Elman, J.L. (1999). Origins of language: A conspiracy theory. In B. MacWhinney (Ed.), *The emergence of language.* Hillsdale, NJ: Lawrence Erlbaum Associates.

# Origins of language: A conspiracy theory

Jeffrey L. Elman

*Department of Cognitive Science*
*University of California, San Diego*
*elman@cogsci.ucsd.edu*

## Introduction

Language is puzzling. On the one hand, there are compelling reasons to believe that the possession of language by humans has deep biological roots. We are the only species that has a communication system with the complexity and richness of language. There are cases of non-human primates who can be taught (sometimes only with heroic effort) some aspects of human language, but their performance comes nowhere close to those of a six-year old child. Second, although languages differ, but there are also striking similarities across widely divergent cultures. Finally, there are significant similarities in the patterns of language acquisition across very different linguistic communities. These (and other considerations as well) all suggest that species-specific biological factors play a critical role in human's ability to acquire and process language.

So what is puzzling? First, it is not at all clear what the biological foundations are. What precisely do we mean when we say that the human propensity for language is innate (or as Stephen Pinker puts it, is an "instinct", Pinker, 1994). Do we only mean, when we say that "language is innate" that one must possess a human genome in order to speak (hear, read, sign)? This is not terribly informative; after all, getting a driver's license also requires a human genome (although driving the freeways of Southern California, one sometimes wonders). But we do not view this as an especially useful explanation of the origin and nature of the skills and competencies which are required to drive a car. Second, when we actually *look* at the genome, we see little that suggests any obvious connection with language. The recurring lesson from recent genetic research is that behaviors typically rest on the interaction of large numbers of genes, each of which may participate in many other processes (e.g., Greenspan, 1995). Claims to the contrary notwithstanding (Gopnik & Crago, 1991; but see also Vargha-Khadem et al., 1995), there are no good examples of selective impairment of language which can be traced to defects in isolated genes.

Yet the fact remains, humans have language and chimpanzees do not. This is true no matter how human-like a chimp's environment and upbringing are made. Thus we remain with a puzzling set of questions:

- *Why does our species have language, and no other? What are the species differences that make language possible?*
- *Why does language take the forms it does, and not others?*
- *How does language emerge in the language-learning child?*
- *How do we account for both global patterns of similarity in language behavior, as well as individual variations on those patterns?*

In this chapter I outline a connectionist perspective on language development with the goal of ultimately (if not now) providing answers to these questions. The particular perspective I put forward is one which has been developed together with my colleagues Elizabeth Bates, Mark Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett (see Elman et al., 1996, for a fuller account). In our view, biology plays a crucial role in determining the outcome of development, and in the case of humans, enabling language. However, rather than viewing the developmental process as one in which biology contributes some portion of the answer, and experience another (much like a jigsaw puzzle, in which biology assembles most of the pieces and experience fills in the rest), we see these two forces as engaged in a complex synergy. The challenge, of course, is to be able to clarify the way in which these interactions occur.

I begin by presenting with a taxonomy for thinking about alternative ways in which behaviors might be constrained by the biology. As it turns out, there are reasons to believe that some of these alternatives are more plausible than others, and that what may be the most widely held view of innateness is highly unlikely in the case of language. After discussing other ways in which biological constraints might constrain outcomes, I will describe two simulations which illustrate how what appears to a much weaker alternative constraint in fact has considerable power. Among other things, these simulations also suggest that domain-specific behaviors can be achieved through mechanisms which are themselves not domain-specific. Finally, I discuss how these results fit in with more general findings regarding development into what I call a “conspiracy” theory of language origins.

## Ways to be innate

At least some of the controversy surrounding the nature/nurture debate arises from lack of clear notions regarding what is meant when it is claimed that a behavior is innate. In the framework outlined by Elman et al, 1996, we found it useful to think about development as a process which can occur at multiple levels (using level here in a heterarchical rather than hierarchical sense), and in which processes at different levels may interact. At all levels, the constraints may crucially depend on interactions with the environment. When we say that an outcome is innate, then, we mean that it is significantly constrained at one or more of these levels, given the expected inputs from the environment. The taxonomy that we developed makes reference to constraints at the levels of *representations*, *architectures*, and *timing*. These levels can be defined in terms of brain development, but we also find it useful to talk about their network analogs.

***(1) Representational innateness.*** If cognitive behaviors are the immediate product of our mental

states, and these are equivalent to brain states, then the most specific way of constraining a cognitive behavior is to constrain the brain states which underlie it. Brain states are patterns of activations across neurons, and their proximal cause lies in the pattern of synaptic connections which generate that activity. Thus, the most direct and specific way of constraining a behavior would be to specify in advance the precise pattern of neuronal connectivity which would lead to that behavior. In brains, then, a claim for representational innateness is equivalent to saying that the genome somehow predetermines the synapses between neurons. In neural networks, representational innateness is achieved by hand-wiring the network and setting the weights prior to learning.

At least some of the discussion regarding the origins of language appear to assume that representational innateness is what is assumed. Thus, for example, Pinker (1994) claims that,

> *It is a certain wiring of the microcircuitry that is essential.…If language, the quintessential higher cognitive process, is an instinct, maybe the rest of cognition is a bunch of instincts too—complex circuits designed by natural selection, each dedicated to solving a particular family of computational problems posed by the ways of life we adopted millions of years ago.* (Pinker, 1994; pp. 93, 97)

Although this scenario is logically possible, and there are some animals for whom the genome appears to constrain the topology and connectivity of specific cells, we shall see below that representational innateness is highly dubious as a mechanism for ensuring language in humans.

***(2) Architectural innateness.*** Outcomes can also be constrained by limiting the architectures which are available. As used here, architecture will refers to organization which is at a higher level of granularity than the prespecified connections between neurons (or nodes) which guarantee representational innateness. Architectural constraints in fact can vary along a large number of dimensions, but in general fall into three broad classes: unit-based, local, and global. *Unit-based* architectural constraints deal with the specific properties of neurons, including firing threshold, refractory period, etc.; type of transmitter produced (and whether it is excitatory or inhibitory); nature of pre- and postsynaptic changes (i.e., learning), etc. In network terms, unit level constraints might be realized through node activation functions, learning rules, temperature, momentum, etc. It is clear that unit level constraints operate in brain development. There are a relatively small number of neuron types, for instance, and they are neither randomly nor homogeneously distributed throughout the brain. The unit level constraints are fundamental to brain organization, since they concern the lowest level of computation in the brain. *Local architectural constraints* operate at the next higher level of granularity. In brains, these describe differences in the number of layers (e.g., the six-layered organization of cortex), packing density of cells, types of neurons, degree of interconnectivity (“fan in” and “fan out”), and nature of interconnectivity (inhibitory vs. excitatory). In network terms, local architectural differences would include feedforward vs. recurrent networks, or the layering of networks. Interestingly, the cortex itself appears to display relatively little in the way of local architectural differences at early stages of development. The much greater differentiation which is found in the adult cortex appears to result from development, and an interesting question is how these differences arise. This, in fact, is one of the goals of the second simulation which will be described below. Finally, *global architectural constraints* specify the way in which the various pieces of a system—be it brain or network—are connected together. Local architecture deals with the ways in which the low-level circuitry is laid out; global architecture deals with the connections at the macro level between areas and regions, and especially with the inputs and outputs to subsystems. If one thinks of the brain as a network of networks, global

architectural constraints concern the manner in which these networks are interconnected.In brain terms, such constraints could be expressed in terms of (e.g., thalamo-cortical) pathways which control where sensory afferents project to, and where efferents originate. Very few network models employ architectures for which this sort of constraint is relevant (since it presupposes a level of architectural complexity which goes beyond most current modeling). One might imagine, however, networks which are loosely connected, such that they function somewhat modularly but communicate via input/output channels. If the pattern of inter-network connections were prespecified, this would constitute an example of a global architectural constraint.

***(3) Chronotopic innateness.*** A third way in which outcomes can be constrained is through the timing of events in the developmental process. Indeed, as Gould (and many other evolutionary biologists) has argued eloquently, changes in the developmental schedule play a critical role in evolutionary change (Gould 1977; see also McKinney & McNamara, 1991).In networks, timing can be manipulated through exogenous means, such as control of when certain inputs are presented. Or timing can arise endogenously, as seen in Marchman's simulations of the critical period (Marchman,1993); in these networks, the gradual loss of plasticity in a network comes about as a result of learning itself. In brains, timing is sometimes under direct genetic control but the control of timing may also be highly indirect and the result of multiple interactions. Hence the onset and sequencing of events in development represents a schedule that is the joint product of genetic and environmental effects. Both of the simulations reported in this chapter deal with the effects of timing.

The differences between the three ways to be innate are shown in Table 1.

**Table 1:**

*Most specific/direct* ↓ *Least specific/indirect*

| Source of constraint | | Examples in brains | Examples in networks |
|---|---|---|---|
| **Representations** | | synapses; specific microcircuitry | weights on connections |
| | ***unit*** | cytoarchitecture (neuron types); firing thresholds; transmitter types; heterosynaptic depression; learning rules (e.g., LTP) | activation function; learning algorithm; temperature; momentum; learning rate |
| **Architectures** | ***local*** | number of layers; packing density; recurrence; basic (recurring) cortical circuitry | network type (e.g., recurrent, feed-forward); number of layers; number of units in layers |
| | ***global*** | connections between brain regions; location of sensory and motor afferents / efferents | expert networks; separate input / output channels |
| **Timing** | | number of cell divisions during neurogenesis; spatio-temporal waves of synaptic growth and pruning / decay; temporal development of sensory systems | incremental presentation of data; cell division in growing networks; intrinsic changes resulting from node saturation; adaptive learning rates |

## The problem with representational innateness

Obviously, the most direct method for guaranteeing an outcome would be for the genome to specific a precise wiring plan for human cortex. Something like this appears to happen with the nematode, *C. Elegans.* This animal has exactly 959 somatic cells, and genetically identical nematodes have virtually identical patterns of cell connectivity. This is quite unlike humans. No two humans, not even monozygotic twins, have identical neuronal connections. And there is abundant reason to believe that representational nativism is simply not an option available for guaranteeing language in humans, and that the cortex of higher vertebrates (and especially humans) has evolved as an "organ of plasticity" which is capable of encoding a vast array of representations.

In a number of recent studies with vertebrates, for example, investigators have changed the nature of the input received by a specific area of cortex, either by transplanting plugs of fetal cortex from one area to another (e.g., somatosensory to visual, or vice-versa, O'Leary, 1993; O'Leary & Stanfield, 1989), by radically altering the nature of the input by deforming the sensory surface (Friedlander, Martin & Wassenhove-McCarthy, 1991; Killackey et al., 1994), or by redirecting inputs from their intended target to an unexpected area (e.g., redirecting visual inputs to auditory cortex (Frost, 1982, 1990; Pallas & Sur, 1993; Roe et al., 1990; Sur, Garraghty & Roe, 1988; Sur, Pallas & Roe, 1990; see also Molnar & Blakemore, 1991). Surprisingly, under these aberrant conditions, the fetal cortex takes on neuroanatomical and physiological properties that are appropriate for the information it receives, and quite different from the properties that would have emerged if the default inputs for that region had occurred. This suggests that cortex has far more representational plasticity than previously believed. Indeed, recent studies have shown that cortex retains representational plasticity into adulthood (e.g., radical remapping of somatosensory cortex after amputation, in humans and in infrahuman primates (Merzenich et al., 1988; Pons et al., 1991; Ramachandran, 1993; see also Greenough, Black, & Wallace, 1993).

In fact, such a situation would seem to be inevitable, given the impossible burden that a direct gene-synapse specification would impose on the genome. Calow (1976) has estimated that the adult human body contains approximately $5 \times 10^{28}$ bits of information (taking into account that cell type, spatial position, and connectivity need to be specified for each of the 100 trillion cells in the body), but the genome contains only about $1 \times 10^{9}$ bits (if it is construed as a bit-map). A better view of what genes do is provided by Bonner (1988), who suggests that much of development occurs through simple inertia of biochemical reactions which drive themselves. Genes play the role of catalysts and regulators which modulate these reactions, so their effects are typically highly indirect and opaque with regard to final outcomes. Furthermore, a very large number of genes may be involved in complex behaviors (e.g., courtship in the fruitfly, Greenspan, 1995), most of which are "re-used" and participate in many other interactions.

Thus, although representational nativism is a logical possibility, it is not likely that it plays any role in the emergence of language.

## The importance of time

If we reject representational nativism, this leads us to seek ways in which architectural and chronotopic (timing) constraints might be responsible for language. Architectural constraints are in fact very powerful, but in this chapter I wish to focus on the role played by time.

Evolutionists have long known that dramatic changes in the timing of developmental events can produce remarkable differences in outcome (e.g., Gould 1977; see also McKinney & McNamara, 1991). The dramatic distortions of body shape which D'Arcy Thompson (1961) described, involving simple allometric changes in Cartesian coordinates, easily arise from altered temporal growth gradients. In other cases, timing may alter the nature of tissue/tissue interaction and tissue induction. In adults, the length of long bones is partially determined by the number of mitotically active founder cells initially available. If the process of bone formation is delayed, these founder cells may in the interim be recruited to form other issue types and so fewer cells will be available, leading to shortened bone length. Or timing may be so altered as to lead to a loss of interactions. The formation of teeth involves a complex interaction between several embryonic tissues. In the case of birds, this interaction has been short-circuited but it can be artificially brought about by bringing together dental ectoderm from the chick and mesenchyme from a mouse (Kollar & Fisher, 1980). The genetic information necessary for tooth formation thus still seems to be present in birds (the last toothed bird dates to the Upper Cretaceous), but has been lost through a change in the timing of developmental events.

These are examples of closed systems, in which timing affects an interaction which is internal to the organism. I would like now to describe two examples in which timing plays a crucial role in enabling an outcome which otherwise would not have occurred, but in which external input from the environment is also necessary.

**The importance of starting small**

One of the most important things human children must learn to do is communicate. Language learning occupies a great deal of a child's time and it takes place over many years. The apparent inexorability of this process has led many people to conclude that there are powerful internal drives at work.

A fascinating feature of this behavior is that its form seems to be quite decoupled from its content. Manipulating words is not like manipulating a bicycle or using chopsticks or learning to walk. In these latter cases the form of the activity is directly related to its function. Language is different in this respect. It is a highly symbolic activity. The relationship between the symbols and the structures they form, on the one hand, and the things they refer to, on the other, is largely arbitrary. This too, has motivated many to seek biological explanations for the behavior, on the assumption that if the structures of language were functional, they could be learned.

Among the many peculiar features of language is the fact that while the sequence of words we speak occur in a simple linear order (one word following another), the relationships between these words are complex and often involve hierarchical organization. Thus, in the sentence *The cat who the dogs chase runs toward me*, the main thrust of sentence is that the cat is running toward me, and the fragment *who the dogs chase* is subsidiary. One way to capture the relationships between the different parts of the sentence is through a tree diagram of the sort shown in Figure 1.

This sort of tree encodes our intuitions about the relative relationship of the words in the sentence by explicitly representing their constituent structure (e.g., *the cat* and *who the dogs chase* are constituents—parts—of the top level NP, which, along with the VP, is a constituent of the top level S).

In traditional linguistic theory, such representations are supposed to do several things for us. First, a theory of meaning (semantics) should be able use these representations to determine how the meaning of the sentences is built up from its constituents, given their structural relationships. Second, these representations should provide a vocabulary for expressing important formal

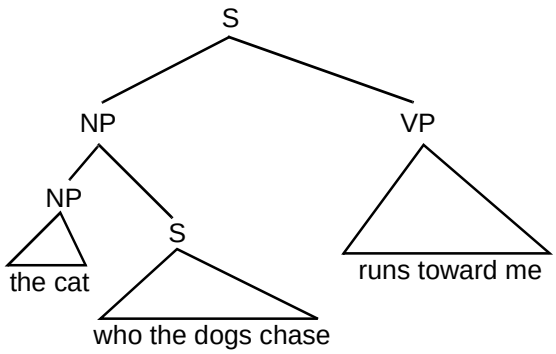



Figure 1. A simplified phrase structure tree corresponding to the sentence *The cat who the dogs chase runs toward me.* Triangles simplify additional structure; S, NP, and VP stand for sentence, noun phrase, and verb phrase.

generalizations about what sorts of structures are grammatical. A very simple but important grammatical generalization in English and many other languages is that the exact form of the verb depends on whether its subject noun is singular or plural. Thus, we rule out *The cat who the dogs chases run toward me* as ungrammatical; *cats* requires that its verb (*runs*) be in the singular, and *dogs* requires that its verbs (*chase*) be in the plural. The generalization which guarantees this agreement between noun and verb can be readily captured by the tree diagram above, because it allows us to appeal to notions of "level" or "clause"; *cats* and its verb are in the same clause, even though embedded material intervenes in the linear string.

Embedding is a basic property of human language. Whatever theory of language one adopts must provide some way to represent the complex hierarchical relationships which occur in many sentences. The ability to maintain such representations would appear to be an ideal candidate for something which must be innate in language users, and absent in non-human species.

Indeed, in a well-known mathematical proof, Gold (1967) was able to show that formal languages of the class which allow embeddings of the sort seen above cannot be learned inductively on the basis of positive input only. A crucial part of Gold's proof relied on the fact that direct negative evidence (e.g., of the explicit form in which the parent tells the child, "The following sentence, 'Bunnies is cuddly', is not grammatical") seems virtually nonexistent in child-directed speech (but see MacWhinney, 1993). Since children eventually do master language, Gold suggested that this may be because they already know critical things about the possible form of natural language. That is, learning merely takes the form of fine-tuning.

Although there are many reasons to believe that Gold's proof is actually not relevant to the case of natural language acquisition, it would be a mistake to take the extreme opposite position and claim that language learning is entirely unconstrained. Children do not seem able to learn any arbitrary language, nor are non-human young able to learn human languages. Or to return to the example at hand, what sort of constraint might permit a language user to represent abstract hierarchical relationships of the sort found in sentences? To study this question, I created an artificial language which possessed a number of characteristics that are presumably problematic (in the above sense), and attempted to teach a simple recurrent network to process them (see Elman, 1993 for full account). The artificial language had the following characteristics:

1. grammatical categories: words belonged to different categories (e.g., noun, verb, etc.);

2. basic sentence structure: simple sentences consisted of a noun followed by a verb; if the verb was transitive then a second noun followed);
3. number agreement between subject noun & verb: singular nouns required the singular form of the form; plural nouns required plural verbs;
4. verb argument structure: some verbs were transitive; others were intransitive; and others were optionally transitive
5. relative clauses: nouns could be modified by a relative clause (e.g., *who the dogs chase*); both subject relatives (*girl who sees the boy*) and object relatives (*girl who the boys see*) were possible.

The words in the language were represented by vectors in which all elements were 0 except for a single bit which was set to 1. Because these vectors are all orthogonal to each other, there was no similarity of form which the network could use to determine that a given vector was a noun or verb, or even that two vectors might be related (as in *boy* and *boys*).

The task of the network was to take one word at a time and predict what the next word would be. Since the grammar which generated the sentences was nondeterministic, any given word might be followed by a number of different possibilities. Short of memorizing the entire training corpus (which was not feasible, given the size of the corpus and the resources available to the network), the optimal strategy would be for the network to predict *all* the possible words which might occur in a given context. Thus, after having heard the sequence *the girl who the dogs see...*, the network should predict all the words which might occur in that position, namely, singular transitive verbs. But in order to do this, the network had to have identified which words were verbs, which were singulars, and which were transitive. Furthermore, and most relevant to the issue at hand, the network must have somehow learned to associate the first verb it encounters in the sequence (*see*) with the second noun it has heard (*dogs*), and that the word which follows *see* must be the verb which goes with the very first noun (*girl*), and therefore a singular. This is exactly what the sort of information tree diagrams are intended to convey. How would a network—or *could* a network—represent this information?

Since the task involved processing a sequence of information presented over time, a simple recurrent network with the architecture shown in Figure 2 was used. The recurrent connections provide the network with the memory that it needs to process the serially ordered inputs.

The results of the first trials were quite disappointing. The network failed to master the task, even for the training data. Performance was not uniformly bad. Indeed, in some sentences, the network would correctly coordinate the number of the main clause subject, mentioned early in a sentence, with the number of the main clause verb, mentioned after many embedded relative clauses. But it would then fail to get the agreement correct on some of the relative clause subjects and verbs, even when these were close together. (For example, it might predict *The boys who the girl* *******chase see the dog*, getting the number agreement of *boys* and *see* right, but failing on the more proximal—and presumably, easier—*girl chase**s***.) This failure, of course, is exactly what might have been predicted by Gold.

In an attempt to understand where the breakdown was occurring, and just how complex a language the network might be able to learn, I devised a regimen in which the training input was organized into corpora of increasing complexity, and the network was trained first with the simplest input. There were five phases in all. In the first phase, 10,000 sentences consisting solely of simple sentences were presented. The network was trained on five exposures ("epochs") to this database. At the conclusion of this phase, the training data were discarded and the network was

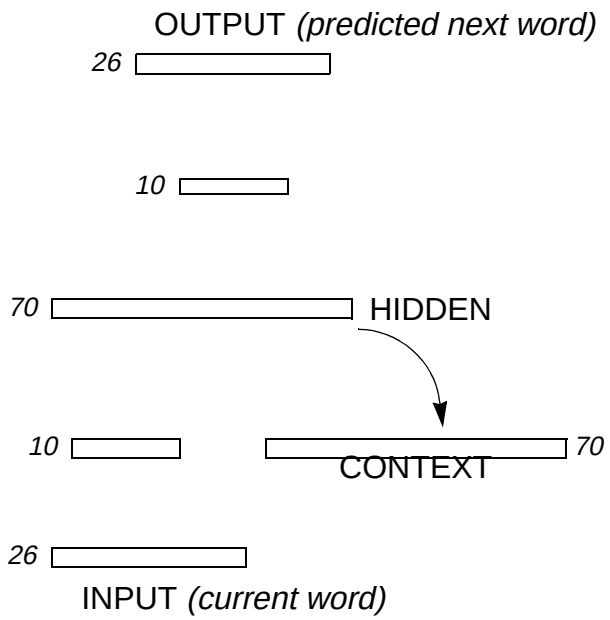



Figure 2. The simple recurrent network used in the prediction task. Rectangles represent groups of nodes; the numbers are shown adjacent to each layer. Lines with arrows indicate connections between layers, and the flow of information. Broken lines represent connections whose weights can be changed by learning; the solid line represents connections which are fixed at the value of 1.0.

exposed to a new set of sentences. In this second phase, 7,500 of the sentences were simple, and 2,500 complex sentences were also included. As before, the network was trained for 5 epochs, after which performance was also quite high, even on the complex sentences. In phase three, the mixture was 5,000 simple/5,000 complex sentences, for 5 epochs. In phase four, the mixture was 2,500 simple/7,500 complex. And in phase five, the network was trained on 10,000 complex sentences. At the conclusion of training, the network's performance was quite good, for complex as well as simple sentence. Furthermore, the network generalized its performance to novel sentences.

This result contrasts strikingly with the earlier failure of the network to learn when the full corpus was presented at the outset. Put simply, the network was unable to learn the complex grammar when trained from the outset with the full "adult" language. However, when the training data were selected such that simple sentences were presented first, the network succeeded not only mastering in these, but then going on to master the complex sentences as well.

In one sense, this is a pleasing result, because the behavior of the network partially resembles that of children. Children do not begin by mastering the adult language in all its complexity. Rather, they begin with the simplest of structures, and build incrementally until they achieve the adult language.

There is an important disanalogy, however, between the way in which the network was trained and the way children learn language. In this simulation, the network was placed in an environment which was carefully constructed so that it only encountered the simple sentences at the beginning. As learning and performance progressed, the environment was gradually enriched by the inclusion of more and more complex sentences. But this is not a good model for the situation

in which children learn language. Although there is evidence that adults modify their language to some extent when interacting with children, it is not clear that these modifications affect the grammatical structure of the adult speech. Unlike the network, children hear exemplars of all aspects of the adult language from the beginning.

If it is not true that the child's environment changes radically (as in this first simulation), what is true is that the *child* changes during the period he or she is learning language. A more realistic network model would have a constant learning environment, but some aspect of the network itself would undergo change during learning. One candidate for a developmental change which might interact with learning is working memory; working memory and attention span in the young child are initially limited, and increase over time. Could such changes facilitate learning?

In order to study a possible interaction between learning and changes in working memory, another new network was trained on the "adult" (i.e., fully complex) data which had initially been problematic. This time, at the outset of learning, the context units (which formed the memory for the network) were reset to random values after every two or three words. This meant that the temporal window within which the network could process valid information was restricted to short sequences. The network would of course see longer sequences, but in those cases the information necessary to make correct predictions would fall outside the limited temporal window; such sequences would effectively seem like noise. The only sequences which would contain usable information would in fact be short, simple sentences. After training the network in this manner for a period of time, the "working memory" of the network was extended by injecting noise into the context units at increasingly long intervals, and eventually eliminating the noise together.

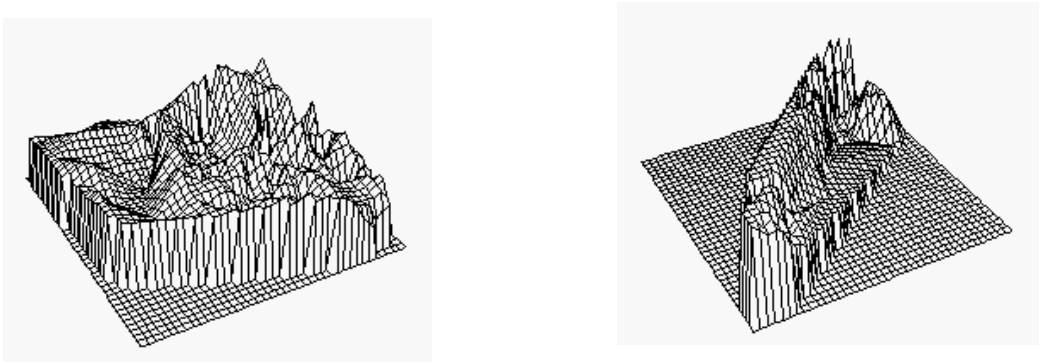
Under these conditions, the performance at the conclusion of training was just as good as when the training environment had been manipulated. Why did this work? Why should a task which could not be solved when starting with "adult" resources be solvable by a system which began the task with restricted resources and then developed final capacities over time?

It helps to understand the answer by considering just what was involved when learning was successful. At the conclusion of learning, the network had learned several things: distinctions between grammatical categories; conditions under which number agreement obtained; differences between verb argument structure; and how to represent embedded information. As was the case in the simulation involving simple sentences, the network uses its internal state space to represent these distinctions. It learns to partition the state space such that certain spatial dimensions signal differences between nouns and verbs, other dimensions encode singular vs. plural, and other dimensions encode depth of embedding.

In fact, we can actually look at the way the network structures its internal representation space. Let us imagine that we do the equivalent of attaching electrodes to the network which successfully learned the complex grammar, by virtue of beginning with a reduced working memory. If we record activations from this network while it processes a large number of sentences, we can plot the activations in a three-dimensional space whose coordinates are the principal components of hidden unit activation space (we shall use the second, third, and eleventh principal components). The plot shown in Figure 3(b) shows the regions of this space which are used by the network.

As can be seen, the space is structured into distinct regions, and the patterning is used by the network to encode grammatical category and number. Once the network has developed such a representational scheme, it is possible for it to learn the actual grammatical rules of this language. The representations are necessary prerequisites to learning the grammar, just because these internal representations are also play a role in encoding memory (remember that the hidden unit acti-

vation patterns are fed back via the context units). Without a way to meaningfully represent the (arbitrarily encoded) inputs, the network does not have the notational vocabulary to capture the grammatical relationships. Subjectively, it's the same problem we would have if we try to remember and repeat back words in an unfamiliar language-it all sounds like gibberish. Note that this creates a bit of a problem, however. If the network needs the right internal representations to work with, where are these to come from? The truth is that these representations are learned in the course of learning the regularities in the environment. It learns to represent the noun/verb distinction because it is grammatically relevant. But we just said it couldn't learn the grammar without having the representations. Indeed, this chicken and egg problem is exactly the downfall of the network which starts off fully developed (but lacking the right representations). If we look at the internal space of this network after (unsuccessful) training, shown in Figure 3(a) we see that the



(a) (b)

Figure 3. View of hidden unit space (in three of 70 dimensions) of a network which fails to learn the grammar (a), and which succeeds (b). The surfaces are plotted by passing a large number of test sentences through each network and recording the hidden unit activation vector following each word. In the case of successful learning, the hidden unit state space is structured and can be interpreted in terms of various dimensions of relevance for the task (e.g., noun vs. verb, singular vs. plural, etc.). In the case of the unsuccessfully trained network, the state space is poorly organized and no clearly interpretable dimensions are found.

space is poorly organized and not partitioned into well-defined areas. The network which starts off with limited resources, on the other hand, actually is at an advantage. Although much of what it sees is now "noise," what remains—the short, simple sentences—are easier to process. More to the point, they provide a tractable (because they are short, and impose fewer demands on a well-developed representation/memorial system) entry point into the problem of discovering the grammatical patterns and categories latent in the environment. Once these categories have been induced, they provide the bootstrap by which the network can go on, as its working memory improves, to deal with increasingly complex inputs and refine its knowledge.

Seen in this light, maturational limitations take on a very positive character. If a domain to be mastered is complex, it helps to have some clues about where to start. Certainly the solution space for inducing a grammar from the data is extremely large, and finding the right grammar might be an intractable problem. It makes sense therefore that children (or networks) might need cues to help guarantee they discover the right grammar. The question is, what do these cues look like?

One possibility is that children (or networks) might be pre-wired in such a way that they

know about concepts such as “noun” and “verb” at birth. We might endow them as well with special knowledge about permissible classes of structures, or grammatical operations on those structures. The role of experience would be to help the learner figure out which particular structures or operations are true of the language being learned. This is the hypothesis of Parameter Theory (Chomsky, 1981).

The simulation here suggests another solution to the problem of finding the needle in the grammatical haystack. Timing the development of memory has the effect of limiting the search space in exactly the right sort of way as to allow the network to solve a problem which could not be solved in the absence of limitations.

Is there any evidence that this positive interaction between maturational limitations and language learning plays a role in children, as it seems to in networks? Elissa Newport has suggested that indeed, early resource limitations might explain the apparent critical period during which languages can be learned with native-like proficiency. Newport calls this the “less is more” hypothesis (Newport, 1988, 1990).

It is well-known that late learners of a language (either first or second) exhibit poorer performance, relative to early or native learner. What is particularly revealing is to compare the performance of early (or native) learners when it is at a comparable level to that of the late learners (i.e., early on, while they are still learning). Although gross error scores may be similar, the nature of the errors made by the two groups differs. Late learners tend to have incomplete control of morphology, and rely more heavily on fixed forms in which internal morphological elements are frozen in place and therefore often used inappropriately. Young native learners, in contrast, commit errors of omission more frequently. Newport suggests that these differences are based in a differential ability to analyze the compositional structure of utterances, with younger language learners at an advantage. This occurs for two reasons. Newport points out that the combinatorics of learning the form-meaning mappings which underlie morphology are considerable, and grow exponentially with the number of forms and meanings. If one supposes that the younger learner is handicapped with a reduced short-term memory, then this reduces the search space (because the child will be able to perceive and store a limited number of forms). The adult’s greater storage and computational skills work to the adult’s disadvantage. Secondly, Newport hypothesizes that there is a close correspondence between perceptually salient units and morphologically relevant segmentation. With limited processing ability, one might expect children to more attentive to this relationship than adults, who might be less attentive to perceptual cues and more inclined to rely on computational analysis. Newport’s conclusions are thus very similar to what is suggested by the network performance: there are situations in which maturational constraints play a positive role in learning. Counterintuitively, some problems can only be solved if you start small. Precocity is not always to be desired.

The starting small/less is more hypotheses suggest a new interpretation to the “critical period” phenomenon. Many people have interpreted the fact that language-learning occurs with greatest success (e.g., learners achieve native fluency) during childhood as evidence for a Language Acquisition Device which operates only during childhood. Once its job is done, it ceases to function. But the simulation here, and Newport’s hypothesis, suggest rather that the ability which children have for learning language derives not from a special mechanism which they possess and adults do not, but just the reverse. It is children’s *lack* of resources which enables them to learn languages fluently.

Finally, how do these hypotheses bear on the issue of innateness? If in fact developmental limitations of the sort discussed here can impose constraints which are crucial for achieving a tar-

get behavior, and these developmental limitations arise from biological factors, then we may say that the network described here is "innately constrained" to discovering the proper grammar. But note that this is a very different sort of innateness than envisioned by the pre-wired linguistic knowledge hypothesis.

**How does the cortex get its architecture?**

One of the arguments I advanced earlier against the hypothesis of representational innateness (i.e., direct specification of cortical microciruity) rested on experimental data which suggest that the regional mapping of functions in the human cortex is not prespecified. Initially, the cortex appears to possess a high degree of pluripotentiality. Over time, however, a complex pattern of spatially localized regions develops, and the pattern of localization is relatively consistent across individuals. The mystery is how the specific functional organization of the cerebral cortex arises. Shrager and Johnson (1996) and Rebotier and Elman (1996), building on earlier work by Kerszberg, Dehaene and Changeux (1992), have offered a preliminary account of at least one factor which might provide an answer to this question. Let me describe these simulations.

Shrager and Johnson began with the assumption that the cortex is organized through a combination of endogenous and exogenous influences, including subcortical structuring, maturational timing, and the information structure of an organism's early environment. Their goal was to explore ways in which these various factors might interact in order to lead to differential cortical function and to the differential distribution of function over the cortex. They began with several simple observations.

First, Shrager and Johnson pointed out that although there are signals which pass through the cortex in many directions, subcortical signals (e.g., from the thalamus) largely feed into primary sensory areas, which then largely feed forward to various secondary sensory areas, leading eventually into the parietal and frontal association areas. Each succeeding large-scale region of cortex can be thought of as processing increasing higher orders of invariants from the stimulus stream. The image is that of a cascade of filters, processing and separating stimulus information in series up toward the integration areas.

Second, Shrager and Johnson noted that a very striking aspect of development of the cerebral cortex is the initial overproduction and subsequent loss of neural connections, resulting in the relatively sparsely interconnected final functional architecture. This process of overproduction of synapses and subsequent (or simultaneous) thinning out of the arbor is thought to be key in cortical ontogeny. As Thatcher (1992) suggests, when initially heavily connected, the cortex is like a lump of stone which in the hands of the sculptor is shaped by removal of bits and pieces into its final form. But curiously, this sculpting does not occur everywhere simultaneously. Instead, there appears to be a general developmental dynamic in which, grossly speaking, the locus of maximum neural plasticity begins in the primary sensory and motor areas and moves toward the secondary and parietal association areas, and finally to the frontal regions (Chugani, Phelps, & Mazziotta, 1987; Harwerth, Smith, Duncan, Crawford, & von Noorden, 1986; Pandya & Yeterian, 1990; Thatcher, 1992). Thus there is a parallelism between the final architecture of the cortex, in which information proceeds from sensory to secondary to association areas, and the dynamics of cortical development, which also proceeds from sensory to secondary to association areas.

Given these observations, Shrager and Johnson posed the question, How might such a developmental wave of plasticity—in which different regions of cortex are more plastic at different points in time—affect the outcome of learning? To study this question, Shrager and Johnson developed a connectionist network which was designed to test the hypothesis that under certain

regimes of wave propagation, we might expect a tendency toward the development of higher order functions in later parts of the cortical matrix. In this way, the model might account for spatial distribution of function in cortex without having to encode the localization directly.

The Shrager and Johnson model is shown in Figure 4. The model consists of an abstract

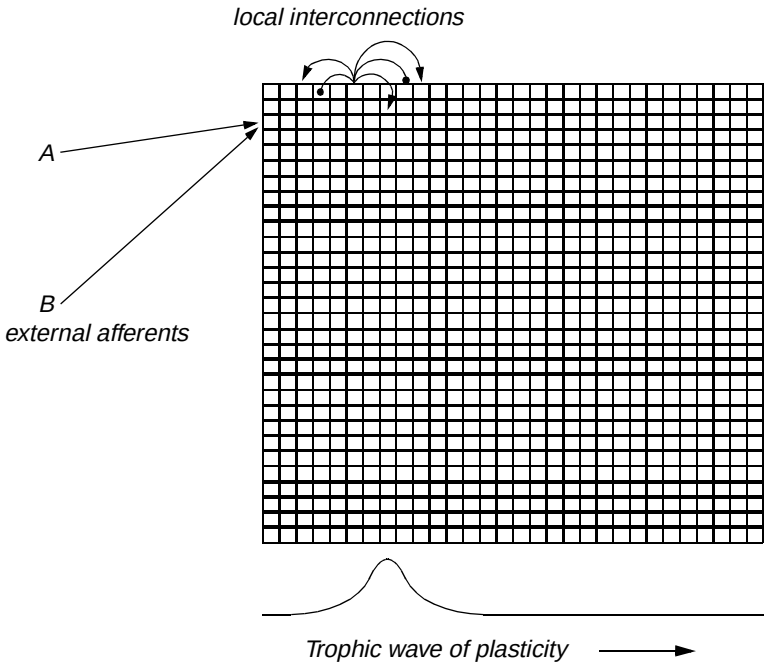



Figure 4. Shrager & Johnson model. Each unit has short local connections (excitatory and inhibitory) to close neighbors, and also receives afferents from the external afferents, A and B (shown here as excitatory, but initially set as excitatory or inhibitory at random). In some simulations, a tropic wave of plasticity spreads from left to right across the matrix and has the effect of modulating learning in the columns under this wave.

"cortical matrix" composed of a 30 by 30 matrix of artificial neurons. Each neuronal unit has afferent and efferent connections to nearby units, and also receives afferent inputs from external signals designated A and B. For our purposes, we shall consider the case in which afferent and efferent connection weights are initially set at random, and in which the external signals A and B provide simultaneous inputs of 0 and 1, also at random.

The matrix weights are changed according to a Hebbian learning rule, so that connection strength grows between units whose activations are more highly correlated. Under the default conditions just described, the outcome after learning is that some units become active only when their A input is on; others become sensitive only to the presence of the B input; and still others become sensitive to A and B simultaneously (logical AND), or to either A or B (logical OR). A very large number of units are always off. No units develop which are sensitive to exclusive OR (XOR), which is not surprising since Hebbian learning does not typically lead to such higher order functions.

Shrager and Johnson then considered what might happen if the Hebbian learning is modulated by a trophic factor (TF) which passed through the matrix in a wave, from left to right. The effect of wave was that the columns of units underneath it, at any given point in time, were more

plastic and therefore able to learn. During the first training cycle, for example, a modulation vector was produced for the 30-column matrix that might be [1.0, 0.86, 0.77, 0.66, 0.53,…,0.0, 0.0]. That is, TF transmission at location 1 in the matrix took place normally, whereas TF transmission at location 2 was reduced to 86% of what would have been moved, etc. On the next cycle, the wave moved to the right a small amount: [0.86, 1.0, 0.86, 0.77, 0.66,…,0.0, 0.0]. The progress of the wave thus modulated the transmission of trophic factor, leading to a dynamic plasticity in the cortical matrix. Leftward columns were plastic early and also lost their plasticity early on; whereas, rightward columns did not become plastic until later on, but were plastic toward the end of the simulation when most of the neurons were reaching asymptote on the stabilization and death curves.

Under certain regimes of wave propagation, Shrager and Johnson expected to observe a tendency toward the development of higher order functions in the cortical matrix. (Higher order functions are those which depend on both A and B inputs; lower order functions are those which depend solely on A or B.) The reason for this may be envisioned by considering two steps in the propagation of the wave from some leftward set of columns to the next set of columns to the right. We shall call these columns COL1 and COL2 (which is immediately to the right of COL1). COL1, initially more plastic than COL2, determines its function during receipt of input from A and B afferents, as has been the case all along. However, COL1 becomes fixated in its function relatively early, as the wave moves on to COL2. Now, however, COL2 is receiving input that is, in addition to the input coming from A and B afferents, includes the combined functions fixated by the earlier plasticity in COL1. Thus, COL2 has, in effect, three afferents: A, B, and COL1.

In fact, Shrager and Johnson found that the number of first order functions (A, ~A, B, and ~B) differed significantly from the number of second order functions (B-AND-~A, A-AND-~B, A-XOR-B, ~[A-AND-B], A-AND-B, A=B, A>=B, A<=B, and A-OR-B), when the wave was present, but not without the wave. Furthermore, as predicted, the density of higher order functions increased in regions of the matrix which were plastic later on, as determined by the propagation of the TF wave. Finally, when the propagation rate of the wave was tripled from the initial rate, a different picture emerged. Again, the first and second order functional densities were significantly different, but this time the mean values were inverted. In the slow wave case the second order functions were emphasized, whereas in the in fast wave case the first order functions were emphasized.

There is another result which is of great significance, and was the focus of a replication and extension by Rebotier and Elman (1996). This result has to do with the problem of how to reconcile the desire for a learning rule which is both biologically plausible, and sufficiently powerful. On the one hand, Hebbian learning has a greater biological plausibility than back propagation learning. Also, Hebbian learning is a form of self-organizing behavior, which is attractive because it means an explicit teacher is not required (as in backpropagation). On the other hand, Hebbian learning cannot be used to learn certain important problems. These include XOR and other functions in which classification cannot be done on the basis of correlations. This is unfortunate, because it means that the learning mechanism which is most natural on biological grounds seems to lack necessary computational properties.

Rebotier and Elman constructed a network of the form Shrager and Johnson devised and allowed Hebbian learning to take place through all parts of the network (“instant cortex”). Not surprisingly, Rebotier and Elman found no units which respond to the XOR of the inputs A and B. Rebotier and Elman then repeated the experiment, but this time allowed learning to be modulated by a spatial wave of trophic factor, which passed over the network from left to right. This time, a

small percentage of units were found which computed XOR. These units tended to be on the right side of the network (i.e., the late maturing regions). The reason they could compute XOR is that they did not learn until later, after early units had developed with learn simpler functions such as AND and OR. These early learning units then became additional inputs to the later learning units. Since XOR can be decomposed into the AND and OR functions, this made it possible to learn a function which could not otherwise have been learned.

There are thus two important lessons to be learned from the Shrager and Johnson and the Rebotier and Elman studies. First, the models demonstrate how the differential functional architecture of the cortex might arise in early development as an emergent result of the combination of organized stimulus input, and a neurotrophic dynamic (whether produced by a natural wave of trophic factor or by some other endogenous or exogenous phenomenon). Second, development provides the key to another puzzle. The studies show how some complex functions which are not normally learned in a static mature system can be learned when learning is carried out over both time and space rather than occurring everywhere simultaneously.

## A conspiracy theory of language

I began at the outset with a set of questions, and I would like to return to them now, if not to provide answers, at least to say how the above simulations suggest we might think about what kinds of answers are likely.

The questions—about species uniqueness, the form of language, language learning, universals and variation—might be answered by simply stipulating that language is an innate property of our species, and takes the form it does "just because it does." This is not only not a very illuminating answer, but to the extent that it relies on representational innateness, is also highly implausible. Yet language does emerge only in our species; it does assume a constrained set of forms; and patterns of acquisition and usage are remarkably similar across languages. How might we account for this?

The two simulations above (and considerable other evidence discussed in detail in Elman et al., 1996) suggests what might be called the *Language as conspiracy* view. This view is in fact consistent with two very robust findings in the embroylogical and developmental genetics literature: (1) the nonlinear effects of small developmental changes on outcome; and (2) the conservative nature of the genome and the importance of interactions in development.

At the turn of the century, the naturalist D'Arcy Thompson published a now classic treatise called *On Form and Growth.* Thompson pointed out that relatively simple transformations of the Cartesian coordinates underlying body plans could produce dramatic differences in body morphology. Thus, the skulls of the human, chimpanzee, baboon, and dog, bear a striking resemblance once the transformation is made apparent (see Figure 5). Thompson suggested that what appear to be large morphological differences in species might be misleading, in that they involve far simpler changes in growth.

We know now that in fact there are a variety of developmental mechanisms which can accomplish transformations of the sort described by Thompson (see McKinney & McNamara, 1991, for an extensive review). It is clear that small changes in a development trajectory can indeed lead to very great differences in outcome. Earlier, in discussing the role of timing, I gave the examples of long-bone growth. Similar accounts have been offered for a variety of other

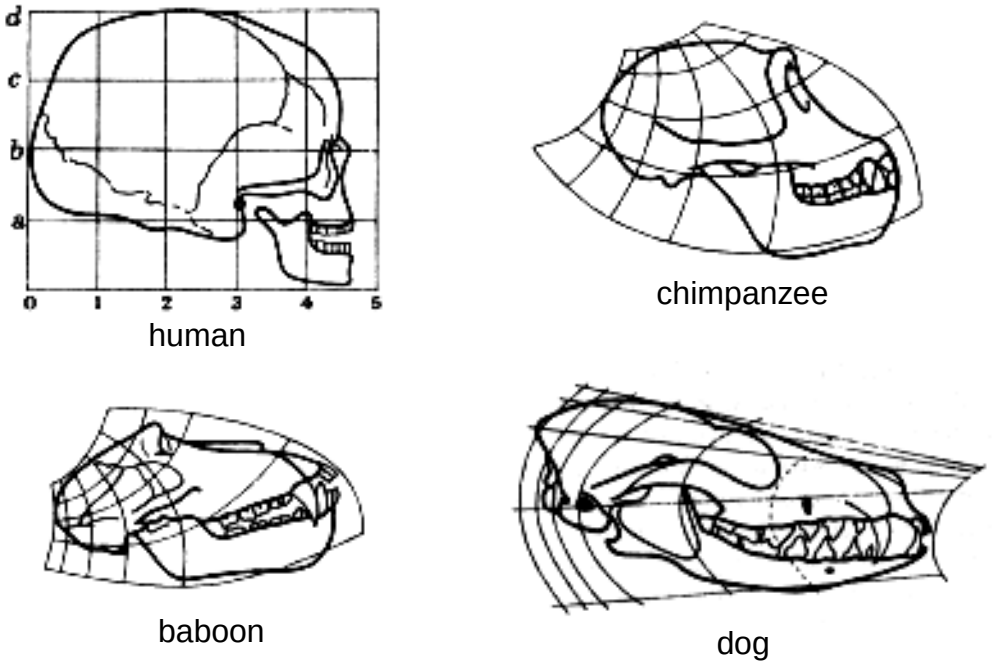



Figure 5. Skulls of human, chimpanzee, baboon, dog, drawn with respect to a single Cartesian coordinate system. Differences in skull size and shape can be produced by transformations on the coordinates. From Thompson (1971), p. 318, 319, 322.

changes associated with speciation. Flightlessness, for example, is common among birds which live in islands without large mammalian predators; maintenance of the bone and muscle mass necessary for flight are energetically expensive and will be selected against unless there is some adaptive advantage to flight. Some groups of birds such as rails have evolved to delay sternum formation until relatively late in development. Delaying this important developmental process until after hatching pre-adapts the group so that further changes leading to flightlessness (in environments where this is advantageous) are easy to achieve.

An even more dramatic example concerns the process of tooth formation (or lack thereof) in the modern bird. The formation of teeth involves a complex interaction between several embryonic tissues. One layer of tissue (epithelium) must be brought into contact with another layer of tissue (mesenchyme). The mesenchyme induces the epithelium to differentiate into an enamel producing organ; the organ-producing epithelium then induces the mesenchyme to differentiate into tissue which secretes dentin. In the absence of this second interaction, the mesenchyme would become spongy bone. Birds are known to descend from ancestral species that possessed teeth, but such toothed birds have not been seen since the Upper Cretaceous. Does this mean the genetic information necessary to form teeth has been lost in birds and replaced by “beak-forming genes?” Apparently not. Rather, it seems that in birds this interaction has merely been short-circuited. The interaction can be artificially brought about by bringing dental epithelium from the chick into contact with mesenchyme from a mouse (Kollar & Fisher, 1980). Under these conditions, the chick epithelium will form enamel organs, and further interactions may lead to forma-

tion of complete teeth.

This last example also illustrates the second major lesson concerning development: In mammals, most important developmental phenonena rest on a complex set of interactions; these include virtually every possible interaction imagineable, e.g., gene/gene, gene/environment, tissue/tissue, tissue/environment, organ/organ, organ/environment, etc.The early view, for example, that complex behaviors might directed by single genes has given way, over the past several decades, to the realization that even apparently simple traits such as eye color reflect the coordinated interaction of multiple genes. For more complex traits, the number of genes involved may figure in the thousands. Furthermore, genes typically play multiple roles, participating in the formation of very different traits.

As an example of this last point, consider courtship in the fruitfly, *Drosophila melanogaster.* Courtship involves at least six distinct phases, each with a different set of behaviors.A great deal has been discovered about the mechanisms which are required for the sequence of behaviors which lead to successful mating. The total repertoire depends on nine or more different regions of the central nervous system. The genetic basis for the behaviors is also beginning to be worked out, with the discovery that the genes which are involved in courtship also play a role in other behaviors. Thus, the *period* gene, which is involved in controlling the rhythm of the courtship song, also plays a role in regulating the fly's circadian rhythms. Other aspects of the courtship require that the male respond adaptively to the female's behaviorl; the *CaMKII* and *eag* genes which are known to play a role in learning and memory in the fruitfly, are then called into play (Greenspan, 1995; see Hall, 1994, for a full review of the genetic basis of courtship behavior).

The complex interactions and the importance of genes as regulators may occasionally give rise to what looks like a one gene-one trait relationship. For instance, two (of the many) species of fruitfly found only in Hawaii differ minimally (and are in fact interfertile), primarily in head shape. One species, *D. silvestris*, has the normal round-shaped head. The other, *D. heteroneura*, has a bizarrely shaped head that looks like a hammerhead shark. This difference is mostly associated with a single gene—but this gene is involved with complex epistatic interactions with three or four other genes and it is a quantitative change in the interaction which gives rise to a qualitative change in trait (Val, 1976).

Or consider the recent discovery of a family in Costa Rica which has a family history, going back over two hundred years, of acquired deafness (Lynch et al., 1997). The deafness onsets around puberty and leads to hearing loss, initially in the low-frequency range but eventually becoming total. Because of the very high family incidence and the long family history (and the total lack of incidence in the family's village), a genetic basis for the disorder was sought and eventually found. It turned out that the deafness could be attributed to a mutation in a single gene. The mutation's effect was that the last 52 of the 1,265 amino acids coded for by the gene were incorrectly specified.

A gene for deafness? Not at all. A nearly similar gene called *Diaphonous* is also found in the fruitfly. The *diaphanous* gene produces a protein which controls the assembly of actin. Actin is one of the most prevalent proteins found in the body; it organizes the tiny fibers found in cell plasma which determine a cell's structural properties (rigidity, ability to move, to deform, etc.) It seems likely that the mutation was selectively producing deafness only because the hair cells in the ear are particularly sensitive to loss of stiffness. Since the mutation was slight, the degenerate form of the protein was sufficient for most of its uses in the rest of the body. It was only in the hair cells that the deficiency could not be tolerated.

The recurring lesson, whenever one looks at complex phenotypic traits in mammals, is

that the traits are produced by a sometimes large number of interactions. The underlying genetic substrate is enormously conservative, evolutionarily. What makes innovation possible is that the interactions are sufficiently complex and that small alterations in developmental pathways can lead to very large differences in outcome.

Seen in this light, we should doubt that the novelty of language lies in having evolutionary and developmental origins which differ radically from those underlying communicative behaviors in similar species. Rather, is makes sense to view language as a behavior which results from allometric transformations (a la D'Arcy Thompson) over a set of behaviors which are present as well in other closely related species. Language is simply the result of a number of tweeks and twiddles, each of which may in fact be quite minor, but which in the aggregate and through interaction yield what appears to be a radically new behavior. It is in this sense that language is a conspiracy.

Of course, in very significant ways, language *is* a radically new behavior. At a phenomenological level, it is quite unlike anything else that we (or any other species) do. It has features which are remarkable and unique. The crucial difference between this view and the view of language as a separable domain-specific module (in the sense of Tooby & Cosmides, 1992) is that the uniqueness emerges out of an interaction involving small differences in domain-nonspecific behaviors.

If this view of language as conspiracy is correct, then it should be possible to list in detail the behaviors which participate in the conspiracy. We should be able to identify the ways in which the human version of those behaviors differ from that in other species. And we should ultimately be able to formulate a theory of interaction which provides an account not only for human language but for other non-human primate communication systems. They too are unique, in their own ways. While we are probably far from having such a full account, I believe that the simulations offered here illustrate such an account's viability.

## References

Bonner, J. T. (1988). *The Evolution of Complexity.* Princeton, NJ: Princeton University Press.

Calow, P. (1976). *Biological Machines: A Cybernetic Approach to Life*. London: Arnold.

Chomsky, N. (1981). *Lectures on Government and Binding.* New York: Foris.

Chugani, H.T., Phelps, M.E., & Mazziotta, J.C. (1987). Positron emission tomography study of human brain functional development. *Annals of Neurology, 22*, 487-497.

Thompson, D. T. (1961). *On Growth and Form.* Cambridge: Cambridge University Press.

Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D., Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development.* Cambridge, MA: MIT Press.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Friedlander, M.J., Martin, K.A.C., & Wassenhove-McCarthy, D. (1991). Effects of monocular visual deprivation on geniculocortical innervation of area 18 in cat. *The Journal of Neuroscience, 11*, 3268-3288.

Frost, D.O. (1982). Anomalous visual connections to somatosensory and auditory systems following brain lesions in early life. *Brain Research, 255(4)*, 627-635.

Frost, D.O. (1990). Sensory processing by novel, experimentally induced cross-modal circuits. *Annals of the New York Academy of Sciences, 608*, 92-109; discussion 109-12.

Gold, E.M. (1967). Language identification in the limit. *Information and Control, 16*, 447-474.

Gopnik, M., & Crago, M.B. (1991). Familial aggregation of a developmental language disorder. *Cognition, 39*, 1-50.

Gould, S.J. (1977). *Ontogeny and Phylogeny.* Cambridge, MA: Harvard University Press.

Greenough, W.T., Black, J.E., & Wallace, C.S. (1993). Experience and brain development. In M. Johnson (Ed.), *Brain Development and Cognition: A Reader* (pp. 290-322). Oxford: Blackwell

Greenspan, R.J. (1995). Understanding the genetic construction of behavior. *Scientific American*, (April, 1995), 72-78.

Hall, J.C. (1994). The mating of a fly. *Science,* 264, 1702-1714

Harwerth, Smith, Duncan, Crawford, & von Noorden, 1986 13

Kerszberg, M., Dehaene, S., & Changeux, J.P. (1992). Stabilization of complex input output functions in neural clusters formed by synapse selection. Neural Networks, 5(3), 403-413.

Killackey, H.P., Chiaia, N.L., Bennett-Clarke, C.A., Eck, M., & Rhoades, R. (1994). Peripheral influences on the size and organization of somatotopic representations in the fetal rat cor-

tex. *Journal of Neuroscience, 14,* 1496-1506.

Kollar, E.J., & Fisher, C. (1980). Tooth induction in chick epithelium: Expression of quiescent genes for enamel synthesis. *Science,* 207, 993-995.

Lynch, E.D., Lee, M.K., Morrow, J.E., , P.L., León, P.E., & King, M. (1997). Nonsyndromic Deafness DFNA1 Associated with Mutation of a Human Homolog of the *Drosophila* Gene *diaphanous. Science,* 278, 1315-1318.

E. D. Lynch, M. K. Lee, J. E. Morrow, P. L. Welcsh, P. E. León, M.

King

MacWhinney, B. (1993). Connections and symbols: Closing the gap. *Cognition, 49(3),* 291-296.

Marchman, V. (1993). Constraints on plasticity in a connectionist model of the English past tense. *Journal of Cognitive Neuroscience, 5(2),* 215-234.

McKinney, M.L., & McNamara, K.J. (1991). *Heterochrony: The Evolution of Ontogeny.* New York and London: Plenum Press.

Merzenich, M.M., Recanzone, G., Jenkins, W.M., Allard, T.T., & Nudo, R.J. (1988). Cortical representational plasticity. In P. Rakic & W. Singer (Eds.), *Neurobiology of Neocortex* (pp. 41-67). New York: John Wiley & Sons.

Molnar, Z., & Blakemore, C. (1991). Lack of regional specificity for connections formed between thalamus and cortex in coculture. *Nature, 351 (6326),* 475-7.

Newport, E.L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences, 10,* 147-172.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science, 14,* 11-28.

O'Leary, D.D. (1993). Do cortical areas emerge from a protocortex? In M. Johnson (Ed.), Brain Development and Cognition: A Reader (pp.323-337). Oxford: Blackwell Publisher

O'Leary, D.D., & Stanfield, B.B. (1989). Selective elimination of extended by developing cortical neurons is dependent on regional locale: Experiments utilizing fetal cortical transplants. Journal of Neuroscience, 9(7), 2230-2246.

Pallas, S.L., & Sur, M. (1993). Visual projections induced into the auditory pathway of ferrets: II. Corticocortical connections of primary auditory cortex. Journal of Comparative Neurology, 337(2), 317-33. Pandya & Yeterian, 1990 13

Pinker, S. (1994). The Language Instinct: How the Mind Creates Language. New York: William Morro

Pons, T.P., Garraghty, P.E., Ommaya, A.K., Kaas, J.H. Taub, E., & Mishkin M. (1991). Massive cortical reorganization after sensory deafferentation in adult macaques [see comments]. Science, 252(5014), 1857-60.

Ramachandran, V.S. (1993). Behavioral and magnetoencephalographic correlates of plasticity in the adult human brain. Proceedings of the National Academy of Sciences, 90, 10413-

10420.

Rebotier, T.P., & Elman, J.L. (1996). Explorations with the dynamic wave model. In D. Touretzky, M. Mozer, & M. Haselmo (Eds.), Advances in Neural Information Processing Systems 8. Cambridge, MA: MIT Press. Pp. 549-556.

Roe, A.W., Pallas, S.L., Hahm, J.O., & Sur, M. (1990). A map of visual space induced in primary auditory cortex. Science, 250 (4982), 818-20.

Shrager, J., & Johnson, M.H. (1996). Dynamic plasticity influences the emergence of function in a simple cortical array. Neural Networks, 8, 1-11.

Sur, M., Garraghty, P.E., & Roe, A.W. (1988). Experimentally induced visual projections into auditory thalamus and cortex. Science, 242, 1437-1441.

Sur, M., Pallas, S.L., & Roe, A.W. (1990). Cross-modal plasticity in cortical development: differentiation and specification of sensory neocortex. Trends in Neuroscience, 13, 227-233.

Thatcher, R.W. (1992). Cyclic cortical reorganization during early childhood. Brain and Cognition, 20, 24-50.

Tooby, J. & Cosmides, L. (1992). The psychological foundations of culture. In *The adapted mind: Evolutionary Psychology and the Generation of Culture* (Ed. J. Barkow, L. Cosmides, & J. Tooby). NY: Oxford University Press. Pp. 19-136.

Val, F.C. (1976). Genetic analysis of the morphological differences between two interfertile species of Hawaiian Drosophila.*Evolution*, 31, 611-620.

Vargha-Khadem, F., Watkins, K., Alcock, K., Fletcher, P., & Passingham (1995). Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. Proceedings of the National Academy of Sciences USA, 92, 930-933.