Arbib, M.A., 2002, The Mirror System, Imitation, and the Evolution of Language, in Imitation in Animals and Artifacts, (Chrystopher Nehaniv and Kerstin Dautenhahn, Editors), The MIT Press, to appear.

The Mirror System, Imitation, and the Evolution of Language

March 15, 2000

Michael Arbib
Computer Science Department and USC Brain Project
University of Southern California
Los Angeles, CA 90089-2520
arbib@pollux.usc.edu
http://www-hbp.usc.edu/

1 Introduction 2
2 The Mirror System Hypothesis: A New Approach to the Gestural Basis of Language 5
Stage 1: Grasping 5
Stage 2: Mirror Systems for Grasping 9
A Mirror System for Grasping in the Monkey 10
A Mirror System for Grasping in Humans 12
Learning, Mirroring, and Vocalization 12
Action = Movement + Goal/Expectation 13
Bridging from Action to Language: The Mirror-System Hypothesis 16
3 Beyond the Mirror: Further Hypotheses on the Evolution of Language 18
Language-Readiness 18
Stages 3 and 4: Simple and Complex Imitation Systems for Grasping 20
Stage 5: A Manual-Based Communication System 23
Stage 6: Speech 25
The Transition to *Homo sapiens* 27
A Multi-Modal System 28
Language Evolving 29
Deep Time 29
From Action-Object Frame to Verb-Argument Structure to Syntax and Semantics 31
4 Conclusion 33
References 36

A dance class in Santa Fe, Sept. 25, 1999:

The percussion is insistent. Dancers move in rows from the back of the hall towards the drummers at the front. From time to time, the mistress of the dance breaks the flow, and twice repeats a sequence of energetic dance moves. The dancers then move forward again, repeating her moves, more or less. Some do it well, others not so well.

Imitation involves, in part, seeing the instructor's dance as a set of familiar movements of shoulders, arms, hands, belly and legs. Many constituents are variants of familiar actions, rather than familiar actions themselves. Thus one must not only observe actions and their composition, but also novelties in the constituents and their variations. One must also perceive the overlapping and sequencing of all these moves and then remember the “coordinated control program” so constructed. Probably, memory and perception are intertwined.

As the dancers perform they both act out the recalled coordinated control program and tune it. By observing other dancers and synchronizing with their neighbors and the insistent percussion of the drummers, they achieve a collective representation that tunes their own, possibly departing from the instructor's original. At the same time, some dancers seem more or less skilled – some will omit a movement, or simplify it, others may replace it with their imagined equivalent. (One example: the instructor alternates touching her breast and moving her arm outwards. Most dancers move their arms in and out with no particular target.) Other changes are matters of motor rather than perceptual or mnemonic skill – not everyone can lean back as far as the instructor without losing balance.

These are the ingredients of imitation.

1 Introduction

I argue that the ability to imitate is a key innovation in the evolutionary path leading to language in the human and relate this hypothesis to specific data on brain mechanisms. The starting point is the discovery of the "mirror system" for grasping in monkey, a region in the monkey brain in which neurons active when the monkey executes a specific hand action are also active when the monkey observes another primate (human or monkey) carrying out that same action. In “Language Within Our Grasp”, Rizzolatti and Arbib (1998) showed that the mirror system in monkey is the homologue of Broca’s area, a crucial speech area in humans, and argued that this observation provides a neurobiological “missing link” for the long-argued hypothesis that primitive forms of communication based on manual gesture preceded speech in the evolution of language. Their “Mirror System Hypothesis” states that the matching of neural code for execution and observation of hand movements in the monkey is present in the common ancestor of monkey and human, and is the precursor of the crucial language property of parity, namely that an utterance usually carries similar meaning for speaker and hearer.[1] Here we refine this hypothesis by suggesting that imitation plays a crucial role in human language acquisition and performance, and that brain mechanisms supporting imitation were crucial to the emergence of *Homo sapiens*.

I stress that imitation – for me at least – involves more than simply observing someone else's movement and responding with a movement which in its entirety is already in one's own repertoire. Instead, I insist that imitation involves "parsing" a complex movement into more or less familiar pieces, and then performing the corresponding composite of (variations on) familiar actions. Note the insistence on "more or less familiar pieces" and "variations". Elsewhere (Arbib, 1981) I have introduced the notion of a coordinated control program, to show how a new behavior could be composed from an available repertoire of perceptual and motor schemas (the execution of a successful action will in general require perceptual constraints on the relevant movements). However, skill acquisition not only involves the formation of new schemas as composites of old ones, it also involves the tuning of these schemas to match a new set of conditions, to the point that the unity of the new schema may over-ride the original identity of the components. For example, if one is acquiring a tennis stroke and a badminton stroke through

[1] Since we will be concerned in what follows with sign language as well as spoken language, I ask the reader to understand that "speaker" and "hearer" may actually be using hand gestures rather than vocal gestures for communication. The claim will be that communication using hand gestures provided the scaffolding for converting a vocal repertoire comprising a limited set of species-specific calls into the open-ended production and perception of sequences of vocal gestures necessary for speech. In other words, speech did not evolve directly within the vocal-auditory domain, but rather was supported by the evolution of a system of manual gestures for communication.

imitation, the initial coordinated control program may be identical, yet in the end the very different dynamics of the tennis ball and shuttlecock lead to divergent schemas. Conversely, a skill may require attention to details not handled by the constituent schemas of the preliminary coordinated control program. Fractionation may be required, as when the infant progresses from "swiping grasps" at objects to the differentiation of separate schemas for the control of arm and hand movements. Later, the hand movement repertoire becomes expanded as one acquires such novel skills as typing or piano playing, with this extension matched by increased subtlety of eye-arm-hand coordination Thus we have three mechanisms (at least) to learn completely new actions: forming new constructs (coordinated control programs) based on familiar actions; tuning of these constructs to yield new encapsulated actions, and fractionation of existing actions to yield more adaptive actions as tuned, coordinated control programs of novel schemas.

Imitation, in general, requires the ability to break down a complex performance into a coordinated control program of pieces which approximate the pieces of the performance to be imitated. This then provides the framework in which attention can be shifted to specific components which can then be tuned and/or fractionated appropriately, or better coordinated with other components of the skill. This process is recursive, yielding both the mastery of ever finer details, and the increasing grace and accuracy of the overall performance.

I argue that what marks humans as distinct from their common ancestors with chimpanzees is that whereas the chimpanzee can imitate short novel sequences through repeated exposure, humans can acquire (longer) novel sequences in a single trial if the sequences are not too long and the components are relatively familiar. The very structure of these sequences can serve as the basis for immediate imitation or for the immediate construction of an appropriate response, as well as contributing to the longer-term enrichment of experience. Of course (as our Santa Fe dance example shows), as sequences get longer, or the components become less familiar, more and more practice is required to fully comprehend or imitate the behavior.

The next section summarizes the basic evidence for the Mirror System Hypothesis for the evolution of language. The rest of the paper will go “Beyond the Mirror” to suggest new considerations that refine the original hypothesis of the 1998 paper. The paper will take us through seven hypothesized stages of evolution,

1. grasping
2. a mirror system for grasping (i.e., a system that matches observation and execution),
3. a simple imitation system for grasping,
4. a complex imitation system for grasping,
5. a manual-based communication system, and
6. speech, which I here characterize as being the open-ended production and perception of sequences of vocal gestures, without implying that these sequences constitute a language
7. language.

At each stage, the earlier capabilities are preserved. Moreover, the addition of a new stage may involve enhancement of the repertoire for the primordial behaviors on which it is based.

Three key methodological points:

(a) We must understand the adaptive value of each of the first six stages without recourse to its role as a platform for later stages.

(b) We will distinguish between “language” and “language-readiness”, stressing that certain biological bases for language may not have evolved to serve language but were selected by other pressures, but then served as the basis for a process of individual discoveries driving cultural evolution which developed language to the richness we find in all present-day societies, from vast cities to isolated tribes. I will argue that the first six stages involved biological evolution that was completed with the emergence of *Homo sapiens*, but that the richness of language reflects cultural evolution with little if any change in the brain of *Homo sapiens* beyond that required to achieve speech in the limited sense described in (6) above.

(c) We will not restrict language to “that which is expressed in speech, or in writing derived therefrom.” By this I mean that language in its fullness may be expressed by an integration of speech, manual gestures and facial movements to which the written record can do at best partial justice.

The argument that follows involves two major sections, "The Mirror System Hypothesis: A New Approach to the Gestural Basis of Language" and "Beyond the Mirror: Further Hypotheses on the Evolution of Language". The first part reviews neurophysiological and anatomical data on Stage 1, Grasping, and Stage 2, Mirror Systems for Grasping, as well as outlining a computational model, the FARS (Fagg-Arbib-Rizzolatti-Sakata) model, for grasping, named for the modelers Andy Fagg and myself, and for the experimentalists Giacomo Rizzolatti and Hideo Sakata whose work anchors the model. The model shows show how the sight of an object may be processed to yield an appropriate action for grasping it, as well as to explain the shifting patterns of neural activity in a variety of brain regions involved in this visuomotor transformation. We then provide a conceptual analysis of how the brain may indeed use a *mirror system*, i.e., one which uses the same neural codes to characterize an action whether it is executed or observed by the agent. A mirror system for grasping in the monkey has been found in area F5 of premotor cortex, while data have been found consistent with the notion of a mirror system for grasping in humans in Broca's area, which is homologous to monkey F5 but in humans is most often thought of as a speech area. After a brief discussion of Learning in the Mirror System, and a conceptual analysis of the equation "Action = Movement + Goal/Expectation", we use the above data to bridge from action to language with the *Mirror-System Hypothesis*, namely that language evolved from a basic mechanism not originally related to communication: the mirror system for grasping with its capacity to generate and recognize a set of actions.

The second half of the paper then goes "Beyond the Mirror", offering further hypotheses on the evolution of language which take us up the hierarchy from elementary actions to the recognition and generation of novel compounds of such actions. The well-known linguist Noam Chomsky (e.g., 1975) has argued that since children acquire language rapidly despite the "poverty of the stimulus" therefore the basic structures of language are encoded in the brain, forming a Universal Grammar encoded in the human genome. For example, it is claimed that the Universal Grammar encodes the knowledge that a sentence in a human language could be ordered as Subject-Verb-Object, Subject-Object-Verb, etc., so that the child simply needs to hear a few sentences of his first language to "set the parameter" for the preferred order of that language. Against this, others have argued that in fact the child does have a rich set of language stimuli, and that there are now far more powerful models of learning than those that Chomsky took into account, allowing us to explain how a child might learn from its social interactions aspects of syntax which Chomsky would see as genetically prespecified. The reader may consult Lieberman (1991) for a number of arguments which counter Chomsky's view. Here I simply observe that many youngsters today easily acquire the skills of "Web surfing" and video-game playing despite a complete poverty of the stimulus, namely the inability of their parents to master these skills. I trust that no one would claim that the human genome contains a "Web-surfing gene"! Instead, we know the history of computers, and know that technology has advanced over the last 55 years to take us from an interface based on binary coding that only a trained scientist could master to a mouse-and-graphics interface so well adapted to human sensorimotor capabilities that a child can master it. My claim is that languages evolved similarly. Deacon (1997) makes a similar point, but blurs it somewhat in the subtitle of his book *The Symbolic Species: The co-evolution of language and the brain.* I agree that communication (but not of the richness that characterizes all present day languages) did provide part of the selective pressures that formed the brain of *Homo sapiens*, but still hold that much of what we regard as the nature of language was formed by a multitude of discoveries that post-dated the overall establishment of the human genome.

Note that the argument is over whether or not the "key grammatical structures of all possible human languages" are all pre-encoded in the human genome, to be selected by parameter setting in early childhood. There is no argument against the view that human evolution yielded genetic specification of some of the structures which *support* language. For example, the human larynx is especially well structured for the clear articulation of vocalization (see Lieberman 1991 for further details) and the human brain provides the necessary control mechanisms for this articulation. However, Lieberman and I reject Chomsky's view that many of the basic alternatives of grammatical structure of the world's current languages are already encoded in the human genome, so that the child's experience merely "sets parameters" to choose among prepackaged alternative grammatical structures. The counter-view which I espouse holds that the brain of the first *Homo sapiens* was "language-ready" but that it required many millennia of invention and cultural evolution for human societies to form human languages in the modern sense.

Given the emphasis on the recognition and generation of novel, hierarchically structured compounds of actions as a key to language, we next come to Stages 3 and 4, simple and complex imitation systems for grasping. With this, we move to a speculative scenario for how Stage 5, A Manual-Based Communication System, broke through the fixed repertoire of primate vocalizations to yield a combinatorially open repertoire, so that Stage 6, Speech, did not build upon the ancient primate vocalization system, but rather rested on the "invasion" of the vocal apparatus by collaterals from the communication system based on F5/Broca's area. In discussing the transition to *Homo sapiens,* I stress that our predecessors must have had a relatively flexible, open repertoire of vocalizations but this does not mean that they, or the first humans, had language. For Stage 7, the transition to language, I offer a scenario for the change from action-object frames to verb-argument structures to syntax and semantics. Finally, I briefly sketch the merest outline of a new approach to neurolinguistics based on these extensions of the mirror system hypothesis.

2 The Mirror System Hypothesis: A New Approach to the Gestural Basis of Language

Stage 1: Grasping

In this section and the next, we will use data on the monkey brain to ground our speculations on what might have been brain structures in the common ancestor of monkey and human that laid the basis for the evolution of a brain which could support imitation and the ability to discover, disseminate and use language as the basis for new patterns of communication. The task of the present section is to take us through Stage 1, grasping, of our hypothesized stages of evolution, reviewing relevant data and presenting useful grounding concepts provided by the FARS model.

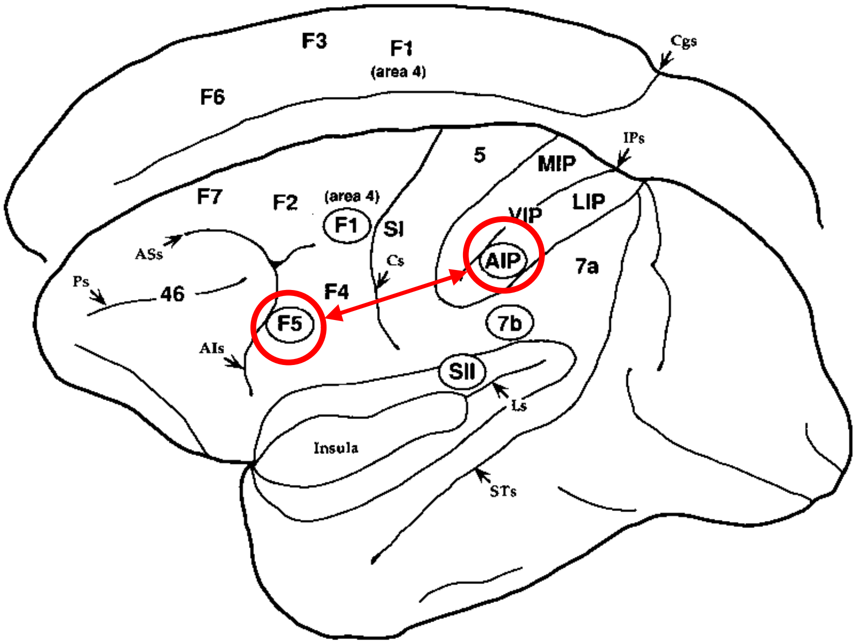

Figure 1. A side view of the left hemisphere of the macaque monkey brain dominates the figure, with a glimpse of the medial view of the right hemisphere above it to show certain regions that lie on the inner surface of the hemisphere. The central fissure is the groove separating area SI (primary somatosensory cortex) from F1 (more commonly called MI, primary motor cortex). Frontal cortex is the region in front of (in the figure, to the left of) the central sulcus. Area F5 of premotor cortex (i.e., the area of frontal cortex just in front of primary motor cortex) is implicated in the elaboration of "abstract motor commands" for grasping movements. Parietal cortex is the region behind (in the figure, to the right of) the central sulcus. The groove in the middle of the parietal cortex, the intra-parietal sulcus, is shown opened here to reveal various areas. AIP (the anterior region of the intra-parietal sulcus) processes visual information relevant to the control of hand movements and is reciprocally connected with F5.

The neurophysiological findings of the Sakata group on parietal cortex (Taira et al., 1990) and the Rizzolatti group on premotor cortex indicate that parietal area AIP (the Anterior Intra-Parietal sulcus) and ventral premotor area F5 in monkey (Rizzolatti et al., 1988) form key elements in a cortical circuit which transforms visual information on intrinsic properties of objects into hand movements that allow the animal to grasp the objects appropriately. See Figure 1 for the anatomy, and Jeannerod et al. (1995) for a review.

Motor information is transferred from F5 to the primary motor cortex (denoted F1 or M1), to which F5 is directly connected, as well as to various subcortical centers for movement execution. For example, neurons located in the rostral part of inferior area 6 (area F5) discharge during active hand and/or mouth movements (Di Pellegrino et al., 1994; Rizzolatti et al., 1996; Gallese et al., 1996). Moreover, discharge in most F5 neurons correlates with an action rather than with the individual movements that form it so that one may classify F5 neurons into various categories corresponding to the action associated with their discharge. The most common are: "grasping-with-the-hand" neurons, "grasping-with-the-hand-and-the-mouth" neurons, "holding" neurons, "manipulating" neurons, and "tearing" neurons. Rizzolatti et al. (1988) thus argued that F5 contains a "vocabulary" of motor schemas (Arbib, 1981). The situation is in fact more complex, and "grasp execution" involves a variety of loops and a variety of other brain regions in addition to AIP and F5.

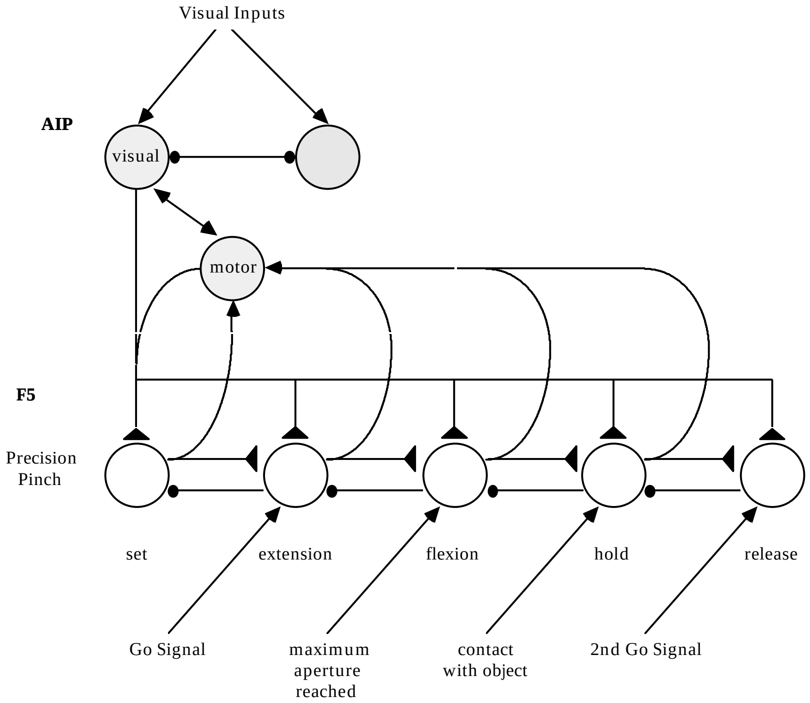

Figure 2. Hypothesized information flow in AIP and F5 in the FARS model during execution of the Sakata paradigm. This neural circuit appears as a rather rigid structure. However, we do not hypothesize that connections implementing the phasic behavior are hardwired in F5. Instead, we posit that sequences are stored in pre-SMA (a part of the supplementary motor area) and administered by the basal ganglia.

The FARS model (Fagg and Arbib 1998) makes clear certain conceptual issues that will be crucial at later stages of the argument. It provides a computational account of what we shall call the canonical system, centered on the AIP → F5 pathway, showing how it can account for basic phenomena of grasping. The highlights of the model are shown in Figures 2 and 3. Our basic view is that AIP whose cells encode (by a population code whose details are again beyond the present discussion) “affordances” for grasping from the visual stream and sends (neural codes for) these on to area F5 – *affordances* (Gibson, 1979) are features of the object relevant to action, in this case to grasping. In other words, vision here provides cues on how to interact with an object, rather than categorizing the object or determining its identity. As figure 2 shows, some cells in AIP are driven by feedback from F5 rather than by visual inputs so that AIP can monitor ongoing activity as well as visual affordances. Here we indicate the case in which the visual input has activated an affordance for a precision pinch, and we here show the AIP activity driving an F5 cell pool that controls the execution of a precision pinch. However, what we show is somewhat complicated because the circuitry is not for a single action, but for a behavior designed by Sakata to probe the time-dependence of activity in the monkey brain. In the Sakata paradigm, the monkey is trained to watch a manipulandum until a go signal instructs it to reach out and grasp the object. It must then hold the object until another signal instructs it to release the object.

In Figure 2, cells in AIP instruct the set cells in F5 to prepare for execution of the Sakata protocol using a precision pinch. Activation of each pool of F5 cells not only instructs the motor apparatus to carry out the appropriate activity (these connections are not shown here), but also primes the next pool of F5 neurons (i.e. brings the neurons to just below threshold so they may respond quickly when they receive their own go signal) as well as inhibiting the F5 neurons for the previous stage of activity. Thus, the neurons which control the extension phase of the hand shaping to grasp the object are primed by the set neurons, and they reach threshold when they receive the first go signal, at which time they inhibit the set neurons and prime the flexion neurons. These pass threshold when receiving a signal that the hand has reached its maximum aperture; the hold neurons once primed will become active when receiving a signal that contact has been made with the object; and the primed release neurons will command the hand to let go of the object once they receive the code for the second go signal.

Karl Lashley (1951) wrote of "The Problem of Serial Order in Behavior", a critique of stimulus-response approaches to psychology. He noted that it would be impossible to learn a sequence like A, B, A, C as a stimulus-response chain because the association "completing A triggers B" would then be interfered with by the association "completing A triggers C", or would dominate it to yield an infinite repetition of the sequence A, B, A, B,.... The generally adopted solution is to segregate the learning of a sequence from the circuitry which encodes the unit actions, the latter being F5 in the present study. Instead, another area (possibly the part of the supplementary motor area called pre-SMA; Rizzolatti, Luppino and Matelli 1998) has neurons whose connections encode an "abstract sequence" Q1, Q2, Q3, Q4, with sequence learning then involving learning that activation of Q1 triggers the F5 neurons for A, Q2 triggers B, Q3 triggers A again, and Q4 triggers C. In this way, Lashley's problem is solved. Other studies lead us to postulate that the storage and administration of the sequence (inhibiting extraneous actions, while priming imminent actions) is carried out by the portion of the supplementary motor area known as pre-SMA and the basal ganglia, respectively.[2]

[2] Here a caveat is in order. Analysis in neuroscience often goes to the extreme of focusing on one circuit or brain region and trumpeting it as "the" circuit implementing some specific function X, just as if we were to claim that Broca's area is "the" region for language. The other extreme is "holism", stressing that X may involve dynamic activity integrating all areas of the brain. Holism may be correct, but seems to me useless as a guide to understanding. My preferred approach is a compromise. I stress that any "schema" or function X involves the cooperative computation of many neural circuits (see Chapter 3 of Arbib, Érdi, and Szentágothai, 1998), but then proceed by using a survey of the literature to single out a few regions for which good data are available correlating neural activity with the performance of function X. I then seek the best model available (from the literature, by research in my own group, or by a combination of both) which can yield a causally complete neural model which approximates both the patterns of behavior seen in animals performing X and the neural activity observed during these performances. As time goes by, the model may yield more and more insight into the data on X (some of which may be new data

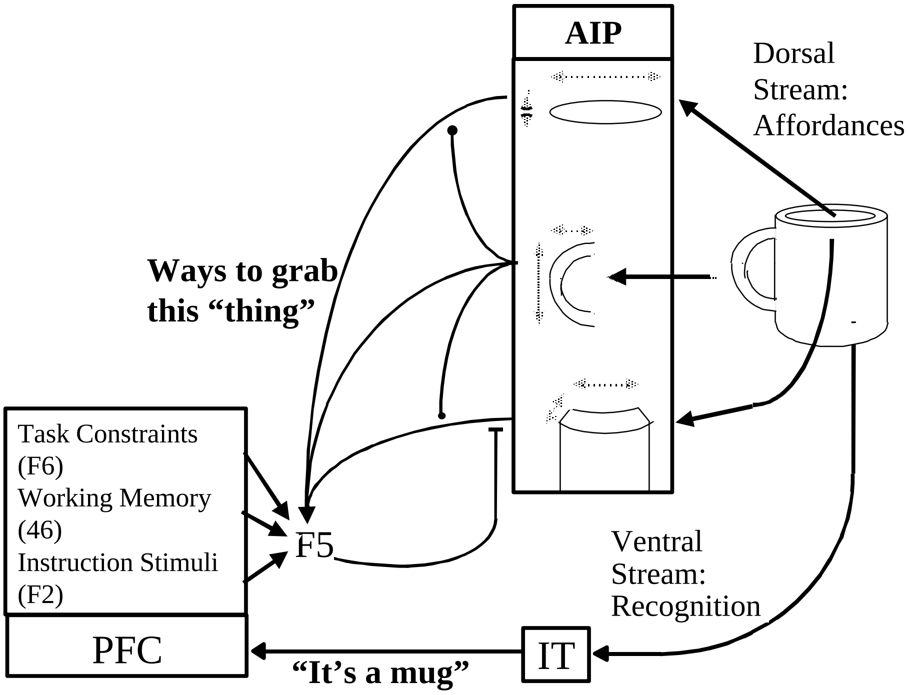

Figure 3. The role of IT (inferotemporal cortex) and PFC (prefrontal cortex) in modulating F5's selection of an affordance.

Note that the solution offered here is a specific case of a far more general solution to Lashley's problem, based on learning a finite automaton, rather than just a sequence (Arbib, 1969; Dominey, Arbib & Joseph, 1995). In the general situation we have a set X of inputs, a set Y of outputs, and a set Q of states. These are augmented by a state-transition function δ: $Q \times X \rightarrow Q$, and an output function β: $Q \rightarrow Y$. When in state q the automaton emits output $\beta(q)$; on receiving input x, it then changes state to $\delta(q,x)$.

We now turn to the crucial role of IT (inferotemporal cortex) and PFC (prefrontal cortex) in modulating F5's selection of an affordance (Figure 3). Here, the dorsal stream (from primary visual cortex to parietal cortex) carries amongst other things the information needed for AIP to recognize that different parts of the object can be grasped in different ways, thus extracting affordances for the grasp system which (according to the FARS model) are then passed on to F5 where a selection must be made for the actual grasp. The point is that the dorsal stream does not know "what" the object is, it can only see the object as a set of possible affordances. The ventral stream (from primary visual cortex to inferotemporal cortex), by contrast, is able to recognize what the object is. This information is passed to prefrontal cortex which can then, on the basis of the current goals of the organism and the recognition of the nature of the object, bias F5 to choose the affordance appropriate to the task at hand. In particular, the FARS

whose collection was prompted by the modeling), whether by conducting new analyses of the model, increasing the granularity of description of certain brain regions, or in extending the model to include representations of more brain regions. Other challenges come from integrating separate models developed to explain distinct functions X, Y, and Z and integrating them to derive a single model of interacting brain regions able to serve all these functions. Thus the above paragraph should not be interpreted as saying "the only parts of the brain relevant to modeling sequential behavior are pre-SMA and the related portions of basal ganglia", but rather that as we extend our F5-based model of sequences of hand movements, our current targets for detailed modeling are pre-SMA and basal ganglia (Bischoff and Arbib, 2000). As we come to understand more fully the roles of

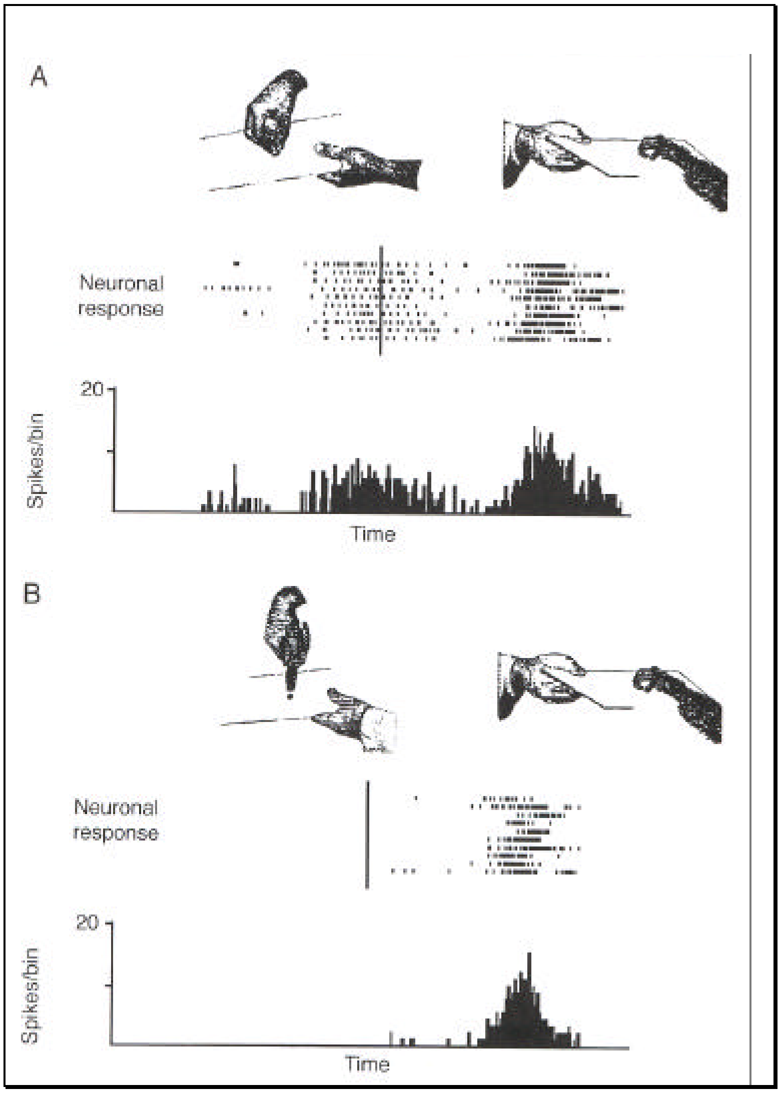

Figure 4. Example of a mirror neuron. Upper part of each panel: behavioral situations. Lower part: neuron's responses. The firing pattern of the neuron on each of a series of consecutive trials is shown above the histogram which sums the response from each trial. A (left): The experimenter grasps a piece of food with his hand, then moves it toward the monkey, who (A, right) at the end of the trial, grasps it. The neuron discharges during observation of the experimenter's grasp, ceases to fire when the food is given to the monkey and discharges again when the monkey grasps it. B (left): When the experimenter grasps the food with an unfamiliar tool, the neuron does not respond, but the neuron again discharges when the monkey grasps the food. The rasters are aligned with the moment when the food is grasped (vertical line). Each small vertical line in the rasters corresponds to a spike. Histogram bin width: 20 ms. Ordinates, spikes/bin; abscissae, time.

these brain regions, we can then turn to more extensive models. However, I fear that if we try to model "everything all at once" we will understand nothing, since the map would then be co-extensive with the whole territory (Borges, 1975).

model represents the way in which F5 may accept signals from areas F6 (pre-SMA), 46 (dorsolateral prefrontal cortex), and F2 (dorsal premotor cortex) to respond to task constraints, working memory, and instruction stimuli, respectively (see Fagg and Arbib 1988 for more details).

Stage 2: Mirror Systems for Grasping

Our task now is to describe neurological data which show that the above "execution system" for grasp in monkeys includes a subset of neurons – the *mirror neurons* – which form an "observation system", then review preliminary data that there is a homologous system in Broca's area, a language area of humans.

A Mirror System for Grasping in the Monkey

Further study of F5 revealed something unexpected – a class of F5 neurons that discharge not only when the monkey grasped or manipulated objects, but also when the monkey observed the experimenter make a gesture similar to the one that, when actively performed by the monkey, involved activity of the neuron. Neurons with this property are called "mirror neurons" (Gallese et al., 1996). Movements yielding mirror neuron activity when made by the experimenter include placing objects on or taking objects from a table, grasping food, or manipulating objects. Mirror neurons, in order to be visually triggered, require an interaction between the agent of the action and the object of it. The simple presentation of objects, even when held by hand, does not evoke the neuron discharge. An example of a mirror neuron is shown in Figure 4. In A, left side, the monkey observes the experimenter grasping a small piece of food. The tray on which the food is placed is then moved toward the monkey and the monkey grasps the food (right side of the figure). The neuron discharges both during grasping observation and during active grasping. B illustrates that when the food is grasped with a tool and not by hand the neuron remains silent. The majority of mirror neurons are selective for one type of action, and for almost all mirror neurons there is a link between the effective observed movement and the effective executed movement. A series of control experiments ruled out interpretations of mirror neurons in terms of monkey's vision of its own hand, food expectancy, motor preparation for food retrieval or reward (Gallese et al., 1996).

The response properties of mirror neurons to visual stimuli can be summarized as follow. Mirror neurons do not discharge in response to simple presentation of objects even when held by hand by the experimenter. They require a specific action – whether observed or self-executed – to be triggered. The majority of them respond selectively in relation to one type of action (e.g., grasping). This congruence can be extremely strict, that is the effective motor action (e.g., precision grip) coincides with the action that, when seen, triggers the neuron (e.g., again precision grip). For other neurons the congruence is broader. For them the motor requirement (e.g., precision grip) is usually stricter than the visual (any type of hand grasping, but not other actions). All mirror neurons show visual generalization. They fire when the instrument of the observed action (usually a hand) is large or small, far from or close to the monkey. They also fire even when the action instrument has shapes as different as those of a human or monkey hand. A few neurons respond even when the object is grasped by the mouth. The actions most represented are: grasp, manipulate, tear, put an object on a plate. Mirror neurons also have (by definition) motor properties. However, not all F5 neurons respond to action observation. We thus distinguish mirror neurons, which are active both when the monkey performs certain actions and when the monkey observes them performed by others, from canonical neurons in F5 which are active when the monkey performs certain actions but not when the monkey observes actions performed by others. Mirror neurons receive input from the PF region of parietal cortex encoding observations of arm and hand movements. This is in contrast with the canonical F5 neurons which receive object-related input from AIP. It is the canonical neurons, with their input from AIP, that are modeled in the FARS model.

The monkey may have a limited "hard-wired" repertoire of basic grasps, such as the precision pinch and the power grasp. These capabilities may then be extended through learning:

1) Developing a further set of useful grasps (extending the repertoire of actions for canonical F5 neurons);

2) Learning to associate view of one's hand with grasp and object and matching this to views of others grasping (linking F5 mirror neurons with the appropriate visual preprocessing and F5 canonical neurons to match the representations of self-generated actions with similarly goal-oriented actions executed by others). An interesting anecdote from the Rizzolatti laboratory (unpublished) is suggestive for further analysis: When a monkey first sees the experimenter grasp a raisin using a pair of pliers, then his mirror neurons will not fire. However, after many such experiences, the monkey's mirror neurons encoding precision grip will fire when he sees the pliers used to grasp a raisin – the initially novel performance has been characterized as a familiar action.

In summary, the properties of mirror neurons suggest that area F5 is endowed with an observation/execution matching system: When the monkey observes a motor act that resembles one in its movement repertoire, a neural code for this action is automatically retrieved. This code consists in the activation of a subset, the mirror neurons, of the F5 neurons which discharge when the observed act is executed by the monkey itself.

A Mirror System for Grasping in Humans

The notion that a mirror system might exist in humans was tested by two PET[3] experiments (Rizzolatti et al., 1996; Grafton et al., 1996b). The two experiments differed in many aspects, but both compared brain activation when subjects observed the experimenter grasping a 3-D object against activation when subjects simply observed the object. Grasp observation significantly activated the superior temporal sulcus (STS), the inferior parietal lobule, and the inferior frontal gyrus (area 45). All activations were in the left hemisphere. The last area is of especial interest – areas 44 and 45 in left hemisphere of the human constitute Broca's area, a major component of the human brain's language mechanisms. Although there is no dataset yet that shows the same activated voxels for grasping execution and grasping observation in Broca's area, such data certainly contribute to the growing body of indirect evidence that there is a mirror system for grasping in Broca's area.

F5 in monkey is generally considered (Rizzolatti and Arbib 1998) to be the homologue of Broca's area in humans, i.e., it can be argued that these areas of monkey and human brain are related to the same region of the common ancestor. Thus, the cortical areas active during action observation in humans and monkeys correspond very well. Taken together, human and monkey data indicate that in primates there is a fundamental mechanism for action recognition: we argue that individuals recognize actions made by others because the neural pattern elicited in their premotor areas (in a broad sense) during action observation is similar to a part of that internally generated to produce that action. This mechanism in humans is circumscribed to the left hemisphere.

[3] PET (Positron Emission Tomography) and fMRI (functional Magnetic Resonance Imaging) are two methods which allow one to measure regional Cerebral Blood Flow (rCBF) in the brain. Since rCBF is correlated with some aspects of neural activity, both

Primate Vocalization

Monkeys exhibit a primate call system (a limited set of species-specific calls) and an oro-facial (mouth and hand) gesture system (a limited set of gestures expressive of emotion and related social indicators), as shown in Figure 5, which includes a linkage between the two systems to stress that communication is inherently multi-modal. (Body posture also plays a role in social communication, not emphasized here.) This communication system is *closed* in the sense that it is restricted to a specific repertoire. This is to be contrasted with human languages which are *open* in two senses: (i) the language is generative or productive, being made up of words and grammatical markers that can be combined in diverse ways to yield an essentially unbounded stock of sentences (you are able to comprehend this sentence even though you have neither seen nor heard it before); and (ii) new words (such as "computer" and "subsidiarity") may always be added to expand the scope of the language.[4]

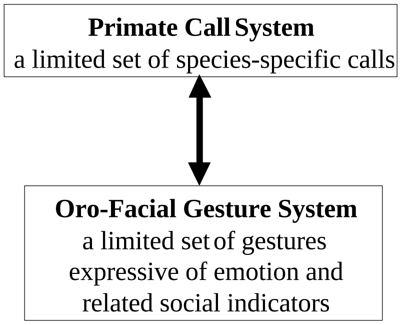
Primate Call System
a limited set of species-specific calls
Oro-Facial Gesture System
a limited set of gestures
expressive of emotion and
related social indicators

Figure 5. The communication system for the monkey is based on a finite set of calls ("vocal gestures") and oro-facial gestures.

What is to be stressed here is that

(i) combinatorial properties for the openness (productivity) of utterances are virtually absent in basic primate calls and oro-facial communication, even though individual calls may be graded.

(ii) the neural substrate for primate calls is in a region of cingulate cortex distinct from F5, which we have seen to be the monkey homologue of human Broca's area.

Our challenge in charting the evolution of human language, which for most humans is so heavily intertwined with speech, is thus to understand why it is F5, rather than the cingulate area already involved in monkey vocalization, that is homologous to Broca's area's substrate for language. But before proceeding to Stage 3, we need to discuss in more detail the nature of the mirror system in monkey which, we presume, carries over into the mirror systems of chimpanzee and human, but which receives further refinements in these species.

methods can be used as the basis of statistical procedures which generate maps showing which brain regions are significantly more active when a human is performing task A rather than task B.

[4] Although each human language is open as to, e.g., nouns and verbs, it is (almost) closed with respect to the stock of prepositions and grammatical markers.

Action = Movement + Goal/Expectation

What makes a movement into an action is that (i) it is associated with a goal, and (ii) initiation of the movement is accompanied by the creation of an expectation that the goal will be met. To the extent that the unfolding of the movement departs from that expectation, to that extent will an error be detected and the movement modified. In other words, an individual performing an action is able to predict its consequences and, therefore, the action representation and its consequences are associated. Thus a "grasp" involves not only a specific cortical activation pattern for the preshape and enclose movements, but also expectations concerning making appropriate contact with a specific object. Elsewhere (Arbib and Rizzolatti, 1997), we have asserted that "an individual making an action 'knows' what action he is performing to the extent that he predicts the consequences of his pattern of movement" but we must be very careful to distinguish "knowledge of action" in the sense of "has a neural representation of Movement + Goal/Expectation" from "has a representation that corresponds to a human's conscious awareness of 'what s/he is doing' ". Indeed, the FARS model contains mechanisms for creating and monitoring expectations even though it only models canonical F5 neurons, not mirror neurons. However, the creation of an expectation associated with one's own action is quite distinct from inferring the action of another from a glimpse of the movement involved.

The data presented earlier show that a major evolutionary development has been established in primates: the individual can recognize ("understand") the actions made by others in the sense that the neural pattern elicited by those actions is similar to that generated by him in doing the action. We suggest that the evolution of mirror neurons extended "knowing" from the individual to the social. Further evolution was required for such a system to mediate imitation. In later sections, we will discuss the importance of imitation not only in and for itself, but also as a crucial step toward the skills needed to mediate language (evolving a "language-ready brain").

It is probably the case that "understanding" involves the cooperation of many brain systems, and cannot be reduced to just the activity in a subset of the monkey's F5 neurons. Similarly for imitation. To say that firing of neurons in the monkey's mirror system is "predictive" means that it "creates a neural representation of a potential future state". Again, while mirror activity may be the key to "understanding", that activity in isolation from other brain systems encodes the ability "to match an external (unknown) event to an internal (known) event". I am not saying that the monkey has no awareness of the results of its actions or is incapable of knowing what another monkey is doing (though I would argue that human language and other cognitive abilities make our awareness very different from that of monkeys; Arbib, 2000), only that F5 activity in isolation is insufficient to mediate that awareness.

Many authors have suggested that language and understanding are inseparable, but our experience of scenery and sunsets and songs and seductions makes clear that we humans understand more than we can express in words. Some aspects of such awareness and understanding, then, may certainly be available to animals who do not possess language. Of course, this does not deny the crucial point that our development, as "modern" humans, i.e., as individuals within a language-based society, greatly extends our understanding beyond that possible for non-humans or for humans raised apart from a language community. Conversely, of course, other species are aware of aspects of their environment and society that we humans can at best dimly comprehend.

Two caveats should be noted:

(i) There is no claim that this mirroring is limited to primates. It is likely that an analogue of mirror systems exists in other mammals, especially those with a rich and flexible social organization. Moreover, the evolution of the imitation system for learning songs by male songbirds is divergent from mammalian evolution, but for the neuroscientist there are intriguing challenges in plotting the similarities and differences in the neural mechanisms underlying human language and birdsong.

(ii) The recognition of consequences may extend to actions beyond the animal's own repertoire, and may in some such cases involve mechanisms not much more complex than classical conditioning, rather than invoking a mirror system. For example, dogs can recognize the consequences of a human's use of a can opener (the sound of the can opener becomes associated with the subsequent presentation of the dog-food from the can) without having a motor program for opening cans, let alone mirror neurons for such a program.

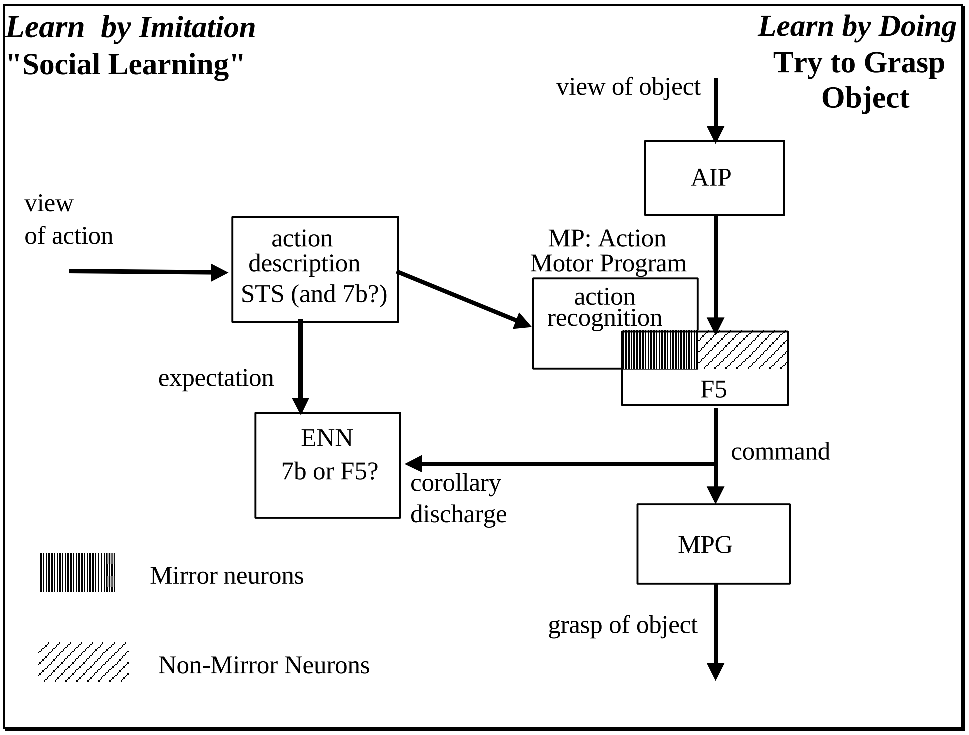

Figure 6. An integrated conceptual framework for analysis of the role of F5 in grasping. The right hand, vertical, path is the **execution system** from "view of object" via AIP and F5 to the motor pattern generator (MPG) for grasping a (seen) object. The loop on the left of the figure provides mechanisms for imitating observed actions in such a way as to create expectations which enable the visual feedback loop to serve both for "social learning" (i.e., learning an action through imitation of the actions of others) and also for (delayed) error correction during, e.g., reaching towards a target. It combines the **observation matching system** from "view of action" via action description (STS) and action recognition (mirror neurons in F5 and possibly 7b) to a representation of the "command" for such a action, and the **expectation system** from an F5 command via the expectation neural network ENN to MP, the motor program for generating a given action. The latter path may mediate a comparison between "expected action" and "observed action" in the case of the monkey’s self-generated movement.

Figure 6 presents a conceptual framework for analysis of the role of F5 in grasping. This combines mechanisms for (1) grasping a seen object (the right hand path from "view of object" to "grasp of object"); and (2) imitating observed gestures[5] in such a way as to create expectations which not only play a role in "social learning" but also enable the visual feedback loop to eventually serve for (delayed) error correction during, e.g., reaching towards a target (the loop on the left of the figure). [A more detailed model, with explicit learning rules, is currently being developed (Oztop and Arbib, 2000).]

The Expectation Neural Network (ENN) is the "Direct Model" of Command → Response which transforms the command into a code for the response. When the animal gives a command (i.e., brain regions issue the neural signals that initiate a movement), ENN generates the expected neural code for the visual signal generated by the resulting gesture. This is different from the FARS model (Fagg and Arbib 1998) which creates sensory expectations of the result of the movement, such as "the feel of the object when grasped", which are "private" to the animal. Here, we look at "public" symptoms of the ongoing movement. The key to the mirror system is that it brings those symptoms for self-movement together with those for other-movement in generating, we claim, a code for "action" (movement + goal) and not just for movement alone. However, there is a subsidiary problem here, namely recognizing which "symptoms" of self-movement correspond to which symptoms of other-movement, since the retinal display for, say, the hand-movement of one's self or another is radically different. In any case, we explicitly label the input to ENN, a copy of the motor command, as a corollary discharge. By contrast, the Motor Program MP provides an "Inverse Model" of Command → Response, going from a desired response to a command which can generate it.

Where are the various stages forming the imitation loop? Here are two, admittedly speculative, possibilities. The first is that the various model stages are located in different anatomical areas. In this case the inverse model which converts the view of an action to a corresponding command could be located along the path leading from STS to F5 (possibly via 7b). The reciprocal path from F5 to superior temporal sulcus would provide the direct model, ENN. It is equally probable, however, that both ENN and MP are located in F5 and the interplay between stages occurs entirely within F5. If the latter interpretation is accepted, the role of STS areas would be that of giving a merely "pictorial", though highly elaborated description, of actions - with the observation/execution system entirely located in the frontal lobe.

The integrated model of Figure 6 thus relates the "grasp an object" system to the "view an action" system. The expectation network is driven by F5 irrespective of whether the motor command is "object-driven" (via AIP) or "action-driven". It thus creates expectations both for what a hand movement will look like when "object-driven" (an instrumental action directed towards a goal) or "action-driven" (a "social action" aimed at making a self-generated movement approximate – by some criterion that does not match body-centered localization – an observed movement). The right hand path of Figure 6 exemplifies "learning by doing", refining a crude "innate grasp" - possibly by a process of reinforcement learning, in which the success/failure of the grasp acts as positive/negative reinforcement. The left hand path of Figure 6 exemplifies another mode of learning (the two may be sequential or contemporary) which creates expectations about actions as well as exemplifying "social learning" based on imitation of actions made by others.

[5] The present discussion is preliminary. A full discussion of imitation involves the recognition that a novel action can be approximated by a composite of variations of familiar actions, and with the improvement of that approximation through practice (recall the Santa Fe dance example). I ask the reader to recognize here that much more work needs to be done to tease apart the neural mechanisms serving for recognition in this extended sense.

Bridging from Action to Language: The Mirror-System Hypothesis

Before proceeding further, we must distinguish "pragmatic action" in which the hands are used to interact physically with objects or other creatures, and "gestures" (both manual and vocal) whose purpose is communication. Our assumption is that monkeys use hand movements only for pragmatic actions. The mirror system allows other monkeys to understand these actions and act on the basis of this understanding. Similarly, the monkey's oro-facial gestures register emotional state, and primate vocalizations can also communicate something of the current situation of the monkey. The monkey exhibits what might be called "involuntary communication" of its current internal state or situation either through its observable actions or through a fixed species-specific repertoire of oro-facial gestures and vocal gestures. We will develop the hypothesis that the mirror system made possible (but in no sense guaranteed) the evolution of the displacement of hand movements from pragmatic action to gestures that can be controlled "voluntarily".

Hewes (1973), Corballis (1991, 1992), Kimura (1993), and Armstrong et al. (1995) are among those who have long argued that communication based on manual gesture played a crucial role in human language evolution, preceding communication by speech. In this regard, we stress that the "openness" or "generativity" which some see as the hallmark of language (i.e., its openness to new constructions, as distinct from having a fixed repertoire like that of monkey vocalizations) is present in manual behavior which can thus supply the evolutionary substrate for its appearance in language. Kimura (1993) argues that the left hemisphere is specialized not for language, but for complex motor programming functions which are, in particular, essential for language production. However, it is not clear (i) whether hominid evolution yielded circuitry used for all forms of complex motor control, so that whether or not specific neurons are involved in control of language behavior is purely a result of experience-based learning, or (ii) evolution at some stage duplicated the circuitry used for complex motor control, with one copy then becoming specialized for the type of motor sequencing mechanisms needed for language (e.g., those involving the coordinated control of larynx, hands and facial expression) while the other copy remained available for other forms of coordinated behavior. If, in case (ii) the two circuits remained adjacent or even overlapping, it would make lesions which dissociate language performance from other forms of motor control very rare.

With our understanding that the mirror system in monkey is the homologue of Broca’s area in humans, we can now appreciate the central hypothesis of “Language Within Our Grasp” (Rizzolatti and Arbib, 1998), namely that this homology provides a neurobiological “missing link” for the long-argued hypothesis that gestural communication (based on manual gesture) preceded speech in the evolution of language. Rizzolatti and Arbib's novel tenet is that the parity requirement for language in humans - what counts for the speaker must count for the hearer - is met because of:

The Mirror-System Hypothesis: Language evolved from a basic mechanism *not* originally related to communication: the *mirror system for grasping* with its capacity to generate *and* recognize a set of actions.[6]

However, it is important to be quite clear as to what the Mirror System Hypothesis does not say

[6] The orofacial mirror neurons seem to be irrelevant here because our argument is that it is the manual system that provides the combinatorial richness (consider the endless possibilities of grooming, feeding, and other forms of manipulation) that leads us toward language in a way in which isolated oro-facial gestures and primate vocal calls do not.

(i) It does not say that having a mirror system is equivalent to having language. Monkeys have mirror systems but do not have language, and we expect that many species have mirror systems for varied socially relevant behaviors.

(ii) It is unclear whether the mirror system for grasping is sufficient for the copying of actions. It is one thing to recognize an action using the mirror system, it is another thing to use that representation as a basis for repeating the action. In an earlier version of this paper, I stated that "monkeys seldom if ever repeat an action but the mirror system helps them recognize the actions of others as a basis for social cooperation and competition." However, Judy Cameron (personal communication) offers the following observation from the Oregon Regional Primate Research Center: "Researchers at the Center had laboriously taught monkeys to run on a treadmill as a basis for tests they wished to conduct. It took five months to train the first batch of monkeys in this task. But they then found that if they allowed other monkeys to observe the trained monkeys running on a treadmill, then the naïve monkeys would urn successfully the first time they were placed on a treadmill." This is not evidence that the monkey mirror system for grasping is part of a system for imitation of hand movements, but does render this likely. We need new experiments to test this possibility and determine to what extent others brain regions besides F5 are required to exploit "mirroring" for the purpose of imitation.

However, I stress again that the ability to copy *single* actions is just the first step towards imitation, since we have earlier insisted that imitation involves "parsing" a complex movement into more or less familiar pieces, and then performing the corresponding composite of (variations on) familiar actions. The subtleties in going from "recognizing a familiar action" to "imitating a complex behavior based on an interweaving of variations on familiar actions" were illustrated in the opening description of a dance class in Santa Fe, and it is such observations that challenge us to go "beyond the mirror", i.e., beyond the recognition of single actions by the mirror system, in later sections of this paper. Note, too, the observation that the synchronization of movements in a group of dancers is enhanced by the rhythm of the music shows that imitation requires the ability for multi-modal associative learning, in this case matching rhythmic auditory and locomotor patterns.

(iii) It does not say that language evolution can be studied in isolation from cognitive evolution more generally. In using language, we make use of, for example, negation, counterfactuals, and verb tenses. But each of these linguistic structures is of no value unless we can understand that the facts contradict an utterance, and can recall past events and imagine future possibilities.

3 Beyond the Mirror: Further Hypotheses on the Evolution of Language

Having established the basic Mirror System Hypothesis, we now go "beyond the mirror" to discuss possible stages in the evolution from a monkey-like mirror system to the human capacity for language. We first distinguish language-readiness from "language hard-wired into the brain", and then examine the next four stages of posited biological evolution – 3 and 4. a simple and complex imitation system for grasping; 5. a manual-based communication system, and 6. speech (in the sense of the ability to produce and comprehend sequences of vocal gestures; I distinguish this from the capacity for spoken language in the modern sense) – which, I claim, provided *Homo sapiens* with a language-ready brain. The argument in part parallels, in part extends, that of Rizzolatti and Arbib (1998). However, I shall also argue that stage 7, the emergence of language, required many millennia of cultural evolution for our ancestors to extend earlier forms of hominid vocal communication into the complex communication systems that we recognize as human languages.

Language-Readiness

Ease of acquisition of a skill does not imply genetic encoding of the skill per se: The human genome does not encode strategies for exploring the Internet or playing video games. But computer technology has evolved to match the preadaptations of the human brain and body.

The human brain and body evolved in such a way that we have hands, larynx and facial mobility suited for generating gestures that can be used in language, and the brain mechanisms needed to produce and perceive rapidly generated sequences of such gestures. In this sense, the human brain and body is *language-ready*.

We thus reframe the old question: “How did language evolve?” as two questions:

1. "What really evolved by natural selection? Brains “equipped” with Language … or Language-Readiness?"

2. "How do we move beyond the mirror system to map changes in the evolutionary tree of primates & hominids in a *variety* of brain structures relevant to language readiness and cognition?"

(A third question, beyond the scope of this article, addresses the dynamics of language on multiple time-scales: "How can the study of language acquisition and of historical linguistics help tease apart biological and cultural contributions to the mastery of language by present-day humans?")

To proceed, we list several criteria for a set of "utterances" to constitute a language in the sense of a present-day human language. This list will help guide our understanding of what it means for the human brain to "have language" or "be language-ready". Note that nothing in this list rests on the medium of exchange of the language, applying to spoken language, sign language or written language, for example.[7]

Naming: The ability to associate an arbitrary symbol with a class of events, objects and actions, etc.;

Parity (*Mirror Property*): What counts for the speaker (or producer) must count for the listener (or hearer);

Hierarchical Structuring: Production and recognition of constituents with sub-parts;

Temporal Ordering: Temporal activity coding these hierarchical structures;[8]

Recursivity: The ability to build utterances using recursive rules, e.g., "a noun phrase is still a noun phrase if you insert an (or another) adjective", in which the generic construct appears as a term in its own definition.

Beyond the Here-and-Now: Verb tenses (or other language constructs for representing past events or future possibilities) demand neural machinery (language-readiness) to recall past events or imagine future ones;

Lexicon, Syntax, and Semantics move us into “language proper”, successfully matching syntactic structures to semantic structures; and

Learnability: To qualify as a human language, a set of symbolic structures must be learnable by most human children.

[7] However, it is only in the case of written language that the boundaries which separate one "word" from another are clearly represented. Even here, the division is somewhat arbitrary, e.g., "can not" versus "cannot" versus "can't".

[8] "Parsing" is the problem of inferring from a linear string of words the hierarchical structure it represents – e.g., determining which words go together to form the subject of the sentence. Ambiguity arises when parsing does not yield a unique structure, as well as when constituent words have multiple meanings. Ambiguity is the enemy of parity. I believe that the structure of each

Our quest to explore the hypothesis that the mirror system provided the basis for the evolution of human language(-readiness) will next lead us to argue that "imitation" takes us beyond the "basic" mirror system for grasping, and that the ability to "acquire novel sequences if the sequences are not too long and the components are relatively familiar" takes us a step further. This leads us to the questions:

What were the further biological changes supporting language-readiness?

What were the cultural changes extending the utility of language as a socially transmitted vehicle for communication and representation?

How did biological and cultural change interact “in a spiral” prior to the emergence of *Homo sapiens*?

Our approach to these issues builds, as I have said on the hypothesis (Rizzolatti and Arbib, 1998) that the homology of the mirror system in monkey with Broca’s area in humans provides a neurobiological “missing link” for the long-argued hypothesis that communication based on manual gesture preceded speech in the evolution of language. It is important here to distinguish "communication using hand gestures" from "sign language". The former can take the primitive form that preceded the evolution of speech; the latter is the use of manual gestures (as distinct from the vocal gestures of spoken language) in a fully expressive human language. My hypothesis is that a sign language exploits many of the same brain mechanisms as does a spoken language, and that the subtlety of modern sign languages rests on a long process of cultural co-evolution with spoken languages across many millennia following the biological emergence of *Homo sapiens*.

In this regard, let me briefly note the issue of non-human use of "sign language". The most impressive example for an ape is the case of the bonobo (pigmy chimpanzee) Kanzi. Savage-Rumbaugh et al. (1998) report that Kanzi and a 2.5 year-old girl were tested on their comprehension of 660 sentences phrased as simple requests (presented once). Kanzi was able to carry out the request correctly 72% of the time, whereas the girl scored 66% on the same sentences and task. Two observations. (a) No non-human primate has exhibited any of the richness of human language that distinguishes the adult human from the two-year-old, suggesting a biological difference in the brain's "language-readiness" between humans and other primates (in addition to the fact that the human brain and larynx together support voluntary control of vocal articulations which the bonobo's cannot – thus the use of signing or other manual-based symbols). (b) No bonobo or other non-human primate has been seen "in the wild" to use symbols in the way that Kanzi has learned to use them. It is Kanzi's exposure to human culture that lets him learn to use signing to communicate, but what he acquires is but a small fragment of the full richness of the sign language of an adult human signer.

language reflects, among many other things the results of an attempt to balance efficiency (getting a message across in as few words as possible) with avoidance of ambiguity *within the context of the current communication.*

Stages 3 and 4: Simple and Complex Imitation Systems for Grasping

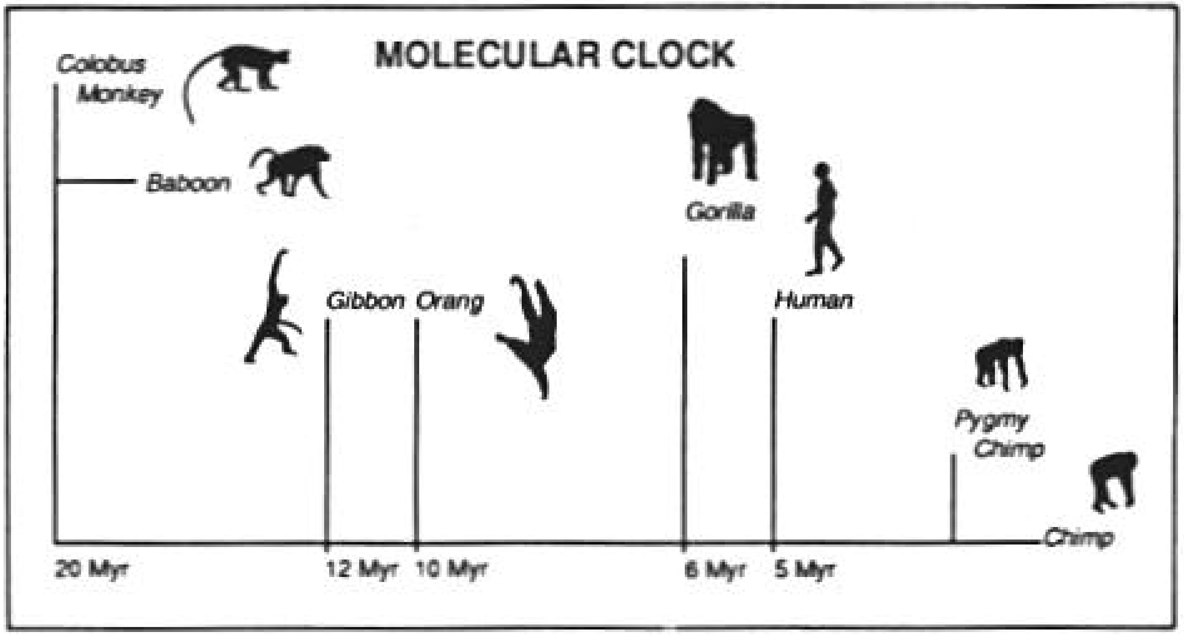

Figure 7. The timetable for hominid evolution inferred from the molecular clock. The crucial dates are 20 million years ago (divergence of monkeys from the line that led to humans and apes), and 5 million years ago (divergence of the hominid line from the line that led to modern apes). The pygmy chimp is also known as the bonobo, a quite different species from chimpanzee. (Adapted from Gamble, 1994, Figure 4.2.)

Figure 7 shows two key branch points in primate evolution. Twenty million years separate monkeys and humans from their common ancestor, while five million years separate chimps and humans from their common ancestor.

We remind the reader again that language played no role in the evolution of monkey or chimp or the common ancestors we share with them. Any changes we chart prior to the hominid line should be shown to be adaptive in their own right, rather than as precursors of language. Overall, the weight of evidence suggests that apes imitate far more than do monkeys. This leads to the following

Imitation Hypothesis: Extension of the mirror system from *recognizing* single actions to *being able to copy* compound actions was the key innovation in the brains of human, chimp and the common ancestor (as compared to the monkey-human common ancestor) relevant to language.

There is no hard and fast distinction between simple and compound actions. An apparently complex behavior may, with sufficient practice, become available as a unitary behavior from which yet more complex behaviors can be compounded. An action is compound relative to the current stock of unitary actions if it may be put together in some way (as a sequence, or some more subtle interleaving) from that stock. In this sense, we would hypothesize that the F5 mirror system would at any time bridge between the execution and observation of the unitary actions currently in the animal's repertoire but that it would require a superordinate system to perceive or execute the relationships between these units that define a compound action. In view of the earlier observations from the Oregon Regional Primate Research Center, it is clear that much more field work needs to be done to more formally characterize the capacities of monkey (and these may well vary from species to species of monkey) and chimpanzee.

In any case, we need top do much more to characterize the relevant monkey neurophysiology and relate it to brain imaging in monkey, chimpanzee and human.

It was noted in the Introduction that imitation generally requires the ability to break down a complex performance into a coordinated control program of pieces which approximate the pieces of the performance to be imitated. This then provides the framework in which attention can be shifted to specific components which can then be tuned and/or fractionated appropriately, or better coordinated with other components of the skill. This process is recursive, yielding both the mastery of ever finer details, and the increasing grace and accuracy of the overall performance. Given this framework, I argued that

i) monkey and, even more so, chimpanzee (and, presumably, the common ancestor of human and chimpanzees) has "simple imitation": they can imitate short novel sequences through repeated exposure, whereas

ii) the human has "complex imitation": s/he can acquire (longer) novel sequences in a single trial.

Important evidence for imitation in chimpanzee is that chimpanzees use and make tools. Different tool traditions are apparent in geographically isolated groups of chimpanzees: Different types of tools are used for termite fishing at the Gombe in Tanzania and at sites in Senegal. Boesch and Boesch (1983) have observed chimpanzees in Tai National Park, Ivory Coast, using stone tools to crack nuts open, although Goodall has never seen chimpanzees in the Gombe do this. The nut-cracking technique is not mastered until adulthood. Mothers overtly correct and instruct their infants from the time they first attempt to crack nuts, at age three years, and at least four years of practice are necessary before any benefits are obtained. To open soft-shelled nuts, chimps use thick sticks as hand hammers, with wood anvils. They crack harder-shelled nuts with stone hammers and stone anvils. The Tai chimpanzees live in a dense forest where suitable stones are hard to find. The stone anvils are stored in particular locations to which the chimpanzees continually return. Chimpanzees also use stones and other objects as projectiles with intent to do harm (Goodall, 1986).

Note that the form of imitation reported here for chimpanzees is a long and laborious process compared to the rapidity with which humans can acquire novel sequences. As long as the sequences are not too complex, their very structure can serve for humans as the basis for immediate imitation or for the immediate construction of an appropriate response, as well as contributing to the longer-term enrichment of experience. Of course (as our Santa Fe dance example shows), as sequences get longer, or the components become less familiar, more and more practice is required to fully comprehend or imitate the behavior.

Figure 6 focused on the generation and observation of a single hand action. We will need to extend this to handle the imitation of compound sequences in a way that meets criteria abstracted from the dance class example. Arbib (1981) showed how to describe perceptual structures and distributed motor control in terms of functional units called schemas which may be combined to form new schemas as coordinated control programs linking simpler (perceptual and motor) schemas. Jeannerod et al. (1995) provide a recent application of schema theory to the study of neural mechanisms of grasping. This raises points to be explicitly addressed in detailed modeling (not provided in this paper; see Bischoff and Arbib, 2000, for a non-adaptive model of non-mirror aspects of this):

We hypothesize that the plan of an action (whether observed or "intended") is encoded in the brain. We have to be a little subtle here. In some cases, a whole set of actions is overlearned and encoded in stable neural connectivity.

In other cases, the whole set of actions is planned in advance based on knowledge of the current situation. In yet other cases, dynamic planning is involved, with the plan being updated and extended as new observations become available. Consider, for example, how one's plan for driving to work may be modified both trivially – changing lanes to avoid slower cars, stopping for pedestrians – and drastically – as when changing traffic conditions force one to take a detour. We earlier spoke of generalizing a sequence to an automaton with a set X of inputs, a set Y of outputs, and a set Q of states, augmented by a state-transition function δ: Q x X → Q, and an output function β: Q → Y. This formalism is broad enough to encompass the above range from overlearned to dynamic plans, but it is still an open question as to how best to distribute the encoding of the various components of the automaton between stable synapses, rapidly changing synapses, and neural firing patterns.

In general, this "automaton" will be event-driven, rather than operating on a fixed clock – different sub-behaviors take different lengths of time, and may be terminated either because of an external stimulus, or by some internal encoding of completion. Neural activity may then encode the current state q as well as priming the code for $\delta(q,x)$ for a small set of "expected" events x. When one of these, say $x_1$occurs, the brain then brings $\delta(q,x_1)$ above threshold – thus releasing output $\beta[\delta(q,x_1)]$ which will be emitted for as long as the neural code for $\delta(q,x_1)$ is sufficiently active – and inhibits q and the other primed states, while priming a small set of candidate successor states. However, if the actual input when in state q is unexpected, say x_2, then $\delta(q,x_2)$ will be unprimed and thus the transition to the new state, and thus new output, will be delayed.

At a basic level, then, we might characterize imitation in terms of ability to "infer automata", recognizing the set of relevant outputs Y (the task of the mirror system) and overt transition signals X, and "inferring" a set of states Q and a set of "covert inputs" X' which allow one to mimic the observed behavior. However, a crucial observation of Arbib (1981) is that complex behaviors may be expressed as coordinated control programs, which are built up from assemblages of simpler schemas. In the corresponding formalism, we thus replace simple automaton inference by concurrent computation in a schema assemblage modeled as a network of port automata (Arbib, 1990; Steenstrup, Arbib and Manes, 1983). The task then becomes to recognize that portions of a novel behavior can be assimilated to existing schemas. Imitation involves, then, the ability to decompose behaviors into constituent schemas and then rapidly encode an assemblage of schemas which yields an approximation of the overall behavior. Further learning can then act both at the level of "assemblage code" (see Arbib, 1990), and at the level of parametric tuning of both the constituent schemas and of the linkages between them. For example, as noted earlier, if one is acquiring a tennis stroke and a badminton stroke through imitation, the initial coordinated control program may be identical, yet in the end the very different dynamics of the tennis ball and shuttlecock lead to divergent schemas.

Stage 5: A Manual-Based Communication System

We have now got to the stage in evolution in which a creature (the postulated common ancestor of human and chimpanzee) had the ability to imitate simple manual performances, and we have suggested that the early stages of hominid evolution yielded the perceptual and motor ability to imitate performances of increasing complexity. I now want to discuss how further hominid evolution might have yielded a manual-based communication system (Stage 5). I will suggest below how this might have led to the further evolution of brain and body mechanisms supporting speech and then will return to the idea that this evolutionary process did not yield, in the first *Homo sapiens*, creatures that already had language in the sense shared by modern human languages. Rather, I shall argue that early *Homo sapiens* was "language ready", and suggest the sort of cultural evolution that might have then led to the diversity of modern languages.

The story of hominid evolution is briefly summarized in Figure 8. Imprints in the cranial cavity of endocasts indicate that "speech areas" were already present in early hominids such as *Homo habilis* long before the larynx reached the modern “speech-optimal” configuration, but there is a debate over whether such areas were already present in australopithecines. This leads us to a related hypothesis: The transition from australopithecines to early *Homo* coincided with the transition from a mirror system used only for action recognition and imitation to a human-like mirror system used for intentional communication.

The function of mirror neurons has been advanced to be that their firing "represents" an action internally (Rizzolatti, et al., 1996a, Gallese et al. 1996, Jeannerod, 1994) as a basis for understanding actions. Here, understanding means the capacity that individuals have to recognize that another individual is performing an action, to differentiate the observed action from other actions, and to use this information in order to act appropriately. According to this view mirror neurons represent the link between sender and receiver that Liberman (1993; Liberman and Mattingly, 1985, 1989) postulated as the necessary prerequisite for any type of communication.

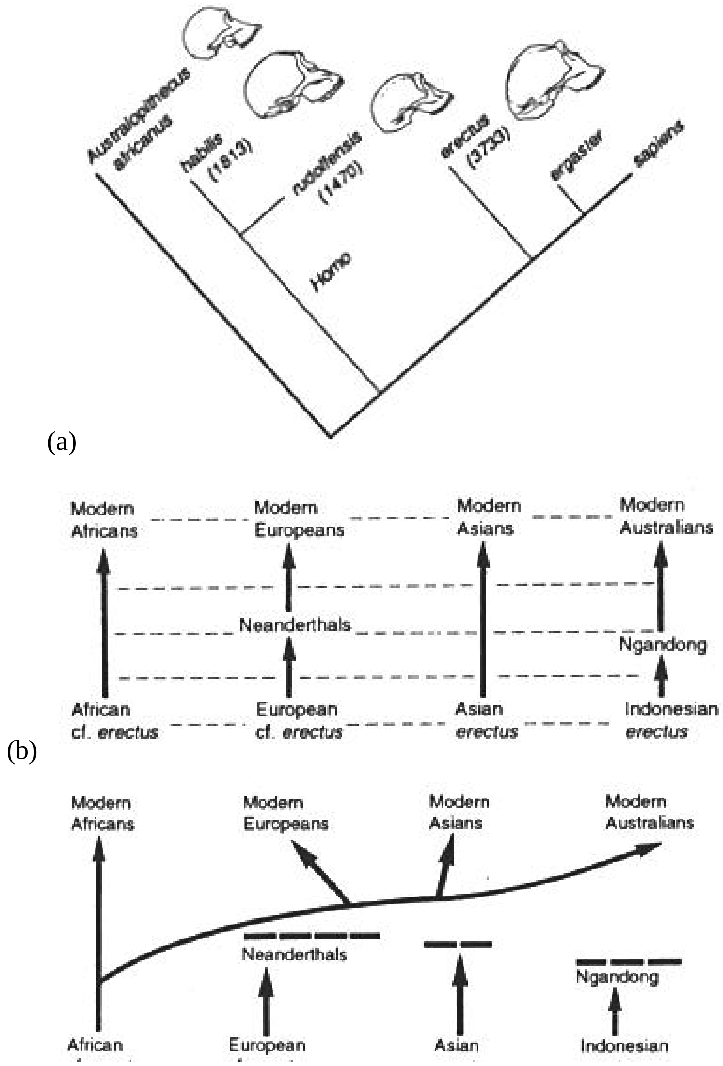

Figure 8. Five million years of hominid evolution. (Adapted from Gamble 1994.) (a) A family tree for hominids, with the common ancestor of humans, *Homo sapiens* (4), and australopithecines (1) preceding that for *Homo sapiens* and *Homo habilis* (2) which in turn precedes the common ancestor of *Homo sapiens* and *Homo erectus* (3). (b) It is widely agreed that *Homo erectus* evolved in Africa, and expanded from there to Europe and Asia. We reject the view (top panel) that erectus evolved independently into Homo sapiens in Africa, Europe and Asia, and instead adopt the view (bottom panel) that Homo sapiens evolved in Africa, and formed a second expansion out of Africa.

We agree, then, with Liberman's "motor theory of perception" – that the basic mechanism appears to be that of matching the neural activity resulting from observation of a gesture with that underlying its execution. However, there are cases of children learning to recognize spoken language without being able to produce it (Giuseppe Cossu, personal communication). In terms of Figure 6, we explain this by noting that the Expectation Neural Network (ENN) is tuned by corollary discharge from F5, and that this may still be available when F5 cannot control appropriate motor pattern generators. Nonetheless, the system can be tuned – for a child motivated enough to pay attention – by matching expectation of what will be said to what actually is said in an overheard conversation or in a classroom setting.

Our hypothetical sequence leading to manual gesture and beyond is then

i. pragmatic action directed towards a goal object;

ii. imitation of such actions;

iii. pantomime in which similar actions are produced in the absence of any goal object. In terms of observable behavior, imitation of an action and pantomime of an action may appear the same. However, imitation is the generic attempt to reproduce movements performed by another, whether to master a skill or simply as part of a social interaction. By contrast, pantomime is performed with the intention of getting the observer to think of a specific action or event. It is essentially communicative in its nature. The imitator observes; the pantomimic intends to be the observed;

iv. abstract gestures divorced from their pragmatic origins (if such existed): in pantomime it might be hard to distinguish a grasping movement signifying "grasping" from one meaning "a [graspable] raisin", thus providing an "incentive" for coming up with an arbitrary gesture to distinguish the two meanings. This suggests that the emergence of symbols occured when the communicative capacities of pantomiming were exhausted;

v. the use of such abstract gestures for the formation of compounds which can be paired with meanings in more or less arbitrary fashion.

My current hypothesis is that stages (iii) and (iv) and a rudimentary (pre-syntactic) form of (v) were present in pre-human hominids, but that the "explosive" development of (v) that we know as language depended on "cultural evolution" well after biological evolution had formed modern *Homo sapiens*. This remains speculative, and one should note that biological evolution may have continued to reshape the human genome for the brain even after the skeletal form of *Homo sapiens* was essentially stabilized, as it certainly has done for skin pigmentation and other physical characteristics. However, the fact that people can master any language equally well, irrespective of their genetic community, shows that these changes are not causal with respect to the structure of language.

Stage 6: Speech

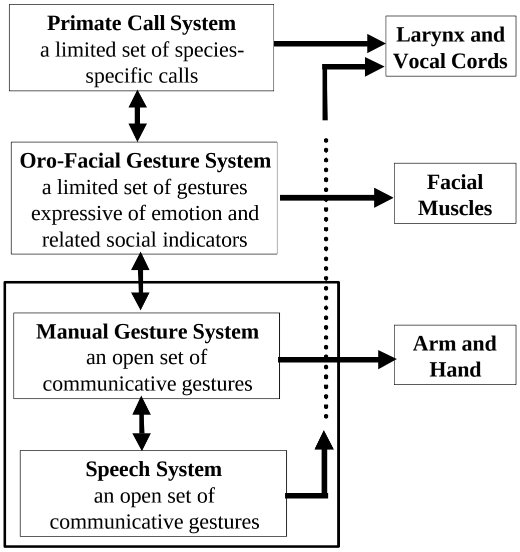
Primate Call System
a limited set of species-specific calls
Larynx and Vocal Cords
Oro-Facial Gesture System
a limited set of gestures expressive of emotion and related social indicators
Facial Muscles
Manual Gesture System
an open set of communicative gestures
Arm and Hand
Speech System
an open set of communicative gestures

Figure 9. A production view of the evolved speech system of early humans. (Perception systems are not shown.)

We earlier noted that the neural substrate for primate calls is in a region of cingulate cortex distinct from F5, which latter is the monkey homologue of human Broca's area. We thus need to explain why F5, rather than the a priori more likely "primate call area", provided the evolutionary substrate for speech and language. Rizzolatti and Arbib (1998) answer this by suggesting three evolutionary stages going beyond the capacities of Figure 5:

1. A *distinct* manuo-brachial (hand-arm) communication system evolved to complement the primate calls/oro-facial communication system.

2. The "speech" area of early hominids (i.e., the area presumably homologous to monkey F5 and human Broca’s area) mediated orofacial and manuo-brachial communication but not speech.

3. The manual-orofacial symbolic system then “recruited” vocalization. Association of vocalization with manual gestures allowed them to assume a more open referential character, and exploit the capacity for imitation of the underlying brachio-manual system.

Thus, we answer the question "Why did F5, rather than the primate call area provide the evolutionary substrate for speech?" by saying that the primate call area could not of itself access the combinatorial properties inherent in the manuo-brachial system.

The claim, then, is that the biological evolution of hominids yielded a mirror system embedded in a far larger system for execution, observation and imitation of compound behaviors composed from oro-facial, manual, and vocal gestures. I also accept that this system supported communication in *Homo erectus* – since otherwise it is hard to see what selective pressure could have brought about the lowering of the larynx which, as Lieberman (1991) observes, makes humans able to articulate more precisely than other primates, but at the precise of an increased likelihood of choking.[9] Clearly, some level of language-readiness and vocal communication preceded this – a core of proto-speech was needed to provide pressures for larynx evolution. However, I do not accept that this means that the earliest *H. sapiens* was endowed with language in anything like its modern human richness. Rather, biological evolution equipped early humans with "language-ready brains" which proved rich enough to support the cultural evolution of human languages in all their commonalities and diversities.[10]

I have schematized the result of the above three evolutionary stages in Figure 9. A key question for later analysis is whether we should consider the manual gesture system as a primitive system atop which evolved the "advanced" speech system, or whether we should view these two as actually different aspects of one multi-modal controller, depending upon which efferent system we focus. (Note that we are here talking of speech as being the open-ended production and perception of sequences of vocal gestures, without implying that these sequences constitute a language.)

Perception systems are not shown in the figure. The mirror system is thus implicit. Extending the Mirror System Hypothesis, we must show how the ability to comprehend and create utterances via their underlying syntactico-semantic hierarchical structure can build upon the observation/execution of single actions. Here I stress, as Rizzolatti and Arbib did not, that the transition to language readiness, with the necessary openness to the creation of compound

[9] The Perth psychologist Colin McLeod [personal communication] quips that "The human vocal tract evolved so that we could cry out 'Help, I'm choking!'")

[10] The reader should be warned of the dangerous methodology that I employ here. My work is anchored in a rigorous knowledge of modern neuroscience, including detailed modeling of many neural systems, and informed by a reasonable knowledge of psychology, linguistics and evolutionary theory. On the other hand, my knowledge of anthropology and human evolution is limited, compounding the severe limitations on the data base on the evolution of language – after all, we have no record of writing that is more than 6,000 years old. Thus, when I assert that *H. sapiens* was not originally endowed with language in anything like its modern human richness, I am not appealing to hard data, but rather forwarding a hypothesis based on, but in no sense implied by, a variety of evidence. However, I do not (nor should the reader) accept my hypotheses uncritically. Rather, each new hypothesis is confronted with new data and competing hypotheses as my reading progresses. The hypotheses presented in this article have thus survived a great deal of "cross-examination", and have been refined in the process. For example, while my reading of historical linguistics impressed me with the rapidity with which languages change (and I view human language as the sum of human languages, not as some abstract entity above and beyond these bio-cultural products) and thus led me to distinguish the notion of a "language-ready" brain from a "language-equipped" brain, further reading and reflection leads me to accept that the dichotomy here is not as sharp as I may have believed earlier.

expressions, required imitation in the sense defined earlier: not just simply observing someone else's movement and responding with a movement which in its entirety is already in one's own repertoire, but rather "parsing" a complex movement into more or less familiar pieces, and then performing the corresponding composite of (variations on) familiar actions.

Having shown why speech did not evolve “simply” by extending the classic primate vocalization system, we must note that the language and vocalization systems are nonetheless linked. Lesions centered in the anterior cingulate cortex and supplementary motor areas of the brain can also cause mutism in humans, similar to the effects produced in muting monkey vocalizations. Conversely, a patient with a Broca's area lesion may nonetheless swear when provoked. But note that "emitting an imprecation" is more like a monkey vocalization than like the syntactically structured use of language. Lieberman (1991) suggests that the primate call made by an infant separated from its mother not only survives in the human infant, but in humans develops into the breath group, i.e., the pattern of breathing in and breathing out that is shaped to provide the contour for each continuous sequence of an utterance. I thus hypothesize that the evolution of speech yielded the pathways for cooperative computation between cingulate cortex and Broca’s area, with cingulate cortex involved in breath groups and emotional shading, and Broca’s area providing the motor control for rapid production and interweaving of elements of an utterance.

Rizzolatti and Arbib (1998) state that “This new use of vocalization necessitated its skillful control, a requirement that could not be fulfilled by the ancient emotional vocalization centers. This new situation was most likely the ‘cause’ of the emergence of human Broca’s area.” I would now rather say that *Homo habilis* and even more so *Homo erectus* had a “proto-Broca’s area” based on an F5-like precursor mediating communication by manual and oro-facial gesture. This made possible a process of collateralization whereby this “proto” Broca’s area gained primitive control of the vocal machinery, thus yielding increased skill and openness in vocalization, moving from the fixed repertoire of primate vocalizations to the unlimited (open) range of vocalizations exploited in speech. Larynx and brain regions could then co-evolve to yield the configuration seen in modern *Homo sapiens*.

Noting that the specific communication system based on primate calling was not the precursor of language, some people (e.g., Bickerton 1995) have claimed that communication could not have been a causal factor in the evolution of language-readiness. They then argue that it was the advantage of being able to represent a complex world that favored language evolution. However, we should not be constrained by such either/or thinking. Rather, the co-evolution of communication and representation was essential for the emergence of human language. Both representation within the individual and communication between individuals could provide selection pressures for the biological evolution of language-readiness and the further cultural evolution of language, with advances in the one triggering advances in the other.

The Transition to *Homo sapiens*

If we look at Figure 1, we can see that the cerebral cortex of monkey can be divided into many different regions. Moreover, the extent of these regions can vary drastically in different primate species. Different species may not only have nuclei (groups of neurons) and cortical regions of different sizes, but may actually have nuclei lacking in other species, or may nuclei present in other species specialized for quite different functions. How can this be? Butler and Hodos (1996) show that the course of brain evolution among vertebrates has been determined by

i) Formation of multiple new nuclei through elaboration or duplication of nuclei present in the ancestral species;

ii) Regionally specific increases in cell proliferation in different parts of the brain can vary the relative size of a given region across species; and

iii) Gain of some new connections and loss of some established connections may change the inputs and outputs of a region, thus changing its computations and their impact on brain and behavior.

These phenomena (and others described by Butler and Hodos) can be influenced by relatively simple mutational events that can thus become established in a population as the result of random variation. Selective pressures determine whether the behavioral phenotypic expressions of central nervous system organization produced by these random mutations increase their proportional representation within the population and eventually become established as the normal condition.

Let me list here a few contrasts between monkey and human brain. There is no space here to go into details. Rather, I want to offer a corrective to any false impression I might have created that that "turning F5 into Broca's area" is the one key neural correlate of human evolution. The differences between monkey and human not only include many bodily changes (including upright gait, a lowered larynx, and increased manual dexterity) but also changes in many neural systems related to new motor and cognitive abilities as well as dramatic extensions of "cognitive reach". Among the changes to be noted are:

(a) Enlargement of the pre-frontal lobe (which uses motivation to evaluate future courses of action) to provide sophisticated memory structures (coupled, e.g., to hippocampus with its key role in episodic memory) to extend "cognitive comprehension" in space and time

(b) Extension of the number, sophistication and coordination of parietal-frontal perceptuo-motor systems

(c) Enlargement of the POT (Parieto-Occipito-Temporal cortex) as a semantic storehouse

(d) Adding prefrontal circuitry with refinements of the basal ganglia and cerebellum keeping pace

(e) An increased ratio of pre-motor cortex to motor cortex.

How did evolution couple the separate parietal↔frontal subsystems into an “integrated state of knowledge”? Fuster (1995) sees prefrontal cortex as evolving to increase working memory capacity. Petrides (1985) argues that we need prefrontal cortex to go beyond single items to keeping multiple objects or events in order. We note again the challenge of embedding the mirror system in a system for handling sequential structure, and hierarchical structure more generally. The detailed consideration of these brain regions and the attendant claims is beyond the scope of this article. Here we note that, irrespective of whether or not they possessed language, all primates possess the ability for visual (and multi-modal) scene perception – perceiving the changing spatial relationships between a number of objects and organisms in the environment, and attending to the key events relevant for ongoing behavior. Events – not objects – are primary in our story, keeping action at the center.

A Multi-Modal System

Our use of writing as a record of spoken language has long since created the mistaken impression that language is a speech-based system. However, McNeill (1992) has used videotape analysis to show the crucial use that people make of gestures synchronized with speech. Even blind people use manual gestures when speaking. As deaf people have always known, but linguists only much later discovered (Klima and Bellugi 1979), sign languages are full human languages, rich in lexicon, syntax, and semantics. Moreover, not only deaf people use sign language. So do some aboriginal Australian tribes, and some native populations in North America. Thus language is more than “that part of speech which can be captured in writing”.

Thus, where Rizzolatti and Arbib (1998) state that “Manual gestures progressively lost their dominance, while in contrast, vocalization acquired autonomy, until the relation between gestural and vocal communication inverted and speech took off”, I would now downplay the autonomy: The above considerations suggest that we locate language in a speech-manual-orofacial gesture complex and that during language acquisition a normal person shifts the major information load of language – but by no means all of it – into the speech domain, whereas for a deaf person the major information load is removed from speech and taken over by hand and orofacial gestures. On this basis, I answer the question following Figure 9 by saying that "The vote is in: one box replaces three", as shown in Figure 10. A warning: Although we claim that there is but one communication system we stress that it involves many brain regions, each with its own evolutionary story. Neither Figure 9 nor Figure 10 shows the neuroanatomy of these mechanisms.

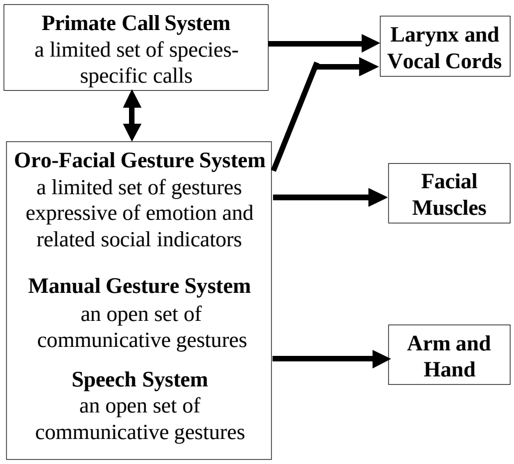
Primate Call System
a limited set of species-specific calls
Larynx and Vocal Cords
Oro-Facial Gesture System
a limited set of gestures expressive of emotion and related social indicators
Facial Muscles
Manual Gesture System
an open set of communicative gestures
Speech System
an open set of communicative gestures
Arm and Hand

Figure 10. The fruit of evolution: Not three separate communication systems, but a single system operating in at least three motor modalities and at least two sensory modalities.

Language Evolving

From Action-Object Frame to Verb-Argument Structure to Syntax and Semantics

The divergence of the Romance languages took about one thousand years. The divergence of the Indo-European languages to form the immense diversity of Hindi, German, Italian, English, etc., took about 6,000 years. How can we imagine what has changed since the emergence of *Homo sapiens* some 200,000 years ago? Or in 5,000,000 years of prior hominid evolution?

I claim that the first humans were language-ready but did not have language in the modern sense. We have already seen that some form of communication, first manual then an integration of oral, manual and facial communication, must have arisen in hominid evolution. What, then, might be the nature of such communication if it does not yet constitute the use of language as humans know it today? A basic component of modern human languages is the verb-argument structure, expressing an "action-object frame" like look(Mary, John) with a sentence such as "Mary looks at John", with "look" the verb, and "Mary" and "John" the arguments. My suggestion is that the first *Homo sapiens* would have been able to perceive that one individual was looking at another, and might even have had the ability to form a distinctive vocalization to draw this event to the attention of another tribe member, but that the vocalization used would be arbitrary, rather than being decomposable into something which would be understood as part of a general class of "something like verbs" and two somethings which would be understood as part of a general class of "something like nouns". In other words, I hypothesize that (i) the ability for visual scene perception that must underlie the present human ability to employ verb-argument structures – the perception of action-object frames in which an actor, an action, and related role players can be perceived in relationship – was well established in the primate line; (ii) the ability to communicate a fair number of such frames was established in the hominid line prior to the emergence of *Homo sapiens*; but that (iii) these "communicative signals" lacked (almost all) syntactic structure and the extraction of semantic structure therefrom.

Our starting point for the biological basis of language-readiness is then that "it is innate to know there are things and events”. Again, we recognize an evolutionary progression:

1. Acting on objects

2. Recognizing acting on objects: an "action-object frame". Here we extend the mirror system concept to include recognition not only of the action (mediated by F5) but also of the object (mediated by IT [inferotemporal cortex]; see Figure 3). This reflects a crucial understanding gained from Figure 3 that is often missing in the study of the mirror system: The canonical activity of F5 already exhibits a congruence between the affordances of an object (mediated by the dorsal stream) and the nature of the object (as recognized by IT and elaborated upon [in a process of "action-oriented perception"] in prefrontal cortex, PFC). In the same way, the activity of mirror neurons does not rest solely upon the parietal recognition (in an area called PF not shown in Figure 3) of the hand motion and the object's affordances (AIP) but also (I here postulate) on the "semantics" of the object as extracted by IT and relayed to F5 via PFC.[11] It is this matching of actions with "object semantics" and "goals", rather than just with affordances, that makes possible the transition to:

3. Creating symbols linked to specific action-object frames. Here I must reiterate the subtle point that the original communicative signal for an action-object frame need not have involved separate lexical entries for the action or the objects. Thus *griffle* might mean "grasp a peanut with a precision pinch", while *tromfok* means "grasp a daisy stem with a precision pinch". Nothing said so far demands a lexical decomposition of these structures. However, the ability to symbolize more and more situations required the creation of a “symbol toolkit” of meaningless elements from which an open ended class of symbols could be generated. The distinction I have in mind here relates to the earlier point that in pantomime it might be hard to distinguish a grasping movement signifying “grasping” from one meaning “a [graspable] raisin”, thus providing an “incentive” for coming up with an arbitrary gesture to distinguish the two meanings. However, it can also be argued that the passage from pantomime did not occur originally in the brachio-manual system, but occurred as speech evolved "atop" manual gesture with the two systems evolving into

[11] This suggests the hypothesis that inactivation of IT which does not disturb AIP will disrupt mirror activity in F5, but not canonical activity in F5.

one integrated system for communication: In this scenario, the ability to create novel sounds to imitate manual gestures in the vocal domain (for example, rising pitch might represent an upward movement of the hand), coupled with a co-evolved ability to imitate novel sound patterns yielding vocal gestures through onomatopoeia that were not linked to manual gestures, created the divorce of gesture from meaning required to create an open-ended vocabulary.

In any case, it is a distinct step in language evolution to proceed from Stage 3 to the ability, e.g., to invoke the same word(s) for " grasp with a precision pinch" irrespective of the specific type of object which is grasped:

4. Naming objects and actions. This involves the creation of symbols (whether vocal or manual) which allow the communicative signal for an action-object frame to become an explicit compound, a verb-argument structure in which a specific action can be denoted by (some variant of) a specific verb no matter what the context, and specific arguments are represented by specific nouns. (Of course, the nouns may themselves be single lexical items, or themselves phrases, but this takes us beyond the present basic evolutionary scenario.)

Stage 4 is then the crucial step in the transition from communication in general to human language as we know it: abstract symbols are grounded (but more and more indirectly) in action-oriented perception; members of a community may acquire the use of these new symbols (the crucial distinction here is with the fixed repertoire of primate calls) by imitating their use by others; and, crucially, these symbols can be compounded in novel combinations to communicate about novel situations for which no agreed-upon unitary communicative symbol exists.

In our example, the ability to differentially signal “grasping” from “a [graspable] raisin” could then have laid the basis for replacing a "unitary symbol" for a verb-argument structure into a compound of two components of what we would now recognize as precursors of a verb and a noun. This could have, in turn, formed the basis for the abstraction and compounding of more generic verb-argument structures. Again, much of this would at first have been based on a limited yet useful set of templates, variations on a few basic themes. It might have taken many, many millennia for people to discover syntax and semantics in the sense of gaining immense expressive power by “going recursive” with a relatively limited set of strategies for compounding and marking utterances, based on a vocabulary that expanded with the kinship structures and technologies of the different tribes, these cultural products themselves enhanced by the increased effectiveness of transmission from generation to generation that the growing power of language made possible. Consistent with the earlier distinction between "language-readiness" and "having language", we need more research to determine what it is about a brain that makes it possible to "go recursive". The hypothesis is that whatever is involved is completely distinct from the encoding of specific recursive rules of syntax within a Universal Grammar.

The cultural evolution of *Homo sapiens* may then have involved an increased ability to name actions and objects to create an rapidly growing set of verb-argument structures, and the ability to compound those structures in diverse ways. I would suggest that many ways of expressing these relationships were the discovery of *Homo sapiens*. That is, many grammatical structures like adjectives, conjunctions such as *but*, *and*, or *or* and *that*, *unless*, or *because* might well have been “post-biological” in their origin. The spread of these innovations rested on the ability of other humans not only to imitate the new actions and compounds of actions demonstrated by the innovators, but also to do so in a way which related increasingly general classes of symbolic behavior to the classes, events, behaviors and relationships that they were to represent. Indeed, consideration of the spatial basis for “prepositions” may help show

how visuomotor coordination underlies some aspects of language, while the immense variation in the use of corresponding prepositions even in closely related languages like English and Spanish shows how the basic functionally grounded semantic-syntactic correspondences have been overlaid by a multitude of later innovations and borrowings. We still have much work to do to understand how the needs of human biology and the constraints of the human brain shaped these basic discoveries, their dissemination, and their stabilization.

4 Conclusion

With this, I conclude my argument that the ability to imitate is a key innovation in the evolutionary path leading to language in the human. However, much has to be done to complete the linkage of this hypothesis to specific data on brain mechanisms. We have taken this argument through seven hypothesized stages of evolution,

1. grasping
2. a mirror system for grasping (i.e., a system that matches observation and execution),
3. a simple imitation system for grasping,
4. a complex imitation system for grasping,
5. a manual-based communication system
6. speech, and
7. language.

Certainly, Stages 1 and 2 are firmly grounded in neurophysiological data for monkey, these data have been modeled in detail for Stage 1 and are now being analyzed by careful modeling of Stage 2. At the other end, Stage 7, we have neurological data on localization of mechanisms related to both spoken and sign language in the human brain. The data on the intermediate stages are still limited. Neurophysiology of the chimpanzee is no longer permitted, but we can hope that advances in the refinement of techniques for noninvasive brain imaging, such as the PET and fMRI studies of humans conducted at present, will help us fill the gap as we seek to link comparative brain imaging studies of a variety of primates with detailed neurophysiological and neuroanatomical studies where available. Computational models of the biological circuitry underlying both simple and complex imitation will play a crucial role, as will the use of mathematical techniques (such as the Synthetic PET technique of Arbib et al., 1994) which link models of circuitry to predictions of relative regional activation in brain imaging studies. However, there is a paucity of satisfactory neurolinguistic modeling which sheds light on the Stage 6 imaging data by tracing the cooperative computation of these regions during language performance.

It is my hope that "Mirror Neurolinguistics" will build on the hypotheses presented above to respond to the challenge of constructing a computational neurolinguistics issued more than 20 years ago (Arbib and Caplan, 1979). The rest of this section offers an early response to this challenge, but the reader will easily see how much more needs to be done. Our evolutionary theory suggests a progression from action to pantomime to (proto)language

1. object → AIP → $F5_{canonical}$: pragmatics
2. action → PF → $F5_{mirror}$: action understanding
3. scene → Wernicke’s → Broca’s: utterance

Goodale, Milner, Jakobson, & Carey (1991) studied a patient (DF) who developed a profound visual form agnosia following carbon monoxide poisoning in which most of the damage to cortical visual areas was apparent in areas 18 and 19, but not area 17 (V1) - still allowing signals to flow from V1 towards PP but not from V1 to IT. When asked to indicate the width of a single block by means of her index finger and thumb, her finger separation bore no relationship to the dimensions of the object and showed considerable trial to trial variability. Yet when she was asked simply to reach out and pick up the block, the peak aperture (well before contact with the object) between her index finger and thumb changed systematically with the width of the object, as in normal controls. A similar dissociation was seen in her responses to the orientation of stimuli. In other words, DF could preshape accurately, even though she appeared to have no conscious appreciation (expressible either verbally or in pantomime) of the visual parameters that guided the preshape. Castiello et al. (1991) report a study of impairment of grasping in a patient (AT) with a lesion of the visual pathway that left PP, IT, and the pathway V→IT relatively intact, but grossly impaired the pathway V→PP. This patient is the "opposite" of DF - she can use her hand to pantomime the size of a cylinder, but cannot preshape appropriately when asked to grasp it. Instead of an adaptive preshape, she will open her hand to its fullest, and only begin to close her hand when the cylinder hits the "web" between index finger and thumb. But there was a surprise! When the stimulus used for the grasp was not a cylinder (for which the "semantics" contains no information about expected size), but rather a familiar object - such as a reel of thread, or a lipstick - for which the "usual" size is part of the subject's knowledge, AT showed a relatively adaptive preshape.

The "zero order” model of AT and DF data is:

4. Parietal “affordances” → preshape

5. IT “perception of object” → pantomime or verbally describe size

which leads to the inference that one cannot pantomime or verbalize an affordance; but rather one needs a "unified view of the object" (IT) to which attributes can be attributed before one can express them. The problem with this is that the “language” path as shown in (5) is completely independent of the parietal → F5 system, and so the data seem to contradict our view in (3).

To resolve this apparent paradox, we must return to the view of the FARS model given in Figure 2, and stress the crucial role of IT and PFC in modulating F5’s selection of an affordance, leading us to include paths from prefrontal cortex to F5 (canonical and mirror) and Broca's area in Figure 11. This figure provides a first speculative attempt to extend the FARS model conceptually to include not only the mirror system for grasping but also the language system evolved "atop" this. The crucial point is that all three paths defined above:

1. object → AIP → $F5_{canonical}$: pragmatics

2. action → PF → $F5_{mirror}$: action understanding

3. scene → Wernicke’s → Broca’s: utterance

are now enriched by the prefrontal system for "scene perception" which combines current IT-input with memory structures combining objects, actions, and relationships. The “lightning bolts” link "grasp boxes" to "language boxes" and are completely speculative.

Much more must be done to take us up the hierarchy from elementary actions to the recognition and generation of novel compounds of such actions. Nonetheless, the above preliminary account strengthens the case that no Universal Grammar need have been encoded in the brain of the first *Homo sapiens*. Rather it was the imitation-enriched mirror system that enabled human societies, across many millennia of invention and cultural evolution, to achieve human languages in the modern sense.

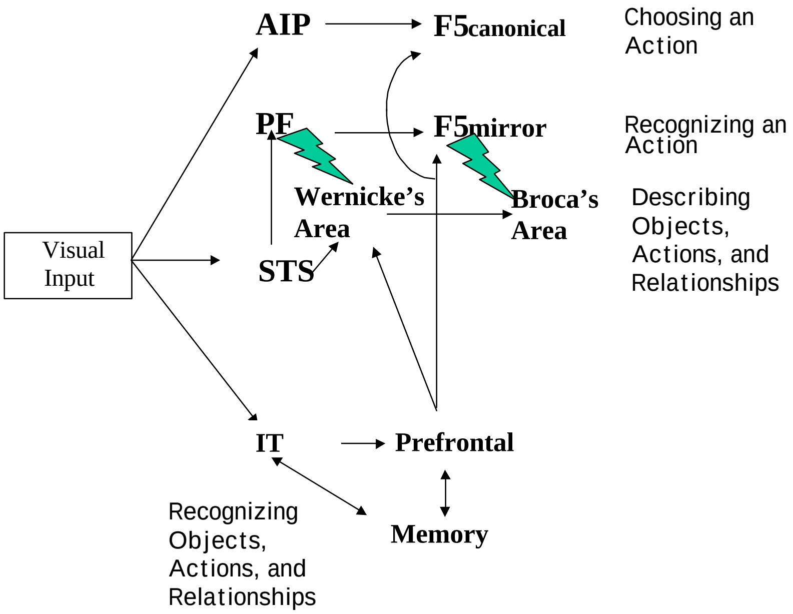

Figure 11. Extending the FARS model to include thee mirror system for grasping and the language system evolved "atop" this. Note that this simple figure neither asserts nor denies that the extended mirror system for grasping and the language-supporting system are anatomically separable, nor does it address issues of lateralization.

Acknowledgements: I recall with pleasure my sabbatical visit to the University of Western Australia in February, March and April of 1999 and the conversations with Robyn Owens and other colleagues there which laid the basis for the present article, as well as the many discussions with Giacomo Rizzolatti and Giuseppe Cossu which helped shape the article's further development while I continued my sabbatical in Rizzolatti's Institute in Parma, Italy, in May, June and July of 1999. I thank Aude Billard and the editors for their constructive comments on recent drafts.

References

M. A. Arbib. Memory Limitations of Stimulus-Response Models. *Psychological Review*, 76:507-510, 1969.

M. A. Arbib. Perceptual Structures and Distributed Motor Control. In V. B. Brooks, editor, Handbook of Physiology, Section 2: The Nervous System, Vol. II, Motor Control, Part 1), American Physiological Society, 1449-1480, 1981.

M. A. Arbib. Programs, Schemas, and Neural Networks for Control of Hand Movements: Beyond the RS Framework. In M. Jeannerod, editor, Attention and Performance XIII. Motor Representation and Control. Lawrence Erlbaum Associates, pp.111-138, 1990.

M. A. Arbib. Co-Evolution of Human Consciousness and Language, Annals of the New York Academy of Sciences (in press), 2000.

M. A. Arbib, P. Érdi, & J. Szentágothai. Neural Organization: Structure, Function and Dynamics. The MIT Press, Cambridge, Massachusetts, 1998.

M. A. Arbib and G. Rizzolatti. Neural expectations: a possible evolutionary path from manual skills to language. *Communication and Cognition*, 29:393-424, 1997.

D. Armstrong, W. Stokoe, & Wilcox, S. Gesture and the Nature of Language. Cambridge University Press, Cambridge, Massachusetts, 1995.

D. Bickerton. Language and Human Behavior. University of Washington Press, Seattle, Washington, 1995.

A. Bischoff-Grethe and M.A. Arbib, Sequential Movements: A Computational Model of the Roles of the Basal Ganglia and Supplementary Motor Area (to appear), 2000.

C. Boesch and H. Boesch, Optimization of nut-cracking with natural hammers by wild chimpanzees, *Behavior*, 83:265-286, 1983.

J. L. Borges. Of exactitude in science. In *A Universal History of Infamy*, p. 131. Penguin Books, 1975.

A. B. Butler & W. Hodos. Comparative Vertebrate Neuroanatomy: Evolution and Adaptation. John Wiley & Sons, New York, 1996.

U. Castiello, Y. Paulignan, & M. Jeannerod. Temporal dissociation of motor responses and subjective awareness: A study in normal subjects. *Brain*, 114:2639-2655, 1991.

N. Chomsky. Reflections on Language. Pantheon, New York, 1975.

M. C. Corballis. The lopsided ape: Evolution of the generative mind. Oxford University Press, New York, 1991.

M. C. Corballis. On the evolution of language and generativity. *Cognition*, 44:197-226, 1992.

T.W. Deacon. *The Symbolic Species: The co-evolution of language and the brain*, W.W. Norton & Company, New York & London, 1997.

G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, & G. Rizzolatti. Understanding motor events: a neurophysiological study. *Experimental Brain Research,* 91:176-180, 1992.

P.F. Dominey, M.A. Arbib, & J.-P. Joseph, A Model of Corticostriatal Plasticity for Learning Associations and Sequences, *J. Cog. Neurosci*., 7:311-336, 1995.

L. Fadiga, L. Fogassi, G. Pavesi, & G. Rizzolatti. Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology*, 73:2608-2611, 1995.

A. H. Fagg, & M. A. Arbib. Modeling Parietal-Premotor Interactions in Primate Control of Grasping. *Neural Networks*, 11:1277-1303, 1998.

J. M. Fuster. Memory in the cerebral cortex: An empirical approach to neural networks in the human and nonhuman primate. The MIT Press, Cambridge, Massachusetts, 1995.

V. Gallese, L. Fadiga, L. Fogassi, & G. Rizzolatti. Action recognition in the premotor cortex. *Brain*, 119:593-609, 1996.

C. Gamble. *Timewalkers: The Prehistory of Global Colonization*. Harvard University Press, Cambridge, Massachusetts, 1994.

J.J. Gibson, *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979.

M. A. Goodale, A. D. Milner, L. S. Jakobson & D. P. Carey. A neurological dissociation between perceiving objects and grasping them. *Nature*, 349:154-156, 1991.

J. Goodall, *The Chimpanzees of Gombe: Patterns of Behavior.* Cambridge, MA: Harvard University Press, 1986.

S. T. Grafton, M. A. Arbib, L. Fadiga, & G. Rizzolatti. Localization of grasp representations in humans by

PET: 2. Observation compared with imagination. *Experimental Brain Research,* 112:103-111, 1996b.

G. Hewes. Primate communication and the gestural origin of language. *Current Anthropology*, 14:5-24, 1973.

M. Jeannerod. The representing brain: neural correlates of motor intention and imagery. *Behav. Brain Sci.*, 17:187-245, 1994.

M. Jeannerod, M. A. Arbib, G. Rizzolatti & H. Sakata. Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences,* 18:314-320, 1995.

D. Kimura. Neuromotor Mechanisms in Human Communication (Oxford Psychology Series No. 20). Oxford University Press/Clarendon Press, Oxford, New York, 1993.

E.S. Klima and U. Bellugi. *The Signs of Language*, Cambridge, MA: Harvard University Press, 1979.

K. S. Lashley. The problem of serial order in behavior. In L. A. Jeffress, editor. Cerebral mechanisms in behavior: The Hixon symposium. Wiley, l951.

A. M. Liberman. Haskins Laboratories Status Report on Speech Research. 113:1-32, 1993.

A. M. Liberman & I. G. Mattingly. The motor theory of speech perception revised. *Cognition*, 21:1-36, l985.

A. M. Liberman & I. G. Mattingly. A Specialization for Speech Perception. *Science*, 243:489-494, 1989.

P. Lieberman. Uniquely Human: The Evolution of Speech, Thought, and Selfless Behavior. Harvard University Press, Cambridge, Massachusetts, 1991.

D. McNeill. *Hand and Mind: What Gestures Reveal about Thought.* Chicago: The University of Chicago Press, 1992.

E. Oztop & M. A. Arbib. Action Recognition in Primates: A Model of the Mirror Neuron System (in preparation), 2000.

M. Petrides. Deficits in conditional associative-learning tasks after frontal and temporal lesions in man. *Neuropsychologia,* 23:601-614, 1985.

G. Rizzolatti & M. A. Arbib. Language Within Our Grasp. *Trends in Neurosciences*, 21(5):188-194, 1998.

G. Rizzolatti, R. Camarda, L. Fogassi, M. Gentilucci, G. Luppino, & M. Matelli. Functional organization of inferior area 6 in the macaque monkey II. Area F5 and the control of distal movements. *Experimental Brain Research*, 71:491-507, l988.

G. Rizzolatti, L. Fadiga, V. Gallese, & L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131-141, 1996a.

G. Rizzolatti, L. Fadiga, M. Matelli, V. Bettinardi, D. Perani, & F. Fazio. Localization of grasp representations in humans by positron emission tomography: 1. Observation versus execution. *Experimental Brain Research*, 111:246-252, 1996b.

G. Rizzolatti, G. Luppino, & M. Matelli. The organization of the cortical motor system: new concepts. *Electroencephalography and Clinical Neurophysiology*, 106:283-296, 1998.

S. Savage-Rumbaugh, S. G. Shanker, T. T. Taylor. Apes, Language, and the Human Mind. Oxford University Press, New York, 1998.

M. Steenstrup, M. A. Arbib, & E. G. Manes. Port Automata and the Algebra of Concurrent Processes. *Journal of Computer and System Sciences*, 27:29-50, 1983.

M. Taira, S. Mine, A. P. Georgopoulos, A. Murata, & H. Sakata. Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Experimental Brain Research*, 83:29-36, 1990.