WHY A BRAIN CAPABLE OF LANGUAGE EVOLVED ONLY ONCE: PREFRONTAL CORTEX AND SYMBOL LEARNING

by Terrence W. Deacon

Abstract. Language and information processes are critical issues in scientific controversies regarding the qualities that epitomize humanness. Whereas some theorists claim human mental uniqueness with regard to language, others point to successes in teaching language skills to other animals. However, although these animals may learn names for things, they show little ability to utilize a complex framework of symbolic reference. In such a framework, words or other symbols refer not only to objects and concepts but also to sequential and hierarchical relationships with other symbols. This process is essential to human mental operations, including language, mathematics, and music. In humans, these operations may have coevolved with the prefrontal area of the cerebral cortex, which is proportionately much larger in humans than in other animals and more intricately linked with other areas of the brain. Analysis of the structure and function of the prefrontal area suggests that it is centrally involved in the operation of higher-order associative relationships involving the subordination of one set of associations to another. This alternate learning strategy apparently appeared at the cost of certain sensory, motor, or limbic abilities. The payoff was symbolic thinking. Humans thus are unique among species, not just for their highly developed language ability but for their odd style of thinking and learning.

Keywords: brain; coevolution; dualism; information processes; language; prefrontal cortex; mind; symbolic reference.

Terrence W. Deacon is Associate Professor of Biological Anthropology at Boston University and Lecturer in Psychiatry at Harvard Medical School. His mailing address is Neuroregeneration Laboratory, Mailman Research Center, MacLean Hospital, MRC 101, 115 Mill Street, Belmont, MA 02178. This paper, except for the "Preface for *Zygon* Readers," will appear in *Human by Nature: Origins and Destiny of Language*, edited by V. Velichrovsky and D. M. Rumbaugh (Hillsdale, N.J.: Lawrence Erlbaum Publishers, in press).

[*Zygon*, vol. 31, no. 4 (December 1996).]
 ISSN 0591-2385

PREFACE FOR *ZYGON* READERS

Discussions of the relationship between biology and human values have a tendency to take place against a dualistic background, even if only in the form of disclaimers. In our scramble to scrap this vestige of a prescientific philosophy, however, we seem far too eager to see it disappear by being explained away and too reticent to address the phenomena that suggested it in the first place. Descartes was not introducing a novel worldview but attempting to base an ancient bit of common sense rigorously in a logical foundation. That commonsense view, implicit in everyday speech, expressed in belief systems, and woven into children's stories in societies around the world extending long into the hidden past, is that human beings originate and take responsibility for their information processes and behaviors in ways that appear to be absent in other species.

For Descartes, the dichotomy between mind and mechanism was intimately bound up with the distinction between humans and animals. Along with a perfunctory disavowal of Cartesian dualism at the beginning of most modern treatments of mind and brain functions, there is often an implicit denial of the animal-human dichotomy as well. For most, the difference is seen merely as a matter of degree. In a phylogenetic sense and even a neurological sense this is indeed accurate. Neurobiologists demonstrate extensive similarities between animal and human brains and their components, and geneticists emphasize that we are part of a phylogenetic continuum, sharing 99 percent DNA sequence similarity with our nearest ape relatives, the chimpanzees. Claims for human uniqueness cast in subtle more-or-less terms appear to contradict the categorical distinction implicit in the Cartesian view. This, however, must be contrasted with linguists' claims of human mental uniqueness with respect to language. Here we find many prominent theorists suggesting that the crucial difference is the possession of a unique neural structure conferring language abilities on humans alone. The presence or absence of this language device in the brain would presumably be the qualifying factor distinguishing human from nonhuman minds. In summary, we are offered either the denial of a categorical difference or the claim that it can be reduced to algorithms in a yet-to-be-identified black box. I must agree with the philosopher John Searle's critical reflection on these trends, that all do a better job of avoiding the question of the mind-brain relationship than answering it. Neither claim offers any foundations upon which to build a theory of human values. So despite the fact that nowadays we are convinced that brains produce minds and that animals also have mental lives of some sort, and that the Cartesian dualistic account is therefore in error, there is little agreement about what part is the baby to be retained and what part is the bathwater to be discarded.

I deal here with the problem of the origins of language. This question might seem to be a very poor place to look for clues to the mystery of the mind-brain relationship, fraught as it is with an unpreserved prehistory, centuries of speculations that have led nowhere, and many persistent confusions about the nature of language and linguistic knowledge. This problem ought to be the last place to turn for insights. These are important caveats to keep in mind, but there is one very good reason to think that this is an ideal place to look for clues to the mind-brain puzzle. Language is an evolutionary anomaly—and a big one at that. Such an elaborate one-of-a-kind cognitive-behavioral adaptation must have left a significant imprint on human brain design. Indeed, I argue that language is responsible for the major global changes in human brain structure (described here) that have evolved over the last 2 million years. With such a major restructuring of the brain producing a correlated major restructuring of function, we are offered one of nature's most blatant hints. We just need to ask the right questions first.

Without giving away the plot of my analysis, let me instead focus on one major conclusion and the implications that make it a relevant issue for readers of *Zygon*. I believe that human brains evolved a means to overcome the nearly insurmountable difficulties involved in learning to recognize symbolic reference; in other words, the ability to refer to things abstractly, by means of semantic as opposed to merely phenomenological associations. I argue that although there was no categorical biological change, no hopeful monster mutation or miraculous neural black box responsible for this shift in ability—only a quantitative rearrangement of parts—the functional *semiotic* change *is* categorical. It opens up a whole new universe of representational possibilities and creates a kind of virtual reference tool that recodes experiences and reorganizes mnemonic processes from the top down, so to speak. Whatever else we might want to claim about the nature of consciousness, what we mean by it has to do with the way the world is *represented* to subjective experience. All must agree, then, that a radical change in the mode of mental representation, from iconic and indexical to symbolic representation, must inevitably constitute a change in the mode of consciousness.

We are conscious of the possibility that there was a Big Bang that created the known universe, we are painfully conscious of our impending end of life, we are conscious of our confusion over the nature of the infinite, and most important, we are conscious of others who are conscious of us. A world of virtual reference is the world in which we discover other minds, because the ground of symbolic reference is fundamentally social and interindividual. We are able to take another's perspective in this virtual world and know something of the consequences

of our actions on others. We experience a sort of empathy available to no other creature, a virtual empathy, with representations of emotions that can be experienced overlaid with our own emotions simultaneously, and which can have substantial impact on them. This empathy is the basis for our most noble acts of self-sacrifice and caring, but it is also the basis for our most detestable and revulsive acts of terrifying and torturing. We intuitively do not hold our cats morally responsible for their playful torture of innocent birds and mice, nor even chimpanzees for dismembering captured monkeys alive as they devour them, but we dare not allow the perpetrators of unspeakable acts of genocide and mass murder to go unaccountable for their actions, even if they pulled no trigger. This representational ability, with all its powers, is what was in the proverbial apple that got us kicked out of the garden in the first place!

But we also take responsibility and authorship for our own actions with the aid of symbolization, because only by means of symbolic reference are we able to be at the same time above and within our own mental processes. We possess a form of agency, unavailable to other species, that is enabled by the representational distance that symbolization provides. The *I* that I identify with is often pitted against the *I* that emerges from my biology moment by moment, emerging from neither social nor biological causes, but from the self-organizing dynamic of an internal symbolic dialogue. This representational ability, this consciousness of a whole new world of abstractions, this intersubjectivity, this ability to run our lives based on counterfactuals, paradoxes, and outright fantasies, this sort of self, all came into the world in the last 2 million years as a result of subtle incremental changes in biology. As Einstein quipped, the most miraculous thing about the universe is its understandability. We may still be far from an understanding of the basis of consciousness, but I think we are close to understanding how and why our consciousness may differ from that of other species. And in many ways this may be the most important aspect of the mystery.

INTRODUCTION

Among the vast multitude of animal species, languagelike communication is the anomaly, not the rule. It's not just unusual or rare, it's essentially nonexistent except in one peculiar species: *Homo sapiens.* And I am not confining my definition of language to verbal communication or communication systems with exactly the grammatical structure that can be found in all human languages. I mean language in a generic sense: a mode of communication based upon symbolic reference and involving combinatorial rules that comprise a system for representing synthetic logical relationships among symbols. Under this definition, sign languages, mathematics, musical scores, and many rule-governed games

might qualify as languagelike, but not bird songs, vervet monkey alarm calls, honeybee dances, or humpback whale songs (some animal communicative behaviors often cited as languagelike), because these nonhuman activities lack both symbolic and combinatorial function, though they resemble language in certain superficial features. No more than a minute vocabulary of meaningful units and two or three combinatorial rules are necessary to fulfill these criteria. A childlike five-or-ten-word vocabulary and a grammar as simple as toddlers' two-word combinations would suffice. And yet even under these loosened criteria, no other species on earth has evolved any form of communication that even remotely qualifies. This is an important and little appreciated paradox, because it indicates that the complexity of language (e.g., the numbers of words and interdependent rules of grammar) is not the issue. Why are there no simple languages in the rest of the animal kingdom?

It is also not just a case of language not evolving in other species because of a lack of need. There is a fundamental difference in the potential for language. Thousands of years of living with domesticated animals and immersing them in human environments has not produced any well-documented cases of pets who understand what is said, except in a very superficial ("rote learning") sense. In addition, three decades of intensive efforts to teach language to apes (and more recently sea lions, dolphins, and a parrot) have shown that it takes almost heroic efforts and a careful choice of subjects and tasks to produce a modicum of symbolic understanding. Even these abilities are far more limited and ambiguous than first thought. At present, there is still considerable legitimate debate over the significance of the bits of languagelike behavior taught to nonhuman animals, and although I take a charitable view of many of these claims, and I do not doubt that given sufficient external support, a number of species might be able to develop some level of symbolic capacity, it is clear that spontaneous abilities to learn symbolic communication beyond rote-level associations (and thus not symbolic) are extremely limited outside our species.

Why did language evolve only once? Why is even a vastly simplified language so difficult for nonhuman species to acquire, whereas even an immensely complicated language appears easy for humans to acquire? I think we tend to gloss over the counterintuitive nature of these questions. Other species are capable of remarkably complex learning and cognitive analysis. Why can they not learn a very simplified language system? This apparent paradox strikes at the heart of what is generally the commonsense notion of the human-nonhuman difference: the assumption that it is the complexity of language that matters. This tacit belief is implicit in the two most common answers to these questions: (1) Humans have larger brains and are therefore smarter than other

species, and this makes them capable of using this much more complex form of communication. (2) Humans possess innate grammatical knowledge embodied in some unique human brain structure, and this makes learning the otherwise unlearnably complex rules of grammar and syntax largely unnecessary.

I have argued elsewhere that neither of these explanations offers anything more than a restatement of the question (Deacon 1992). On the one hand, to argue that language requires more intelligence merely restates the fact that other species are not capable of learning language because of some unspecified cognitive limitation. It ignores both the peculiarity of many features of language and the multiple dimensions of cognitive processes that might be differentially involved in language abilities as compared to other cognitive processes. On the other hand, to argue that language can be explained only by postulating some uniquely human brain structure with a set of rules capable of specifying any language merely passes the buck to some hypothetical black box wherein the answer to all questions about language structure and human language abilities can be found. But there is a more serious criticism of these answers.

The force of both arguments is undermined when we stop trying to explain complicated modern languages and instead ask why simple languages do not occur in other species. If only a dozen or more words and a couple of grammatical rules were involved, a vast learning ability would be unnecessary; and if the rules for the grammar were not so many and so intricately interdependent, they would no longer seem unlearnable, making an innate universal grammar irrelevant. Both arguments address a question that has little to do with language origins and the core human-nonhuman difference. Language processing must place some unusually intense demands on neural computations that are not well supported in nonhuman brains, otherwise there would be many other species with languages. However, which neural computations are these, if they do not involve complex grammar or vocabulary?

A Simple Working Hypothesis: Human Brains Are Adapted to Language

Language is not just some superficial part of human thinking. We are not just apes that have dabbled with some special communication trick. Language is totally integrated into every aspect of human mental functioning. We are linguistic savants, lightning calculators of semantic and syntactic arithmetic, and although people differ in linguistic abilities, it is a remarkable fact that only the most severely brain-damaged children fail to develop spontaneously some level of language competence. This

rare and anomalous cognitive ability is thus one of the most robust and irrepressible characteristics of our species. This is hardly the mark of an evolutionary afterthought, of a function that arose secondary to general intelligence or tool use. And it shows none of the awkwardness, inflexibility, inefficiency, stereotypicality, or mismatch with other social and cognitive functions that might be expected of an ability that arose by accident, without honing from natural selection. The most obvious interpretation of these facts is that the human brain evolved with respect to language, not independent of it, and did not develop language abilities as secondary spin-offs of some other adaptation.

I suggest that the anatomical changes that make language so nearly effortless for modern humans arose as adaptations to 2 million years of cognitive demands imposed by languagelike communication. This does not require that modern language per se predated the changes in brain structure that facilitate it, only that some languagelike system of communicating was present throughout the major period of hominid brain evolution; that the human brain and language coevolved. The brains of transitional australopithecines and early hominines would have been no better suited to meet language demands than are the best nonhuman brains today. But if forced, generation upon generation, to accomplish a nearly impossible cognitive task, natural selection would have inevitably played a role to ease the burden and decrease the probability of failing.

Early forms of languagelike communication would have recruited brain structures that evolved previously for other functions. Their overlap with these novel cognitive tasks would have been incidental, but all other brain structures would have been even less well suited. But ultimately those brain structures most impacted by these new computational burdens would be subject to the most intense effects of natural selection. This is an important hint. The effects of this adaptational process ought to have produced some of the most marked deviations of human brain structure from what is found in a typical primate brain. Reversing this logic: We would predict that those brain structures that are most deviant in human brains offer the best indices of the peculiarities of language-processing demands.

Ontogenetic Constraints on Human Brain Evolution

The most salient comparative feature of the human brain is its comparatively large size. The majority of claims about what is different about the human brain focus on this one trait. Although the change in size of the human brain must be a prominent component of any theory of human brain evolution, it is not necessarily the case that brain size, itself, is the trait that needs to be explained. Large human brain size almost certainly is not a simple trait with simple consequences (e.g., increased

intelligence), though many theories tacitly assume it is. Bigger brains are not just bigger, they are inevitably different.

Although it might seem a simple matter to assess the anatomical differences between human and nonhuman brains, it is far from a trivial project. The large size of the human brain makes comparisons with other species' brains problematic. Larger brains have different proportions among their parts than do smaller brains, so determining which of the myriad differences between brains are significant requires more than directly comparing structures, measurements, or lists of connections. The key to this problem is that the relative sizes of different brain structures are highly correlated. This pattern of correlated size changes is not surprising, given the systemic interconnectedness of different brain regions and the variety of ways that developing brains dynamically match cell populations and connection patterns in interdependent structures. Though comparative anatomists have labored for a century to produce data on surfaces, volumes, and neuronal population counts for the various brain structures in humans and other primates, only recently have we begun to understand the details of the embryonic mechanisms that determine these structural differences.

Remarkably, one can predict the size of most large forebrain structures in primate brains on the basis of brain size alone. But in the case of the human brain, this predictability breaks down in complex and interesting ways. Within each major "organ" of the brain, such as the cerebral cortex or thalamus, there are numerous structural and functional subdivisions. Many of these cortical areas, nuclei, and subdivisions diverge from primate predictions to varying degrees. This has given anatomists the impression that individual brain structures can grow and evolve in a piecemeal mosaic fashion. Many researchers have consequently theorized about independent adaptational functions for each of these apparent differences. But are these independent changes, or are many or all of these deviations superficial expressions of some more global underlying cause? We can begin to distinguish between these possibilities by analyzing the developmental mechanisms that are most likely to affect them.

Although quantitative data comparing the growth of embryonic structures in human and nonhuman primate brains are unavailable, one can gain a fair picture of the human deviations by comparing adult brain structures in groups that correspond to the major embryological growth fields. During development, the sizes of brain structures are determined hierarchically as the brain differentiates from major structural components into progressively smaller subdivisions. Extrapolating back from structural components that correspond to some of the earliest structural divisions to be formed, some large-scale patterns can be discerned. There appear to be two broad moieties of embryonic brain regions in humans

that are out of proportion with respect to each other. The cerebral and cerebellar cortices as well as the tectum apparently are larger than would be expected compared with most remaining brain structures, including the diencephalon, basal ganglia, brain stem, and spinal cord, along with many other structures. However, within either of these two groupings of structures, the components seem well proportioned with respect to each other (Deacon 1984; 1988). In general, the enlarged structures are all cortical-like surface structures located roughly on the anterior dorsal surface of the brain, and the comparatively smaller structures are all nuclear structures located ventrally and in the interior of the brain. The different structures that scale according to the same pattern are not associated by common connectivity or by common function, and represent all levels of the brain and nearly the entire range of sensorimotor modalities. They are, however, associated by position and by similarities of their laminated cell architectures.

What, then, are the causes of this unprecedented break in the typical primate growth allometry of these early-appearing brain regions? The mechanisms determining numbers of cell divisions, and thus target cell numbers, are as yet unknown. There are, however, some clues of a correlational nature. The comparatively enlarged and nonenlarged structures within the human brain divide roughly along suggestive embryonic lines. The embryonic neural tube is initially divided into dorsal and ventral halves by a tiny sulcus along each side of its interior wall, the sulcus limitans. This marks a developmental boundary that is respected by generative events all along the neuraxis. The three major divisions of the brain that are comparatively enlarged in humans are located on the dorsal anterior surface of the neural tube above this dorsal-ventral dividing line. The nonenlarged structures derive from the ventral half.

Recently, breakthroughs in developmental genetics have provided further clues to the significance of this pattern. Using in situ hybridization to discover when and where in the embryonic body selected genes are activated, it has become possible to map the sequence of genetic events that initially establish many of the major divisions of brain structures. Those that appear to play the crucial roles in initially partitioning the relatively undifferentiated neural tube into major brain regions produce proteins that bind to DNA and probably serve to regulate expression for suites of other genes. These "homeotic" genes, named for the whole-body segment modifications that often result from mutations affecting them, are highly conserved in all animals and are the initial determinants of cell lineage groups and cell fates. The expression domains of these homeotic genes appear to be essential for specifying the extent of progenitor cell regions that will become distinct brain structures. Although knowledge is still very incomplete concerning their functions in the

developing brain and the patterning of their expression, the regions of enlarged and unenlarged cell populations in human brains appear to respect some of these boundaries, suggesting that the proportional differences might be traceable to changes in the expression of certain homeotic genes. It is particularly relevant that the division between the dorsal enlarged and ventral unenlarged structures of the telencephalon follows the gene expression boundary respected by many genes. More work is needed with human and primate embryos in order to test this apparent gene/allometry correlation and to determine at what stage mitotic differences between these regions begin to be evident. Nevertheless, a shift in cell proliferation patterns at an early stage in neuraxis development could be sufficient to produce a subsequent systematic restructuring of circuitry in human brains as they mature.

In order to trace the consequences of such an early quantitative change in brain development, we must recognize that most of the information ultimately employed to build a functioning brain does not derive directly from genetic sources. It is rather the result of cell-cell interactions that incorporate spatial and, eventually, experiential information into the differentiation process (reviews in Purves and Lichtman 1985; Purves 1988; Deacon 1990b). This slight alteration in proportions of cells in human embryogenesis likely produces a cascade of other developmental consequences (fig. 1). These ensue because the patterns of axonal connections between structures are determined by competitive processes among developing axons. Because of this, we should expect that differences in numbers of projections from various areas competing for the same targets will be biased in favor of projections from larger structures.

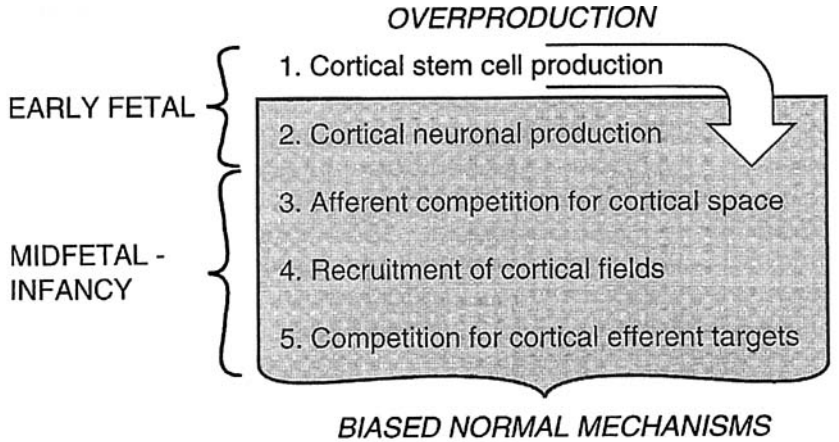

Fig. 1. List of the proposed sequence of embryological events that determine the unique proportions and connectional relationships of human brain structures.

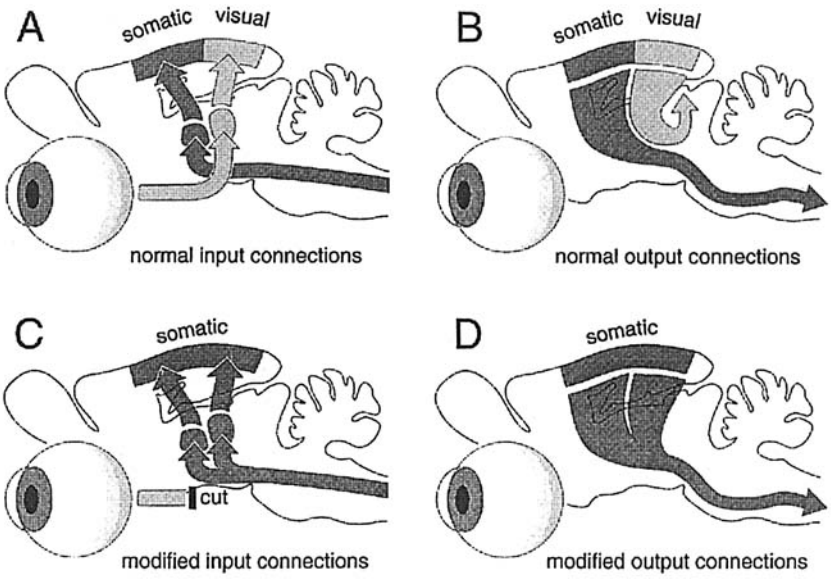

Fig. 2. Graphic depiction of an experiment described by D. O'Leary (1992) in which visual and somatic cortex and their projection systems are modified by prenatal elimination of visual inputs.

A: Normal visual and somatic sensory input pathways to the thalamus and from the thalamus to the cortex. The visual pathways are shown in lighter gray, and the somatic pathways are shown in darker gray. *B*: Normal cortical output pathways into the spinal cord (somatic) and tectum (visual). Although in early stages of development both visual and somatic cortical outputs project to both tectal and spinal targets in an undifferentiated pattern, during later development the tectal connections of somatic cortex and the spinal connections of visual cortex are competitively eliminated. *C*: Cutting the projections from the retina eliminates the visual inputs that would ordinarily have recruited space in the lateral geniculate nucleus of the thalamus and instead allows other afferent projections (e.g., ascending somatic projections) to recruit this abandoned target, thereby expanding the cortical representation of somatic responses. *D*: The alteration of thalamic inputs and cortical fields in *C* produces a different pattern of cortical outputs as well, due to the loss of fetal tectal connections and retention of spinal connections for both cortical regions. Thus, one sensory system replaces the other.

The effects of cell-proliferation allometry appear to be particularly important for the development of connections of the cerebral cortex. For example, projections from peripheral organs like the eye and tactile sensory system normally recruit target populations of neurons within the thalamus and cortex that are appropriate for the number of afferent

inputs they supply. This has been demonstrated by many experiments that manipulate numbers and patterns of peripheral inputs during early life. This sort of developmental displacement of some projections by others has been demonstrated by a number of experiments in which these relationships are directly manipulated; an example is shown in figure 2. Extrapolating this effect to the human brain/body relationship suggests an interesting possibility. Since the human body is only a fraction as large as would normally carry a brain the size of the human brain, recruitment of central neuronal populations should be significantly reduced with respect to the size of the brain. This is exemplified by the allometric relationships exhibited in the human visual pathways. Both the lateral geniculate nucleus, which receives primary inputs from the retina, and area 17 of the cerebral cortex, which receives lateral geniculate projections, are significantly smaller than would be predicted in a typical primate brain that reached human proportions. This follows inevitably from the fact that a brain of this size would actually be expected only in an ape of immense proportions, and this ape would have had much larger eyes and retinas than we have. But with input and output systems recruiting less synaptic "space" than expected, some other systems must stand to benefit in their recruitment. In contrast, those thalamic nuclei and cortical areas that receive predominantly centrally originating projections, especially from enlarged structures, are not similarly constrained and appear to inherit what cortical space is not taken up by the comparatively reduced sensory and motor maps. One region of the of the brain appears to have benefited most from this bias in favor of central versus peripheral projections: the prefrontal cortex. According to extrapolations derived from two different data sources (see Deacon 1984; 1988), prefrontal cortex is at least twice the size that would be predicted in an ape brain of this size (fig. 3).

The brain is not simply a collection of independent functioning anatomical modules, but a network. Many prior theories describing how brains evolved can be characterized as mosaic theories, suggesting that new brain structures were progressively added to old during evolution. But this view has become untenable in the face of recent embryological data that indicate that cortical areas do not develop from an intrinsic protomap but rather reflect a dynamic differentiation process, partly determined by geometric patterns of input and output connections and partly by competitive elimination of nonspecific connections. Cell production within the cerebral cortex precedes differentiation of its functional subdivisions, such as the prefrontal cortex. As a result, determination of which cells are destined to become prefrontal cells is a matter of dividing up a fixed total surface. Thus, the enlargement of the prefrontal cortex is an indirect conse-

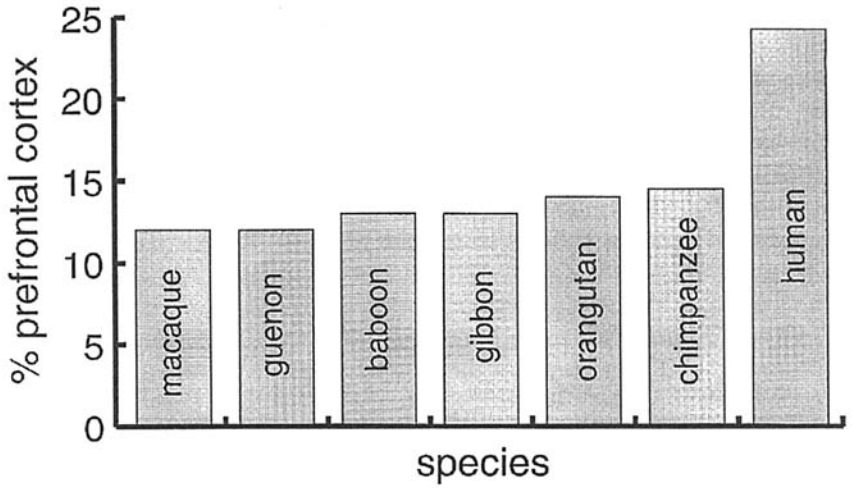

Fig. 3. One of many quantitative analyses (See Deacon 1984; 1988 for others) that shows prefrontal enlargement in humans with respect to monkeys and apes. Data are graphed in percentage of cortical surface area (as reported in Blinkov and Glezer 1968).

quence of the convergence of a number of systemic developmental processes, not the addition of extra neurons to this part of the brain (reviewed in Deacon 1990b).

Because of the magnitude of this difference, expansion of the prefrontal lobes during human evolution has been noticed by brain evolution researchers since the late nineteenth century (although there have been studies purporting to show that prefrontal lobes are not comparatively enlarged, each has suffered from either errors in correcting for allometric trends or problems of confounding different cortical regions in different species). The differences between previous views about prefrontal enlargement and the approach taken here derive from the two misunderstandings about brain evolution that have motivated this embryonic approach: the tendency to view it as a mosaic enlargement of an isolated "organ" of the brain and a predisposition to interpret it in terms of increased cognitive power of the prefrontal lobes. Not only must we view the enlargement in systemic terms, but we must understand the functional consequence in these terms as well. We must address these two formidable problems before we can hope to understand the significance of an enlarged prefrontal cortex. First, we need to consider the plausible mechanisms by which brain structure size differences could influence brain function. Second, we must sort through the considerable disagreement about the nature of prefrontal functions.

Irrespective of the developmental mechanisms that produced it, prefrontal enlargement and its correlated connectional effects clearly stand out as the most extensive differences distinguishing human brains from other primate brains. This major restructuring is our best clue concerning what is unique about human brains and their overall function, as well as the best source of information about the kind of cognitive demands that drove human brain evolution in the first place. We are, reluctantly, forced to try to make some sense of these two most enigmatic problems in one.

Overview of Prefrontal Anatomy and Connectivity

Just as prefrontal expansion can only be understood as a function of dynamic systemic interactions between many brain structures during development, its structural and functional consequences also require an understanding of systemic consequences of changes in size. Its comparative enlargement with respect to the majority of other brain structures in the cortex and elsewhere is a consequence of the developmental competitive advantage that its afferents have over other types of cortical projections, which include a very wide range of other cortical areas including all sensory and motor modalities and numerous subcortical regions. Of particular interest are its widespread cortical connections with nearly every modality of cortex. In an anatomical sense it stands between sensory-motor cortical areas and limbic cortex. Given its very large size compared to the sizes of its targets, it can be expected to occupy a far greater proportion of available synapses in these structures during development than do other structures that send competing afferents to these targets. Consequently, compared with more typical primate and mammal brains, prefrontal information processing will likely play a more dominating role in nearly every facet of sensory, motor, and arousal processes in humans. Irrespective of whether this structure has more "capacity" in some information-processing sense because of its size, it simply has more "votes" in whatever is going on in those regions of the brain to which it projects. In general terms, human information processing should be biased by an excessive reliance on and guidance by the kinds of manipulations that prefrontal circuits impose upon the information they process. We humans should therefore exhibit a cognitive style that sets us apart from other species, a pattern of organizing perceptions, actions, and learning that is peculiarly "front-heavy" so to speak. But how can this be described in neuropsychological terms?

Although during development the prefrontal region is probably carved out as a single projection field, in the mature brain the prefrontal cortex is not a single homogeneous structure with a single function. As a result there is a danger of overgeneralizing from studies based on one prefrontal

area to the whole prefrontal region. Different prefrontal regions receive diversely different cortical inputs and outputs that provide hints concerning their functional differences. Many regions receive inputs from specific sensory or motor modalities, and others receive converging inputs from more than one modality. No prefrontal area, however, receives direct input from primary sensory or motor cortices. One reason the prefrontal region remains to some extent mysterious is that its "map" structure is difficult to discern. Unlike the cortical topography of most sensory areas, positions within prefrontal regions do not seem to correlate with the peripheral topography of any sensory receptor surface. Nor is there a clear map of motor topography. One hint concerning the sort of "mapping" of functions within the prefrontal regions, however, comes from studies of visual attention and the subcortical structures that underlie it. Patricia Goldman-Rakic and her colleagues (1977) have demonstrated that one portion of the prefrontal cortex in monkeys (the dorsal lateral prefrontal region, or principalis region, named for its location surrounding the principal sulcus) is organized according to the direction of attention-driven eye movements with respect to the center of gaze. Damage to some sector of this region can selectively block the ability to learn to produce or inhibit directed eye movements in a particular direction or in response to cues in a particular direction. It is not surprising that this subregion of the prefrontal cortex is located adjacent to a region known as the frontal (motor) eye field, which directs eye movement. The eye-movement–attentional features of this region of the prefrontal cortex are consistent with its input-output association with the deep layers of the superior colliculus (a midbrain structure associated with visual orienting). This same sector of prefrontal cortex also shares extensive corticocortical connections with temporal and parietal visual areas (Barbas and Mesulam 1981). Dorsally and ventrally adjacent to this zone are regions that are reciprocally connected to auditory and multimodal auditory-somatic cortical areas of the temporal lobe (Barbas and Mesulam 1981; Pandya and Barnes 1987; Deacon 1992). These regions likely also send projections to other collicular regions where there is auditory representation (in the deep layers of the superior colliculus and the inferior colliculus). These areas may be expected to "map" auditory orienting processes in ways analogous to the way in which the principalis cortex "maps"' visual orienting. It is less easy to find map correlations for orbital and medial regions of the prefrontal cortex. These have predominantly limbic and adjacent prefrontal cortical connections and output pathways that include structures more associated with visceral and arousal functions than with sensorimotor functions.

The function of the lateral divisions of prefrontal cortex must partly be understood in terms of attentional mechanisms, with respect both to

collicular systems and to cortical systems to which they project outputs, whereas the function of orbital and medial divisions of prefrontal cortex must, in contrast, be more involved with arousal, visceral, and autonomic functions. These two systems are not only structurally interconnected but are probably functionally interdependent as well. Arousal, orienting, and attending are all part of the same process of shifting motivation to regulate adaptive responses to changing conditions. The lateral divisions may provide a substrate for intentionally overriding collicular orienting reflexes, using orienting information as cues for working memory about alternative stimuli or to select among many sensory configurations for further sensory analysis. The orbital and medial divisions may provide correlated shifts in arousal and autonomic readiness both to support shifts in attention and to inhibit the tendency for new stimuli to command attentional arousal.

Family Resemblances between Prefrontal Functions

So is there a common theme? Is there something that all these prefrontal cortical regions do similarly? This is no simple question. In fact, it remains one of the more debated questions in neuropsychology (for in-depth reviews see Fuster 1980; Perecman 1987; Stuss and Benson 1986). The reason it is difficult is that the explanation cannot be tied directly to any sensory or motor function. When prefrontal areas are damaged, there are no specific sensory or motor problems. Surgeons who performed prefrontal lobotomies earlier in this century used to point out that it didn't reduce their patients' IQs either. Consequences of prefrontal damage only show up in certain rather specific sorts of learning contexts. Nevertheless, these can be extensive and ultimately debilitating. And there is not just one type of prefrontal deficit but variants that correspond approximately to distinct prefrontal subdivisions. Because different prefrontal areas are connected to different cortical and subcortical structures, they produce slightly different types of impairments when they are damaged. Not only are there numerous competing theories attempting to explain individual types of prefrontal impairments, but there also is no account of the family resemblances that link the many different deficits associated with different prefrontal subareas.

I start by taking a global overview of prefrontal functions, not by treating the prefrontal cortex as homogeneous, but rather by searching among prefrontal areas, connections, and deficits for common themes and family resemblances. I am encouraged that there are some common threads because of the global similarities in connectional architecture that link these areas with the rest of the brain. I suspect that like the numerous subareas of the visual cortical system, the different prefrontal

regions share a common computational problem but have broken it up into dissociable subtasks in large brains, perhaps separated according to modality differences.

Let me begin by sampling a variety of interesting tasks affected by damage to the prefrontal cortex in monkeys. Figure 4 provides a schematic depiction of a number of learning tasks that have specific association with distinct prefrontal subdivisions. Beginning with the classical prefrontal task identified by Jacobsen (1936) many decades ago, figure 4*A* depicts the delayed response or delayed alternation task. In this task a food object is placed in one of two covered containers as the monkey watches. Then the experimenter distracts the monkey by pulling down a blind for a few seconds and then raising it to allow retrieval of the food by uncovering it. This is no problem, but on a succeeding trial, the hidden food object is placed in the alternative container, again in full view. Now, however, after the delay period, rather than looking in the new hiding place, the prefrontal-damaged monkey again tends to look in the place where he found food before, not where he saw it being hidden (this is similar to the hidden-object problems demonstrated in young children by Jean Piaget 1952). Some have explained this as a problem with short-term memory. The monkey might be unable to use information from a past trial to influence its choice in a future trial. A simple memory problem, however, would tend to produce random performance. In general, the animal's perseveration indicates that it does remember the previous successful trial, all too well, it would seem. Apparently it either can't inhibit the tendency to return to where he got rewarded the last time or can't subordinate this previously stored information to the new problem. Historically, interpreters of prefrontal deficits have split evenly over whether they interpret them in terms of memory or response inhibition. But before taking sides, let's consider a few additional examples.

Another more sophisticated version of the same task has been investigated by Richard Passingham (1985; see fig. 4*B*). His work offers some insight into how this task might have real-world adaptive consequences. As in the simple delayed-response experiment, food is placed in food wells while the monkey watches (although the observation is not a necessary factor), but unlike the simpler task, in this one food is hidden in all or many of a large number of wells. No delay is necessary. The monkey must simply sample through the wells to retrieve all the food objects. Monkeys with efficient sampling strategies tend not to sample the same wells twice. Once food has been located in one place and taken, there is no reason to go back and check it out. Prefrontal-damaged monkeys, however, fail to efficiently sample. They perseverate, by returning more often to previously sampled wells and failing to sample others. Again it is

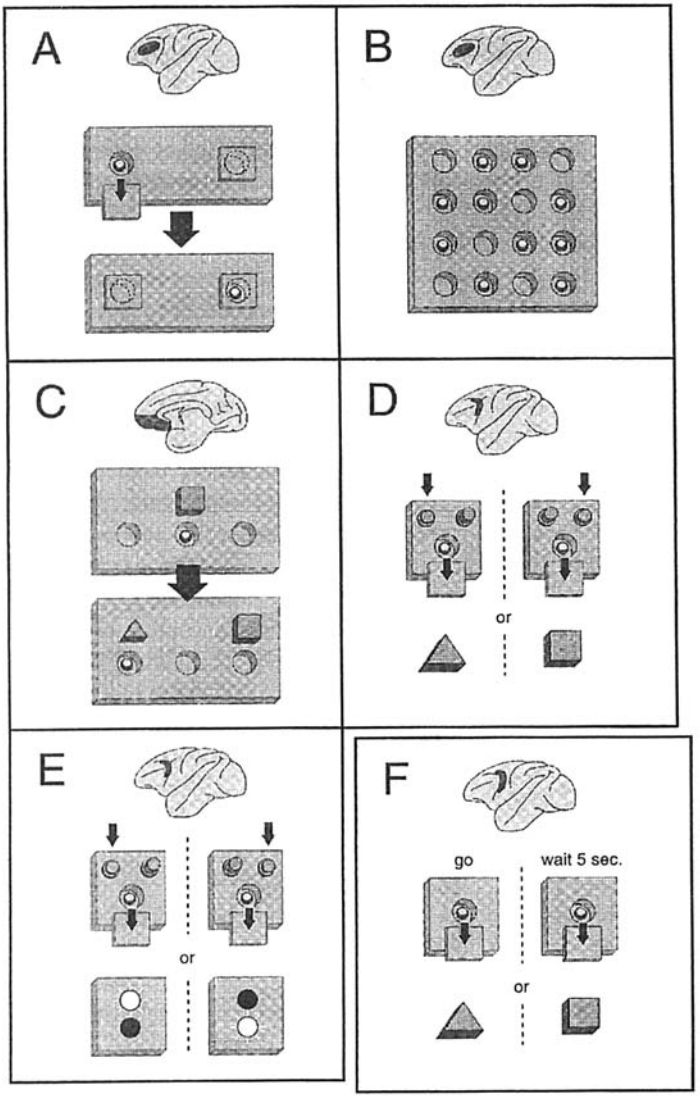

Fig. 4. Diagrammatic depiction of six different cognitive deficits shown in monkeys with frontal lobe damage to different subregions. *A*: Delayed response (delayed alternation) task associated with dorsal lateral prefrontal damage (Jacobson, 1936). *B*: Self-ordered sampling task associated with dorsal lateral prefrontal damage (simplified from Passingham, 1985). *C*: Delayed nonmatch to sample task associated with ventral medial prefrontal damage (Mishkin and Manning 1978). *D* and *E*: Conditional association tasks (spatial versus non-spatial cues, respectively) associated with periarcuate prefrontal damage (Petrides, 1982, 1985). *F*: Go/no-go task associated with periarcuate premotor damage (Petrides 1986).

not clear whether to consider this forgetfulness or failure to inhibit repetition of past responses. The practical significance of such an ability is clear, however. This is precisely the sort of problem that might be faced by an animal foraging in many different places. Once all the food has been eaten in one place, it makes no sense to go back looking for more later, even if the food tasted particularly good there.

Turning now to medial frontal damage, we find a slightly different kind of deficit. Monkeys with medial prefrontal damage succeed at the hidden object type tasks, but they fail at tasks where the shift in food location is cued by a shift in stimulus as well. Figure 4*C* depicts this sort of task. The food object is hidden and unlike the hidden object task, the location is cued by some stimulus. After the monkey succeeds at this task, the food object is rehidden and the hiding place is marked by a new stimulus, with the previous stimulus now marking no food. Thus, the monkey must learn that the food will always be hidden where the newer stimulus of the two is placed. Medial prefrontal damage appears to preferentially affect this kind of learning but not delayed response, delayed alternation, or sampling tasks that are sensitive to dorsal lateral prefrontal damage.

Compare this to tasks sensitive to lesions of posterior lateral prefrontal regions. These tasks have a multipart form. Figures 4*D* and 4*E* provide examples of these tasks. Common to these tasks is a dependency between two classes of cues or between alternative cues and behavior options. In 4*D* one cue stimulus indicates that the food is hidden in the well with the lighted light, whereas the alternative cue indicates that the food is hidden in the well with the unlighted light. The pattern is "if *X* then *Z*, if *Y* then not *Z*." It is a conditional relationship in which one stimulus indicates the relationship between another stimulus and the position of the food. In 4*E* there is a similar dependency relationship, but this time between a stimulus pattern and a choice of buttons that open the food well. Unlike the previous task, there is no spatial difference in food position, but like the last task, there is a conditional relationship in which the stimulus indicates which of two alternatives is associated with the food. Depending on what stimulus is presented, the monkey has to reverse his expectation about the association between the reward and some behavioral option.

These tasks are all different and yet they share a number of common features. Although all involve some apparent inability to inhibit responding, this may be secondary. Prefrontal-damaged animals do not show a problem with simple go/no-go tasks that require withholding a response, as in 4*F*. In this sort of task the presentation of one of two different stimuli indicates whether an immediate response or the same response just delayed a few seconds will produce a reward. Animals that

have difficulty suppressing a response will be unable to learn such tasks because they will fail at the no-go task. Prefrontal damage generally does not produce significant impairment on simple go/no-go tasks. This difficulty is demonstrated by premotor lesions, which additionally produce problems with motor sequences and skilled movements.

One might also argue that all of these tasks involve holding information in mind while not acting on it, a function some have called working memory. The deficit pattern is not simply a failure of short-term memory, however, because the perseverative failures are themselves the best indications of acting on the basis of prior information. In fact, information in short-term memory seems to inappropriately dominate the tendency to respond. What must be held in mind in these tasks is not just prior information, but information about the applicability of that information in a different context. One of the most salient common features of these tasks sensitive to prefrontal lesions is that they all, in some way or other, involve shifting between alternatives or opposites, alternating place from trial to trial, shifting from one stimulus to a new one, or from one pair-wise association to another, depending on the presence of different cues. Tasks sensitive to prefrontal damage thus all involve short-term memory, attention, suppression of responses, and context sensitivity, but they all have one other important feature in common. Each involves a kind of negation relationship between stimuli or stimulus-behavior relationships. They all have to do with using information about something one has just done or seen against itself, so to speak, to inhibit the tendency to follow up that correlation and, instead, shift attention and direct action to alternative associations. Precisely because one association works in one context or trial it is specifically excluded in the next trial or under different stimulus conditions. An implicit "not" must be generated to learn these tasks, not just an inhibition.

Similar deficits are well known in human patients (Kolb and Whishaw 1990; Stuss and Benson 1986), even though associations between specific tasks and different prefrontal subareas are not worked out as well in humans. For example, human prefrontal patients often fail at card-sorting tasks that require them to change sorting criteria. They also tend to have trouble generating lists of words. In trying to generate word lists according to some criterion or instruction, they hardly get past the first few names of things before getting stuck or repeating items already named. These two skills are formally similar to conditional association and sampling tasks, respectively. In addition, prefrontal patients also often have difficulty learning mazes based on success-failure feedback, making plans, and spontaneously organizing behavior sequences, and with tasks that require taking another perspective (allocentric vs. egocentric). Analogous to using a mirror, thinking in allocentric terms requires

a systematic reversal of response tendencies. In general, tasks that require convergence on a single solution are not disturbed by prefrontal damage, whereas those that require generating or sampling a variety of alternatives are. This capacity has been called "divergent thinking" by Guilford (1967), and may explain why prefrontal damage does not appear to have a major effect on paper-and-pencil IQ tests. Like the logic shared by tasks sensitive to different frontal lobe defects in monkeys, the many human frontal lobe signs also crucially involve difficulties in using information negatively. Prefrontal patients show a generalized tendency to be controlled by immediate and simple correlative relationships between stimuli and rewards, which essentially blocks the ability to entertain higher-order associative relationships, because of the inability to subordinate one set of associations to another.

These insights about prefrontal functions, although far from solving the riddle of the prefrontal lobes, may offer sufficient information for understanding the significance of the remarkable expansion of this structure in human evolution. They do, however, beg this question: What crucial adaptation demanded such a premium on the ability to learn complex conditional and negational relationships?

The Symbol Acquisition Problem

If prefrontal expansion—and by implication the increasing influence of the functions of the prefrontal cortex over other cognitive and sensorimotor processes—are both consequence and cause of human cognitive evolution, then it is reasonable to suspect that the functions of the prefrontal cortex ought to provide insight into our most divergent cognitive ability—language. It does not follow that prefrontal cortex is the locus of language functions, the repository for grammatical knowledge, or the basis for increased intelligence. I think it is none of these. Rather, I believe it addresses a learning problem that lies at the heart of language: the problem of the missing simple languages. I suggest that we have been looking at the wrong level of the phenomenon for answers. Other animals' brains are not just abysmal at performing the computations required for analyzing the grammatical relationships between symbols; they cannot even be tested adequately, because they are unable to perform the necessary computations for learning even a simple symbolic reference system. In other words, I think that it is not grammar that is holding other species back. It is something much more basic and more subtle: symbolic reference.

What is so hard about learning symbolic reference? Why should symbolic associations be different from other associations? One possibility is that they might involve more complicated stimuli. For example, there might be more details to remember or fewer clues to help one learn the

associations. Another possibility is that one may need to learn many more associations for any of them to be useful. For example, in language it seems necessary to combine words into sentences in order for them to serve any purpose. Isolated words have meaning only in special kinds of utterances, such as giving commands, naming objects, or identifying people. However, there is a third possibility that I wish to explore: the possibility that symbolic associations are different in more fundamental ways from other kinds of associations, different in ways that nonhuman brains are poorly equipped to handle and that human brains have become specialized to overcome.

This difference is the basis for an old and stubbornly resistant philosophical question: What is special about the way we represent and understand meanings in language? This question addresses the difference we refer to by distinguishing understanding from mere rote learning. There is a crucial difference that distinguishes symbolic associations from other forms of learned associations, but we tend to ignore this difference because we usually find the transition between nonsymbolic associations and symbol learning so effortless. The tacit assumption is that word reference is learned in essentially the same way as are other associations. The commonsense idea is that a symbolic association is formed when we learn to pair a sound or inscription with something else in the world. The idea or concept of the thing associated with the sign constitutes the symbolic link. According to this view, the association between a word and what it represents is not essentially distinguished from the kind of association made by an animal in a Skinner box when it learns that there is a correlation between a red light and the availability of food. The conditioned stimulus takes on referential power in this process: It represents something about the state of the apparatus for the animal. It is, technically, an *index* of a change in state of the Skinner box. When the light is off, there is no food available, but when it is illuminated, food is available. When the light is off, no action the rat can perform will induce the apparatus to deliver the food, but when it is illuminated, the rat can perform a particular associated behavior (e.g., pressing a bar), and food will be delivered. Although common sense suggests that word meaning is more complicated than this and that conditioned association is somehow more mechanical and nonsemantic, it has been curiously difficult to find a clear exposition of the difference between them in either the psychological or the philosophical literature.

The development of stimulus generalization or learning sets has also been compared to symbol learning, but this is also not sufficient to explain the difference between symbolic and nonsymbolic associations. A similarity is often suggested because terms for things usually name classes of things rather than individual things. Transference of learning

from stimulus to stimulus or from context to context occurs as a natural incidental consequence of learning. This is the case because there is always some ambiguity as to the essential parameters of events that are antecedent to the conditions the subject is seeking to reproduce, and because the learning process is essentially a statistical estimate of the sufficient stimulus. Thus, to the extent that other stimuli or stimulus contexts are physically similar in some respect to the implicit subsample used for training, they are also incidentally learned. Although this may be formally represented in psychological models as though the subject has learned rules for identifying associative relationships, the generalization effect is not so much the result of listed criteria as it is a failure to distinguish, a tendency to gloss over differences. Transference of learning can be broadened by training that purposely varies stimulus and response conditions along certain parameters. This kind of generalization is still essentially based on one-to-one pairings of stimuli, but what constitutes a stimulus is ambiguous in certain dimensions.

Simple conditioned stimuli are ultimately symptomatic or indexical of the stimuli with which they have become associated. Stimuli linked by learned association are acquired because of their contiguity or coincidence in space or time, by the fact that they habitually "go together" in some way or other. In the same way that the presence of a fever indicates disease, or that smoke indicates combustion, the conditioned stimulus is a signal to the subject that in the present state the device or the experimenter will likely provide the associated reward. The stimulus is evidence of this change in state. To distinguish this sort of associative relationship from others, particularly symbolic associations, I identify it as indexical association.

Understanding the difference between this sort of learned association and symbolic association is fundamental to my argument, so I digress slightly from the brain-language problem to deal with this most ensnaring philosophical problem. Given the history of failures to solve the conundrums it poses, I hope the reader will be charitable if in this short exposition I do not fully plumb the depths of the problem. I do hope that at least the skeleton of the approach will become clear. The answer I propose can be paraphrased by saying that symbols are essentially about indexical associations, not about objects directly.

Take, as a starting point, words and objects. The source of the problem in understanding the difference between the symbolic and nonsymbolic relationships involved is that terms for objects can be paired with things in a way that superficially resembles conditioned association. However, by virtue of the fact that words also represent relationships to other words (think of the way a dictionary works), this pairing is far from the whole story. In fact, it is by virtue of this sort of dual reference,

to objects and to other words (or at least to other semantic alternatives), that a word conveys the information necessary to pick out objects of reference. Unlike a colored light in a Skinner box, a word doesn't refer to some thing or condition by virtue of habitually being associated with it—in fact, the physical association between a word and an appropriate object of reference can be quite rare or even an impossibility—but rather by virtue of carving out a kind of logical space. Words superimpose pragmatic logical boundaries on the physical continuities and discontinuities found among real stimuli and events. This is what provides the power of symbolic relationships, because by virtue of the possible combinatorial interrelationships between symbols, there can be an exponential growth of reference with each new added element.

Even without struggling with the philosophical subtleties of this relationship, we can immediately see the significance for learning. The learning problem associated with symbolic reference is a consequence of the fact that what determines the pairing between a symbol (such as a word) and some object or event is not their probability of co-occurrence, but rather some complex function of the relationship that the symbol has to other symbols. Learning is, at its base, a function of the probability of correlations between things, from the synaptic level to the behavioral level. Past correlations tend to be predictive of future correlations, and so it is a powerful if simple recipe for adaptation. In order to comprehend a symbolic reference, however, you have to selectively ignore certain habitual associations and correlations between symbols as stimuli and their objects of reference as stimuli and instead focus on the relationships between different symbols and how these modify the probabilities of symbol-object co-occurrence. This is a troublesome shift of focus. The correlations between symbols and objects are merely the clues for determining the more crucial relationships between the symbols themselves. And these clues are not highly correlated. Let me offer an extended example to help demonstrate this problem.

One of the most insightful examples of the difference between conditioned associations and symbolic associations is offered by a set of experiments that attempted to test symbolic abilities in chimpanzees. This study was directed by Sue Savage-Rumbaugh and Duane Rumbaugh, then at the Yerkes Primate Center (Savage-Rumbaugh and Rumbaugh 1978; Savage-Rumbaugh et al. 1980). The chimps in this study were taught to use a special computer keyboard made up of lexigrams—geometric drawings on large keys. Though previous experiments had shown that chimps have the ability to learn a large number of paired associations between lexigrams (and in fact other kinds of symbol-tokens) and objects or activities, some problems arose when they were required to use these in simple combinatorial relationships. In order to test the chimps'

symbolic understanding of the lexigrams, they were trained to chain lexigram pairs in a simple verb-noun relationship (e.g., a sequence glossed as meaning "give," which caused a dispenser to deliver a solid food, and "banana" to get a banana). There were initially only two "verb" lexigrams and four food-or-drink lexigrams to choose from, and each pair had to be separately taught. But after successful training of each pairing the chimps were presented with all the options they had learned independently and were required to choose which combination was most appropriate on the basis of food availability or preference. Curiously, this task was not implicit from their previous training. This was evidenced by the fact that some chimps tended to stereotypically repeat only the most recent single learned combination, whereas others chained together all options, irrespective of the intended meanings and what they knew about the situation. Thus, they had learned the individual associations but failed to learn the system of relationships of which these correlations were a part. Although the logic of the combinatorial relationships between lexigrams was implicit in the particular combinations that the chimps learned, the converse exclusive relationships had not been learned. Though implicit for those of us who treat them symbolically from the start, the combinatorial rules of combination and exclusion that underlie the symbolic use of these lexigrams was vastly underdetermined by the training experience.

It is not immediately obvious exactly how much exclusionary information is implicit, but it turns out to be quite a lot. Think about it from the naive chimpanzee perspective for a moment. Even with this ultrasimple symbol system with six lexigrams and a two-lexigram combinatorial grammar, the chimpanzee is faced with the possibility of sorting among 720 possible ordered sequences (6*5*4*3*2*1) or 64 possible ordered pairs. The training has offered only four prototype examples, in isolation. Though each chimp may begin with many null hypotheses about what works, these are unlikely to be in the form of rules about allowed and disallowed combinations, but rather about possible numbers of lexigrams that must be pressed, their positions on the board, their colors, or shape cues that might be associated with a reward object.

Recognizing this limitation, the experimenters embarked on a rather interesting course of training. They set out to explicitly train the chimps which cues were irrelevant and which combinations were not meaningful. This poses an interesting problem that every pet trainer faces. You can't train what not to do unless the animal first produces the disallowed behavior and it can be immediately punished or at least explicitly not rewarded. So the chimps had to first be trained to produce the incorrect association (e.g., mistaking keyboard position as the relevant variable) and then subsequently have only that aspect explicitly not rewarded. By

a complex hierarchic training design involving thousands of trials, it was possible to systematically exclude all inappropriate associative and combinatorial possibilities, leaving the animals able to produce the correct lexigram strings essentially every time. Remarkably, after this training regimen, when a new food item and new lexigram was introduced some of the chimps were able to respond correctly the first time or with only a few errors, instead of hundreds as before. What had happened to produce this difference? How had they graduated from what we would recognize as rote learning to what we would call an understanding of the meaning of the lexigrams by this process?

What the animals had learned was not only an association between lexigrams and objects or events. They had learned a set of logical relationships between the lexigrams, defined by exclusion and inclusion. More important, these lexigram-lexigram relationships formed a complete system in which each possible relationship of adjacency, substitutability, or opposition was defined. In fact, they had to learn that the relationship that a lexigram has to an object is a function of the relationship it has to other lexigrams, not a simple function of the correlated appearance of both lexigram and object. Reference is determined indirectly. This subordination of associative relationships to combinatorial relationships between symbols is schematically depicted in figure 5. Indexical associations are one-to-one, and the indexical reference is achieved as a function of the correlations between some token (i.e., the sign stimulus) and some object (shown as a solid arrow). In contrast, the system of token-token interrelationships, such as those between lexigrams or words (shown as solid arrows interconnecting symbols), is essentially independent of their indexical functions. Tokens indicate one another in the sense that their presence or position in a communicative activity influences the admissibility or nonadmissibility of others. This, however, is a purely conventional token-token indexicality, because it constitutes a closed group of "pointing" relationships (i.e., determines reference to objects collectively as a function of relative position within this token-token reference system). Symbolic reference emerges from the hierarchic relationship between these levels of indexicality. Although the indexical reference of symbol-tokens to objects is maintained, it is no longer determined by a simple correlational relationship between sign and object. The subordination of indexical reference to the lexigram-lexigram relationship, however, makes a new kind of generalization possible: logical generalization, as opposed to stimulus generalization. This is what made the no-trial learning of new lexigram-object relationships possible for the chimps Sherman and Austin.

The system of lexigram-lexigram interrelationships is a source of implicit knowledge about how novel lexigrams must be incorporated. Add-

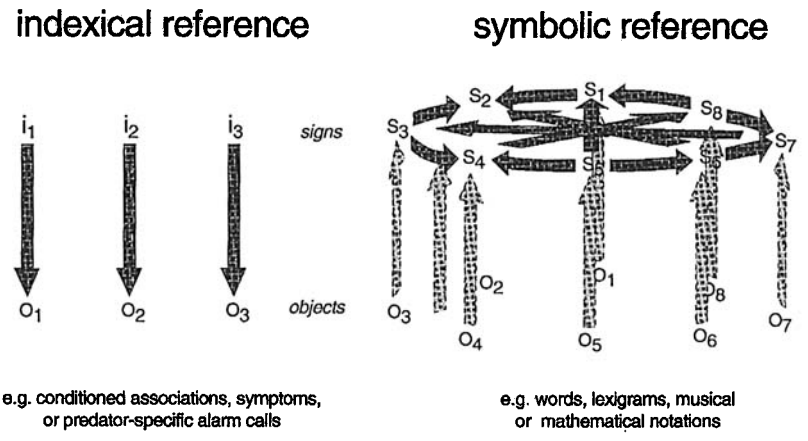

Fig. 5. Schematic depiction of the difference between indexical reference, as might be created by correlative associative learning, and symbolic reference. Simple associative links between sign stimuli and objects are one-to-one and essentially independent of one another (indicated by dark arrows in the left figure), except insofar as the stimulus parameters might overlap. Correlation between sign stimulus and object is the basis for indexical reference relationships. Symbolic reference is based primarily on the system of combinatorial inclusion/exclusion relationships between stimulus signs (symbols) and the way these pick out categorical boundaries of classes of indexical associations (indicated by dark solid arrows connecting symbols). Although indexical links still exist between objects and symbols that represent them, these are entirely secondary to symbol : symbol relationships and no longer are a function of correlation and co-occurrence (indicated by light gray dashed arrows).

ing a new food lexigram, then, does not require the chimp to learn the correlative association from scratch each time. The referential relationship is no longer a function of lexigram-food co-occurrence, but rather a function of the relationship this new lexigram has with the system of other lexigrams. There is a shift in analysis from relationships among stimuli to relationships among lexigrams as logical operators. A new food or drink lexigram must fit into a predetermined slot in this system of relationships. There are only a few possible alternatives to sample, and none requires the chimps to assess the probability of paired lexigram-food occurrence, because lexigrams need no longer be treated as indices of food availability. As with words, the probability of co-occurrences may be quite low. In a real sense, the food and drink lexigrams are nouns and are defined by their potential combinatorial roles. Testing the chimps' ability to extrapolate to new lexigram-food relationships is a way

of demonstrating whether or not they have learned this logical-categorical generalization. It is a crucial defining feature of symbolic reference.

The experimenters provided a further, and in some ways more dramatic, demonstration of the difference between the rote learning of lexigram-object correlations and symbolic learning by comparing the performance of the two symboling apes, Sherman and Austin, to a previous subject, Lana, who had been trained with the same lexigram system but not in the same systematic fashion. Lana had learned a much larger corpus of lexigram-object associations, though by simple paired associations. All three chimps were able to learn a task that required sorting food items together in one pan and tool items together in another. In fact, Lana learned in far fewer trials than did either Sherman or Austin. After this, when presented with new foods or tools, the chimps were able to generalize from their prior behavior to sort these new items appropriately as well. This is essentially a test of stimulus generalization, although it is based on some rather abstract qualities of the test items (e.g., edibility). A second task, though, in which the chimps had to associate each of the same food items with a single lexigram for *food* and the same tool items with a single lexigram for *tool* provided different results. This task clearly distinguished the symboling chimps from Lana. Even though all three chimps were able to learn the new associations (this took as many trials to learn as did the original training), when asked to generalize the referential scope of these lexigrams to two new foods and two new tools, only Sherman and Austin were able to do so essentially without errors. Lana not only failed to add the new items as referents of the lexigrams, but also became unsettled by the errors, and in successive testing ignored what she had previously learned. She treated these new trials as independent from the preceding trials, and essentially assumed that she needed to learn the new associations from scratch.

Sherman and Austin, as a result of their experience with a previous symbol system, had recruited the individual lexigram-object associations they had acquired by rote and used them to create two new symbolic categories that superseded the individual associations. It took hundreds or thousands of trials to learn the first simple one-to-many associations, because there was no systemic relationship in the chimps' small existing lexigram set into which a general reference for *food* and *tool* would fit. Once the chimps had established the new symbolic relationship, though, it was easily expandable. Because of this, they generalized to new associations, not by stimulus features, but by what amounted to semantic features. They were eventually able to fully integrate these categories with their existing system by learning associations between these two lexigrams and other lexigrams, eventually associating lexigrams of individual food items directly with the lexigram for food. Although it typi-

cally took hundreds, even thousands, of trials for the chimps to acquire a new rote association, once a systemic relationship was established, new items could be added essentially without any trial and error. This difference translated into more than a hundredfold increase in learning efficiency. This difference is the key to understanding the apparent leap in human intelligence as compared to other species. Increased intelligence didn't produce symbols; symbols increase effective intelligence. The power of the symbolic step in this learning process derived from the fact that the chimps essentially knew something that they had never explicitly learned. Implicit knowledge is an inevitable spontaneous product of symbolic representation. This fact plays a critical role in many facets of language acquisition that are often attributed to innate foreknowledge.

I have chosen to recount this "ape-language" study not because it portrays any particularly advanced abilities in chimpanzees, nor because I think it is somehow representative. I have focused on it because of the clarity with which it portrays the special nature of symbol learning, and because it provides an example of the hierarchic relationship between symbolic and indexical (simple correlative) referential relationships. Symbolic referential relationships are constituted by relationships among indexical referential relationships. Indexical associations are necessary stepping-stones to symbolic reference but must ultimately be overcome and ignored for symbolic reference to work.

The temporal-spatial correlations between the sign stimulus and object do not mean what they predict, i.e., that one is causally related to the other. In learning symbolic associations, the apparent causal implications of correlative associations must be ignored and associated causal expectations must be suppressed in service of the search for a higher-order relationship between the sign stimuli irrespective of their causal correlations with other objects. This higher-order relationship is not determined by any physical properties of the sign stimuli. It is a logical relationship defined by allowed and disallowed combinations. Before symbolic reference is possible, one must first learn many nonsymbolic associative relationships that are, in effect, only symptoms of a higher-order symbol system. The association between a sign stimulus and an object must be understood not as pointing to that object but as pointing to the place that this associative relationship occupies in a system of other associative relationships and by virtue of which it is identified.

The problem with symbol systems, then, is that there is a lot of both learning and unlearning that must take place before even a single symbolic relationship is available. Symbols cannot be acquired one at a time the way other learned associations are, except after a reference symbol system is established. A logically complete system of relationships among the symbols in the set must be learned before the symbolic association

between any one symbol and an object can be thereby determined. The learning step occurs prior to recognizing the symbolic function, and this function emerges only from a system; it is not vested in any individual sign-object pairing. For this reason, it's hard to get started. To learn the first symbolic relationship requires holding a lot of potential combinations in mind at once in order to discover how any one fits in with the others. Even with a very small set of symbols, the number of possible combinations is vast, and so sorting out which combinations work and which don't requires sampling and remembering a large number of possibilities. Moreover, remembering by rote which combinations worked in which situations may work against the need to decompose the combinatorial relationships in order to discover the underlying rules of logical exclusion and inclusion that they encode. The problem with learning to reproduce symbolic material by rote, i.e., as indexical associations (like learning to reproduce a mathematical calculation by memory), is that the information is not generalizable except with respect to perceptual parameters of the stimuli. It is the essence of symbolic associations that their reference is determined by general rules—logical relationships that have application across all possible combinations in the system.

The Contribution of the Prefrontal Cortex to Symbol Learning

This difference between associative learning and symbol learning has some interesting consequences so far as the evolution of intelligence and language is concerned. The ability to acquire learned associations between stimuli enables an animal to adapt more efficiently and flexibly to the cause-effect contingencies of a complex changing environment. The ability to learn quickly to discern and predict the most highly correlated spatial-temporal relationships among events is a powerful strategy for internalizing the structure of the world around us. It is, however, a poor strategy for learning symbolic relationships.

In fact, it is probably the case that an ability to rapidly discover and memorize the simple correlative relationships among stimuli would interfere with discovery of abstract rules of logical combination among these same stimuli that could be the basis for symbolization. Since the probability of correlating a symbol with a given object depends entirely on which other symbols it precedes, follows, or co-occurs with, the statistics of correlation provide a poor predictor of the relationship across all possible occurrences. Not only that, the smarter the learning device, brain or otherwise, the quicker it will tend to jump to such conclusions, because it will discover the subtle differences in the weightings of statistical associations more quickly. The faster the statistical weighting of cor-

relations is discovered, the faster the learning process will proceed. But such a statistical best guess must ignore any large-scale distributed logic that might be exhibited combinatorially, and this is precisely what a symbolic system must be built from. In short, increased intelligence defined in this way would likely be counterproductive to symbolic learning.

This is why the evolution of the human brain is explicitly not well described as the evolution of increased intelligence. Building a smarter brain is not sufficient to get over the threshold separating simple associative learning abilities and symbolic learning abilities; in fact, it makes the threshold higher. It may also partially explain why children, with their somewhat limited learning rates and memory spans, may be so much better than adults at learning symbolic systems from scratch. Their intelligence does not get in their way, so to speak.

So what took human brains over this hump? I suggest that this is the significance of the enlargement of the prefrontal cortex and expansion of its projection systems. Not because we have a smarter prefrontal cortex, but rather because the prefrontal system has become much more involved in the activities of all other brain systems. Abraham Maslow once quipped that if the only tool you have is a hammer, you will tend to treat everything like a nail. This offers an analogy I would apply to this prefrontal change. The prefrontal propensity to inhibit the tendency to act on simple correlative stimulus relationships and instead sample possible higher-order sequential or hierarchic associations has come to dominate the human learning process more than in any other species. In simple terms, much more control of the brain is vested in the prefrontal cortex in human brains. The way the parietal cortex handles tactile and movement information, and the way the auditory cortex handles sound information, the way the visual cortex handles visual information are now much more constrained by prefrontal activity than in other species.

The contributions of prefrontal areas to learning all involve, in one way or another, the analysis of higher-order associative relationships. More specifically, judging from the effects of damage to prefrontal regions, prefrontal regions are necessary for learning associative relationships in which one association is in some way subordinated to another. These mental computations address the most critical learning problem faced during symbol acquisition. The more complicated the combinatorial relationships or the more easily confused the correlated relationships, the more that prefrontal systems are taxed. This is clearly demonstrated by cerebral blood flow and PET imaging studies of the metabolic correlates of different cognitive tasks in human subjects. Complex sorting problems and difficult word-association tasks have been shown to particularly activate prefrontal metabolism (see also Deacon 1989 on

language tasks). There is also indirect evidence that task difficulty determines how much prefrontal cortex gets recruited to the task. Electrical stimulation studies of awake neurosurgery patients have shown that patients with lower verbal IQs tend to have larger regions of prefrontal cortex susceptible to disruption of language tasks (Ojemann 1979).

BRAIN-LANGUAGE COEVOLUTION

Expansion of the prefrontal cortex and its projection fields in human evolution can be interpreted in the context of these special learning problems. It is not clear whether we should interpret prefrontal expansion as an enhancement of these classes of mental computations or merely a predisposition to treat most learning contexts as involving combinatorial and conditional relationships. Either would contribute significantly to symbol learning, possibly at some cost in terms of simple associative learning. However we interpret this difference, it cannot be doubted that such a major change in brain structure reflects some rather special learning demands faced by our ancestors, but no other species. It can hardly be a coincidence that the most salient differences of human brain structure and human cognitive abilities from those of other animals converge on the same learning problem. The magnitude of prefrontal enlargement and the nearly 2-million-year time-course of this evolutionary change suggests that these capabilities were under powerful selection for a considerable period during hominid evolution.

This may also provide an explanation for the failure of languagelike (symbolic) communication to evolve in all but one species. A simple improvement of learning rates or memory capacity, etc., cannot account for the transition to symbolic communication that took place in hominid evolution. One cannot extrapolate some general tendency toward more complex communication or higher intelligence and arrive at language evolution. This is because the cognitive requirements for efficient associative learning are in many ways in conflict with those that would enhance symbol learning. Selection for the one would tend to be countered by selection for the other. The transition from associative forms of learning and communication to symbolic forms requires the crossing of a high threshold in terms of learning costs. The organism must invest immense learning effort in acquiring associative relationships that make no sense until the whole system of interdependent associations is sorted out. In other words, for a long time in this process, nothing useful can come of it. Only after a complete group (in the mathematical sense) of interdefined symbols is assembled can any one of them be used symbolically. Until then, their indexical associations will be useful in only a limited set of stereotypic contexts. To approach most learning problems

with the expectations and biases that would aid symbol learning would be very inefficient for most species.

The time course of brain-language coevolution can be estimated unambiguously because of the way this difference in brain structure correlates with features that can be discerned from fossils. Prefrontal enlargement and the enlargement of the projection fields of the prefrontal cortex are determined systemically by competitive processes during development that are reflected in global brain parameters, specifically the relationship between brain and body size. The size of the cortical region recruited by the medial dorsal and anterior thalamic nuclei (which project to prefrontal cortex in adult brains) is a function of competition with thalamic projections associated with peripheral sensory and motor systems. Consequently, the relative size of the brain with respect to the body (which correlates with the size of peripheral organs and their projections) should be an accurate index of the proportions of cortical areas, including the size of the prefrontal cortex. As hominid brains first began to enlarge significantly with respect to body size approximately 2 million years ago, they were not merely increasing in brain size, but in the proportion of prefrontal cortex and the proportion of prefrontal projections into target fields that in other brains would be occupied with other sensory, motor, or limbic projections.

Hominid brain expansion can therefore be used as an index of the change in its internal structure and for the degree of functional change associated with incremental prefrontal expansion. The increase in brain size traced from *Homo habilis* to *Homo sapiens* therefore is a symptom of prolonged selection favoring an alternative learning strategy. Almost certainly this reflects an increasing need for combinatorial and hierarchic learning, even at the expense of more basic correlative learning strategies. These hominids were not getting smarter in any simple sense. They were likely getting dumber when it came to the correlative-associative learning that is so critical for solving problems posed by physical or social circumstances.

The cognitive abilities favored and enhanced by this evolutionary trend, however, were the sine qua non of symbol acquisition. Even learning the simplest symbolic relationships places heavy demands on these particular cognitive abilities. Attention to higher-order distributed associations and away from those based on temporal-spatial correlations tends to render these other forms of associative learning somewhat less efficient. It is difficult to imagine what other practical domain could benefit from such a shift in learning style. No other species evolved this ability, because incremental change in learning abilities that would enhance symbol acquisition would be counterproductive to learning in the absence of symbolic communication. It is hard to imagine, then, that

anything other than the significant advantages of symbolic communication (in whatever form) could account for selection pressures that would drive such an unusual course of brain evolution. Some simple symbolic communication must, therefore, have preceded and driven hominid brain evolution, not followed it.

This theory of brain-language coevolution forces us to entirely rethink hominid origins. The restructuring of the hominid brain was not sudden, nor was it merely a quantitative expansion. It took place incrementally (though I leave it to paleontologists to quibble about the number and size of the "increments," it at least can be certain that it was not a one or two step process) over the course of approximately 1.5 million years beginning approximately 2 million years ago with the species *Homo habilis*. The impetus behind this restructuring of the brain appears to have been the unusual nature of the cognitive demands imposed by symbolic communication, not some generalized demand on intellectual capacity. Selection for prefrontal expansion derives from the incredible demands symbol learning places on combinatorial and hierarchic learning processes.

This neurological adaptation does not directly account for the evolution of complex grammar and offers no support for the idea of a modular innate universal grammar. If anything, it suggests that the evolution of grammatical systems is at most a secondary issue. To the extent that symbolic associations are inherently and irreducibly combinatorial and hierarchic, any adaptation that increases the facility for producing and analyzing such relationships will contribute to the ability to become skilled at handling the sorts of computations that syntactic processes require. The human facility for constructing and analyzing complex sequential and hierarchic relationships may also offer some insight into other related abilities and predispositions nascent in human brains, from art and music to mathematics and game playing. It shows humans to be peculiarly unique among species, not just for their language abilities but for their odd style of thinking and learning.

REFERENCES

Barbas, H., and M.-M. Mesulam. 1981. "Organization of Afferent Input to Subdivisions of Area 8 of the Rhesus Monkey." *Journal of Comparative Neurology* 200: 407–31.

Blinkov, S., and I. Glezer. 1968. *The Human Brain in Figures and Tables*. New York: Plenum.

Deacon, Terrence W. 1984. "Connections of the Inferior Periarcuate Area in the Brain of *Macaca fascicularis*: An Experimental and Comparative Investigation of Language Circuitry and Its Evolution." Ph.D. diss., Harvard University.

_____. 1988. "Human Brain Evolution: II. Embryology and Brain Allometry." In *Intelligence and Evolutionary Biology*, ed. H. Jerison and I. Jerison. Berlin: Springer-Verlag.

_____. 1989. "The Neural Circuitry Underlying Primate Calls and Human Language." *Human Evolution* 4: 367–401.

_____. 1990a. "Fallacies of Progression in Theories of Brain Size Evolution." *International Journal of Primatology* 11: 193–236.

_____. 1990b. "Rethinking Mammalian Brain Evolution." *American Zoologist* 30: 629–705.
_____. 1992. "Brain-Language Co-evolution." In *The Evolution of Languages*, ed. J. Hawkins and M. Gel-Man. Redwood City, Calif.: Addison-Wesley.
Fuster, J. 1980. *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe.* New York: Raven.
Goldman-Rakic, Patricia R. 1977. "Circuitry of the Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory." *Handbook of Physiology*, ed. Stephen R. Geiger. Bethesda, Md.: American Physiological Society, pp. 373–418.
Guilford, J. 1967. *The Nature of Human Intelligence.* New York: McGraw-Hill.
Jacobsen, C. 1936. "Studies of Cerebral Function in Primates." *Comparative Psychology Monographs* 13: 1–68.
Kolb, B., and I. Whishaw. 1990. *Fundamentals of Human Neuropsychology.* 3d ed. New York: W. H. Freeman.
Linden, R. 1990. "Control of Neuronal Survival by Anomalous Targets in the Developing Brain." *Journal of Comparative Neurology* 294: 594–606.
Mishkin, M., and F. Manning. 1978. "Nonspatial Memory after Selective Prefrontal Lesions in Monkeys." *Brain Research* 143: 313–23.
Ojemann, G. A. 1979. "Individual Variability in Cortical Localization of Language." *Journal of Neurosurgery* 50: 164–69.
O'Leary, D. 1992. "Development of Connectional Diversity and Specificity in the Mammalian Brain by the Pruning of Collateral Projections." *Current Opinions in Neurobiology* 2: 70–77.
Pandya, D., and C. Barnes. 1987. "Architecture and Connections of the Frontal Lobe." In *The Frontal Lobes Revisited*, ed. E. Perecman. New York: IRBN Press.
Passingham, Richard. 1985. "Memory of Monkeys (*Macaca mulatta*) with Lesions in Prefrontal Cortex." *Behavioral Neuroscience* 99: 3–21.
Perecman, E., ed. 1987. *The Frontal Lobes Revisited.* New York: IRBN Press.
Petrides, M. 1982. "Motor Conditional Associative Learning after Selective Prefrontal Lesions in the Monkey." *Behavioral and Brain Research* 5: 407–13.
_____. 1985. "Deficits in Nonspatial Conditional Associative Learning after Periarcuate Lesions in Monkey." *Behavioral and Brain Research* 16: 95–101.
_____. 1986. "The Effect of Periarcuate Lesions in the Monkey on the Performance of Symmetrically and Asymmetrically Reinforced Visual and Auditory Go, No-Go Tasks." *Journal of Neuroscience* 6: 2054–63.
Piaget, Jean. 1952. *The Origins of Intelligence in Children.* New York: International Universities Press.
Purves, D. 1988. *Body and Brain. A Trophic Theory of Neural Connections.* Cambridge: Harvard Univ. Press.
Purves, D., and J. Lichtman. 1985. *Principles of Neural Development.* Sunderland, Mass.: Sinauer Associates.
Savage-Rumbaugh, E. S., and D. M. Rumbaugh. 1978. "Symbolization, Language and Chimpanzees: A Theoretical Reevaluation Based on Initial Language Acquisition Processes in Four Young *Pan troglodytes.*" *Brain and Language* 6: 265–.
Savage-Rumbaugh, E. S.; D. M. Rumbaugh; S. T. Smith; and J. Lawson. 1980. "Reference: The Linguistic Essential. *Science* 210: 922–25.
Stuss, D., and D. Benson. 1986. *The Frontal Lobes.* New York: Raven.