning language. Christiansen did later address the question of inconsistent input in terms of statistical learning (e.g. Christiansen and Curtis, 1999), and other work has been done in this area, including studies of deaf children learning ASL – and developing native proficiency – with only non-native input (e.g. Singleton and Newport, 2004).

Infinite generalization is the problem of language learners needing to parse and organize vast amounts of essentially arbitrary information. The sets of information received are different for every learner, and yet every learner deduces the same underlying regularities. Here Christiansen invokes his findings regarding maturational learning in connectionist models, in which the models' learning benefited from gradually lessening memory restrictions designed to be analogous to the processing limitations of a maturing child. If initially restricted processing provides a beneficial scaffold for

early learning, that could explain how very different pools of input are distilled into comparable, coherent systems at early stages of language learning.

Christiansen criticizes the early emergence of language as an argument for Universal Grammar by addressing a few specific features such as categorical perception. As with several of the other poverty of stimulus arguments, he again explains early emergence features as a direct result of biological features such as cochlear physiology, dismissing the need to invoke a pre-existing Universal Grammar to explain the phenomena in question. Finally, he describes the inadequacy of learning argument as claiming that only Universal Grammar can explain an ability to learn an entire set of linguistic rules that would be difficult for even a trained scientist to deduce empirically. However, here again, by shifting to the perspective of a processing-constrained, statistical approach to language acquisition rather than a rule-constrained deductive one, Christiansen rejects the final poverty of stimulus argument along with its fellows.

LANGUAGE AS NON-OBLIGATE SYMBIANT

In addition to addressing the character of language in the C/PD question and the acquisition of language in the discussion of poverty of stimulus, Christiansen also tackles the topic of language evolution. He skirts the issue of *why* language evolved, merely speculating that it developed from "sequential learning" mechanisms, probably starting out as a manual system before developing into a vocal one. Rather than focusing on these questions, he concentrates on the evolutionary mechanism itself. Theories of language evolution are generally either exaptationist (favored by Chomsky) or adaptationist (favored by Pinker). Exaptationist theories involve the evolution of language from some earlier traits that developed for other purposes; adaptationist theories argue that language evolved directly for its current purpose. As an alternative to these perspectives, Christiansen proposes that language behaves as a *non-obligate symbiant*, "a kind of beneficial parasite…that confers some selective advantage onto its human hosts without whom it cannot survive." (p. 26). This means that, rather than viewing language evolution as being driven by potentially unrelated factors (exaptationist) or by natural selection for effective communication (adaptationist), language itself is perceived as the evolving organism, adapting to the environment of the human processing system. Christiansen states, "I contend that language is the product of an evolutionary process in which lan-

guage had to adapt to the human learning mechanism (with all its developmental idiosyncrasies) in order to 'survive'" (p.125); and that "language learning is therefore more appropriately explained as a product of natural selection of linguistic structures, rather than natural selection of biological structures, such as [Universal Grammar]" (p.126). While not inseparable from connectionism, this perspective is clearly consistent with an overall learning- and processing-centered view of language.

PROBLEMATIC POINTS

There are a number of criticisms that can be made of Christiansen's work, some of which he raises in the thesis, such as the need for further modeling of additional language universals, creolization, and maturationally constrained learning. A number of other issues are also apparent, however. For example, he gives only limited attention to comparing his networks' performance to that of humans. One strong advantage of connectionist modeling is its ability to replicate and predict actual human behavior, but Christiansen only briefly touches on demonstrated capacities for processing of recursion and garden path structures during his discussion of C/PD. His presentation of experimental data from the models would have been greatly strengthened had he presented analogous behavioral data along with it to directly support the claim that the models' performance was comparable to that of humans.

In his section on poverty of stimulus, Christiansen seems to pick and choose the arguments that he describes to represent the viewpoints he opposes. For example, the section on the problem of infinite generalization is almost exclusively centered on refuting the claims of a 1967 paper by Gold. This is symptomatic of a broader criticism, which is that the scope of the thesis did not allow for more in-depth analyses of the numerous topics touched on (of which the infinite generalization argument is just one example). Outside the central network simulations, most issues in the thesis are only discussed for a tantalizing few pages before the author moves on to his next subject.

Another area that suffered from this breadth-to-the-detriment-of-depth quality was Christiansen's section on the evolution of language. His description of the original sequence of language evolution is frustratingly sparse, with little concern for how the transition was made from "sequential learning mechanisms" to language.

Christiansen also does not delve into the issue of how streams of vocalizations could have evolved into syntactically organized and learnable grammars, other than in the discussion of the benefits of maturationally constrained learning. His networks begin with specific words already assigned to particular roles within the model; the thesis' questions had to do with whether the models could learn to produce recursive output, and the topic of how roles became assigned to particular words was not relevant. Others have addressed this question, however, including Kirby (2002), who constructed a connectionist model that essentially generated novel recursive grammars from undifferentiated streams of input within a few hundred generations.

Perhaps the most challenging question faced by connectionist paradigms is how to reconcile the statistical austerity of the models with the experimentally observed messiness of localized functions in the brain. Certainly some tasks, such as memory and naming, are distributed diffusely throughout the brain; but others, such as Broca's and Wernicke's language areas in the left hemisphere, appear more localized (the functions of these areas are not restricted to speech and language; nevertheless they do appear to play important and distinct roles in language expression and comprehension (e.g. Kober et al., 2001). If connectionist models are truly analogous to biological neural networks, even to a limited degree, then they must be able to account for the fact that different areas of the brain have relatively specific roles for different functions. If connectionism is to be truly upheld, models must be designed wherein localized lesions preferentially affect some functions but not others. Extensive work on lesioning networks to model aphasia, for example, have begun this process (e.g. Gordon and Dell, 2003). While it is certainly beyond the scope of the current project to address this issue in depth, further work in the area is clearly needed.

CONCLUSION

Despite its weaknesses, Christiansen's thesis work provides valuable contributions to several areas of linguistic theory, including discussions of the role of recursion in human language, the balance of innateness versus learning in language acquisition, and possible mechanisms of language evolution.

Taken as a whole, Christiansen's experimental and theoretical work may be interpreted as suggesting a new way of framing ques-

tions about language evolution. Traditional perspectives (exemplified by Hauser et al.) prompt questions such as, "Was recursion necessary for the evolution of language?" or "Can other species do recursion?" Christiansen's work urges a paradigm shift toward asking questions such as, "What are the minimum processing requirements to produce recursive output?" or "Would a processor that could *not* produce recursive output still be able to learn language?" and conversely, "Would any processor complex enough to produce recursive output *not* be able to learn language?" These questions arguably address the same issues raised by Hauser et al. in their attempt to define empirically testable problems, but approach them from a somewhat novel direction. Indeed, despite his ardent refutation of important conventions like C/PD and Universal Grammar, the drive for empiricism is highly compatible with Hauser et al.'s stated goals.

Overall, based on Christiansen's experimental findings and theoretical discussions, it is possible that concepts such as recursion and potential infinity are not inherent (prescriptive) features of human language as such; rather, they are meta-linguistic concepts that humans use to describe observed features of language. It seems appropriately ironic that language must be used to describe this emergent linguistic phenomenon of recursion – and to frame the arguments that recursion is indeed an emergent property rather than a prescriptive sculptor of language.

REFERENCES

Christiansen M (1994) Infinite languages, finite minds: connectionism, learning and linguistic structure. Ph.D. thesis, University of Edinburgh

Christiansen M, Curtis S (1999) Transfer of learning: rule acquisition or statistical learning? Trends Cogn Sci 3:289-290

Gold E (1967) Language identification in the limit. Inform Control 16:447-474

Gordon J, Dell G (2003) Learning to divide the labor: an account of deficits in light and heavy verb production. Cognitive Sci 27:1-40

Hauser M, Chomsky N, Fitch T (2002) The faculty of language: what is it, who has it, and how did it evolve? Science 298:1569-1579

Hurford J (2004) Human uniqueness, learned symbols and recursive thought. Eur Rev 12:551-565

Kirby S (2002) Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe EJ (ed) Linguistic evolution through language acquisition: formal and computational models. Cambridge University Press

Kober H, Möller M, Nimsky C, Vieth J, Fahlbusch R, Ganslandt O (2001) New approach to localize speech relevant brain areas and hemispheric dominance using spatially filtered magnetoencephalography. Hum Brain Mapp 14:236-250

McLeod P, Plunkett K, Rolls E (1998) Introduction to connectionist modelling of cognitive processes. Oxford University Press, Oxford

Plunkett K, Elman J (1997) Exercises in rethinking innateness: a handbook for connectionist simulations. The MIT Press, Cambridge, MA

Singleton J, Newport E (2004) When learners surpass their models: the acquisition of American Sign Language from inconsistent input. Cogn Psychol 49:370-407