Arbib, M.A., 2000, 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

DRAFT: December 10, 1999

of a Chapter to appear in ***Imitation in Animals and Artifacts***,

Chrystopher Nehaniv and Kerstin Dautenhahn, Editors, to be published by MIT Press

(c.l.nehaniv@herts.ac.uk, kerstin@ai.mit.edu)

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/

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 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.

Introduction

The starting point for the present approach to language evolution is provided by the paper "Language Within Our Grasp" (Rizzolatti and Arbib, 1998). Briefly, the thesis of that paper is that the mirror system in monkey is the homolog of Broca's area in humans, and that this observation provides a neurobiological "missing link" for the long-argued hypothesis that sign language (based on manual

gesture) preceded speech in the evolution of language. The next section summarizes the basic evidence for this "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. In particular, I will argue that the ability to imitate is one of the evolutionary stages that marks the evolutionary path from mirror neurons in the common ancestor of monkey and human to language in the human. The paper will take us through five hypothesized stages of evolution,

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

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 above five 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. (c) We will not restrict language to "that which is expressed in speech, or in writing derived therefrom."

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 offering a detailed model, the FARS model, for grasping, and 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 a mirror system for grasping in humans has been found 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, ands 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 takes us "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. We first make a vital distinction between the view that the basic structures of language are encoded in the brain, as in the Universal Grammar of Chomsky, and the view – which I espouse – 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 possess 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 Stage 3, An Imitation System for Grasping, suggesting that this evolved in the 17 million years between the common ancestor of monkey and human and the common ancestor of chimpanzee and human. With this, we move to a speculative scenario for how Stage 4, A Manual-Based Communication System, broke through the fixed repertoire of primate vocalizations to yield a combinatorially open repertoire, so that Stage 5, 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. I speculate on the transition from action-object frame to verb-argument structure 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.

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

Stage 1: Grasping

The task of the present section is to take us through Stage 1 of our five hypothesized stages of evolution, grasping ("before the mirror"), reviewing relevant data, and presenting useful grounding concepts provided by the FARS model. The neurophysiological findings of the Sakata group on parietal cortex and the Rizzolatti group on premotor cortex indicate that parietal area AIP (the Anterior Intra-Parietal sulcus) and ventral premotor area F5 in monkey 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 (Taira et al., 1990; Rizzolatti et al., 1988; see Jeannerod et al., 1995 for a review). The FARS (Fagg-Arbib-Rizzolatti-Sakata) model (Fagg and Arbib 1998) 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 1 and 2.

Our basic view is that AIP computes "affordances" for grasping from the visual stream and sends (neural codes for) these on to area F5 – *affordances* 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. 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. In what follows, we briefly outline the FARS model, for it makes clear certain conceptual issues that will be crucial at later stages of the argument.

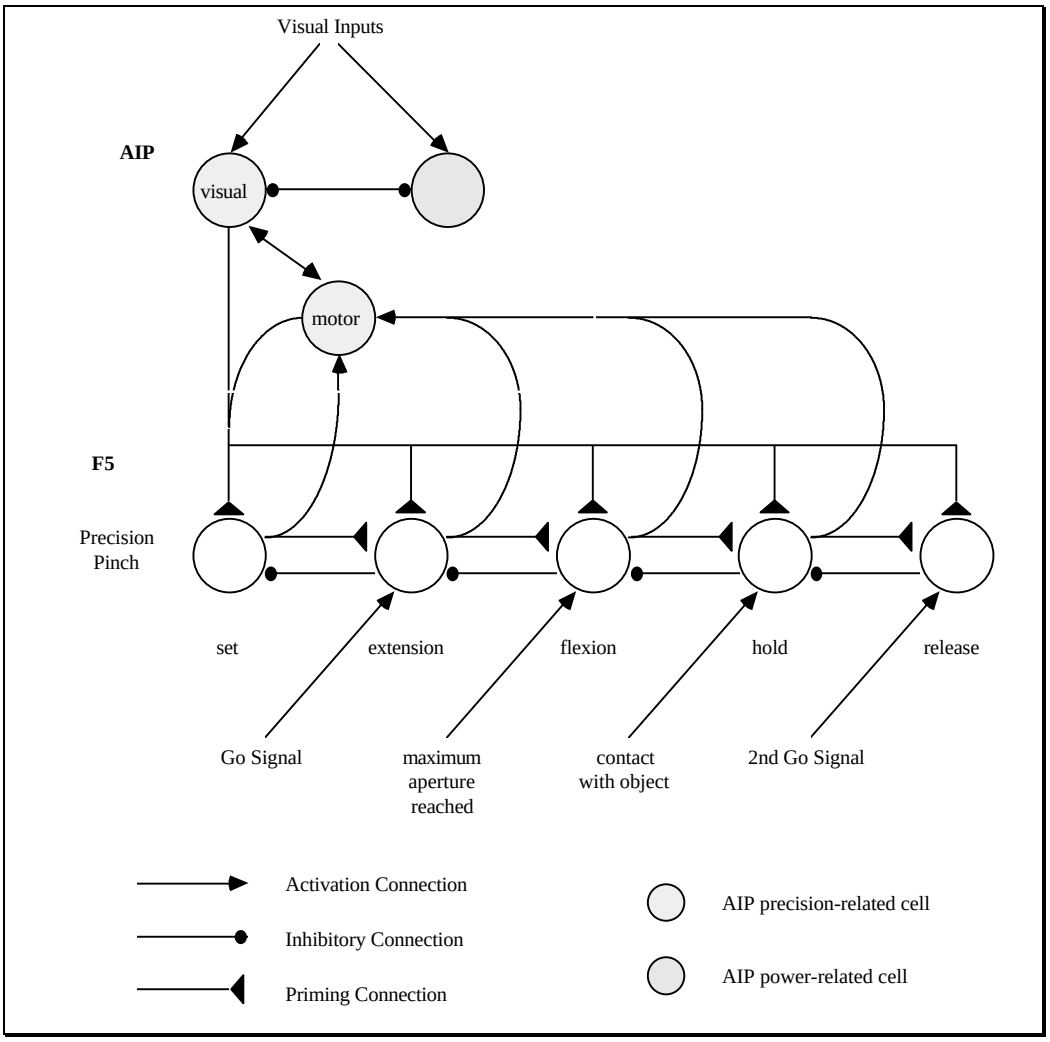


Figure 1. Hypothesized information flow in AIP and F5 in the FARS model during execution of the Sakata paradigm. This neural circuit appears as a rather rigid structure. However, we do not hypothesize that connections implementing the phasic behavior are hardwired in F5. Instead, we posit that sequences are stored in pre-SMA (a part of the supplementary motor area) and administered by the basal ganglia.

The basic logic shown in Figure 1 is that the visual inputs are processed (via various stages whose details are beyond the present discussion; see Fagg and Arbib, 1998, for details) and then passed to AIP whose cells code (by a population code whose details are again beyond the present discussion) the

various affordances seen in the object. As the figure 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 1, then, we see that 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 (our review of the literature suggests that it is part of the supplementary motor area called pre-SMA) 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 actual administration of the sequence (inhibiting extraneous actions, while priming imminent actions) is carried out by the basal ganglia.

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). 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 2). 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 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).

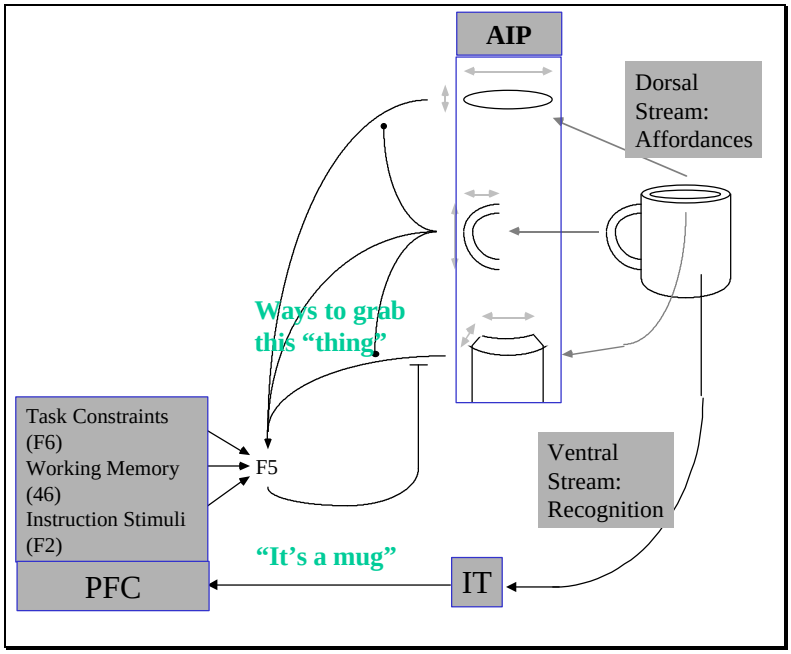


Fig. 2. The role of IT (inferotemporal cortex) and PFC (prefrontal cortex) in modulating F5's selection of an affordance.

Stage 2: Mirror Systems for Grasping

Our task now is to provide a conceptual framework which extends the above "execution system" to include an "observation system", and then to discuss the possibility that the combined system provides a substrate for the evolution of language.

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 3. 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).

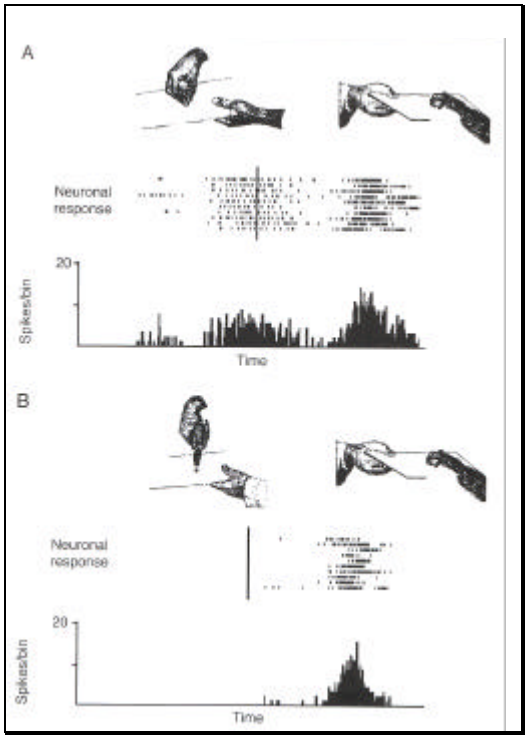


Figure 3. Example of a mirror neuron. Upper part of each panel: behavioral situations. Lower part: neuron's responses. The firing pattern of the neuron on each of a series of consecutive trials is shown above the histogram which sums the response from each trial. A (left): The experimenter grasps a piece 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.

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. It is the canonical neurons, with their input from AIP, that are modeled in the FARS model. Mirror neurons receive different input from the parietal cortex, i.e., not from AIP, encoding observations of arm and hand movements.

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 was tested by two PET experiments (Rizzolatti et al., 1996; Grafton et al., 1996b). The two experiments differed in many aspects, but both had a condition in which subjects observed the experimenter grasping a 3-D object. Object observation was used as a control situation. (This condition also controlled for verbalization.) 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.

F5 is generally considered to be the homologue of Broca's area (see Rizzolatti and Arbib 1998 for the details). 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.

Learning in the Mirror System

For both oro-facial and grasp mirror neurons we may have a limited "hard-wired" repertoire that can then be built on through learning:

1) Developing a set of basic grasps that are effective;

2) Learning to associate view of one's hand with grasp and object;

3) Matching this to views of others grasping;

4) Learning new grasps by imitation of others.

We do not know if the necessary learning is in F5 or elsewhere. In any case, our working hypothesis is that Properties (1) - (3) are present not only in monkey but in the common ancestor of human and monkey, whereas rudimentary forms of imitation were not available to this ancestor, but were available to the common ancestor of chimp and human. We shall discuss this further in the section entitled "Stage 3: An Imitation System for Grasping", but here we want to contrast the manual system of the monkey with the vocalization system of the monkey.

For want of better data, we will assume that the common ancestor of humans and monkeys shared with monkeys primate call system (a limited set of species-specific calls) and an oro-facial gesture system (a limited set of gestures expressive of emotion and related social indicators), as shown in Figure 4. I include a linkage between the two systems to stress that communication is inherently multi-modal. Body posture also plays a role in social communication, but I shall not emphasize this here.

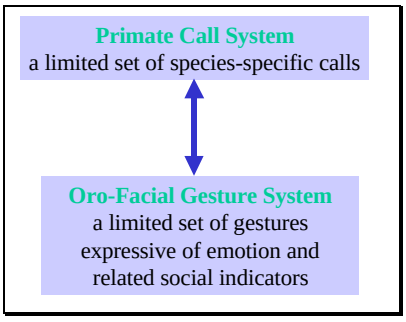


Figure 4. Hypothesized communication system for common ancestor of human and monkey.

What is to be stressed here is that

(i) combinatorial properties for the openness of communication 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 area already involved in

vocalization, which 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 chimp and monkey, but which receive (as we shall see) further refinements in these species.

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 their action 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. As we shall argue in the next section, in human evolution this may have occurred somewhere between the common ancestor of monkey and human, and the common ancestor of chimpanzee and human. 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").

"Prediction" means "creates a neural representation of a potential future state" rather than "is aware of this potential future state". Similarly "understanding", at the level of the monkey's mirror system, means "to be able to match an external (unknown) event to an internal (known) event", without any assumption as to who or what knows the internal event. Similarly for imitation. 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. Of course, this does not deny the crucial point that our development, as "modern" humans, as individuals within a

language-based society greatly extends our understanding beyond that possible for humans raised apart from a language community.

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 here 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 5 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 in such a way as to create expectations which, as we shall shortly see, 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 together those symptoms for self-movement 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.

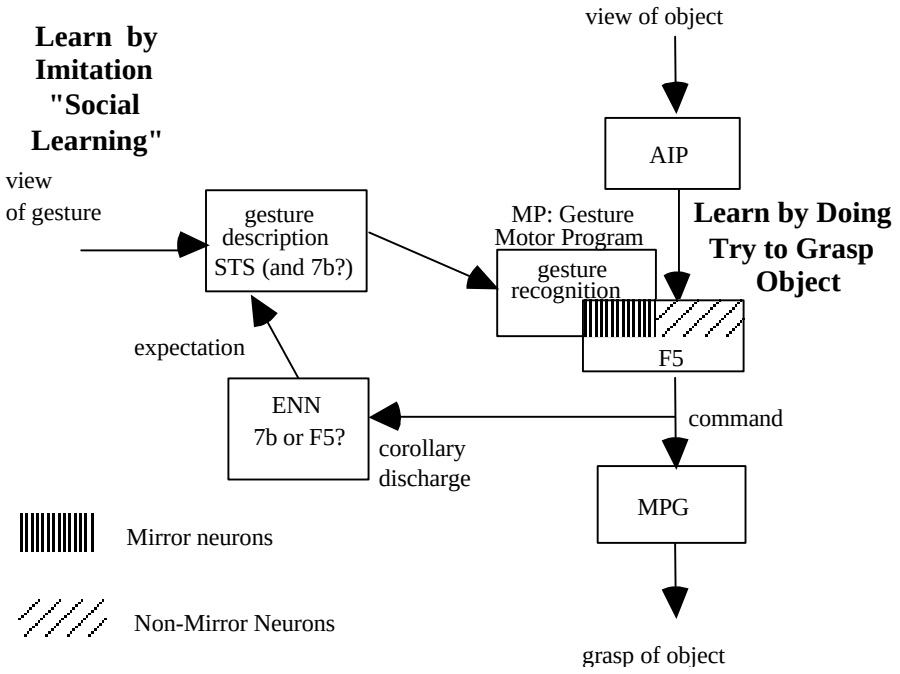


Figure 5. An integrated conceptual framework for analysis of the role of F5 in grasping. The right hand, vertical, path is the **execution system** from "view of object" via AIP and F5 to the motor pattern generator (MPG) for grasping a (seen) object. The loop on the left of the figure provides mechanisms for imitating observed gestures 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 gesture" via gesture description (STS) and gesture recognition (mirror neurons in F5 and possibly 7b) to a representation of the "command" for such a gesture, and the **expectation system** from an F5 command via the expectation neural network ENN to MP, the motor program for generating a given gesture. The latter path may mediate a comparison between "expected gesture" and "observed gesture" in the case of the monkey's self-generated movement.

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 a gesture 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 gestures - with the observation / execution system entirely located in the frontal lobe.

The integrated model of Figure 5 thus relates the "grasp an object" system to the "view a gesture" system. The expectation network is driven by F5 irrespective of whether the motor command is "object-driven" (via AIP) or "gesture-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 "gesture-driven" (a "social action" aimed at making a self-generated movement approximate an observed movement). The right hand path of Figure 4 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 4 exemplifies another mode of learning (the two may be sequential or contemporary) which creates expectations about gestures as well as exemplifying "social learning" based on imitation of gestures made by others. Note that the expectation network is here driven by F5 irrespective of whether the motor command is "object-driven" (via AIP) or "gesture-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 "gesture-driven" (a "social action" aimed at making a self-generated movement which approximates - by some criterion that does not match body-centered localization - an observed movement).

Bridging from Action to Language: The Mirror-System Hypothesis

Hewes (1973), Corballis (1991, 1992), Kimura (1993), and Armstrong et al. (1995) are among those who argued earlier that gestural communication played a crucial role in human language evolution. In this regard, we stress that the "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 motor 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, language may require its own "copy" of motor sequencing mechanisms, with the adjacency of these to "old" mechanisms. This makes lesions which dissociate the two very rare.

With this understanding that the mirror system in monkey is the homolog 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 sign language (based on manual gesture) preceded speech in the evolution of language. Their 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.

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

(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 does not say that the ability to match the perception and production of ***single*** gestures is sufficient for language. In fact, it is not even sufficient for imitation. The subtleties in going from "recognizing a familiar action" to "imitating a complex behavior based on an interweaving of variations on familiar actions" was illustrated in the opening description of a dance class in Santa Fe, and it 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, how 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, i.e. 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.

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 monkey-like human 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 3 stages of posited biological evolution – 3. an imitation system for grasping; 4. a manual-based communication system, and 5. speech – 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 it 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 then list several criteria for language to help guide our understanding of what it means for the human brain to "have language" or "be language-ready":

Naming: The ability to associate an arbitrary symbol with a class of objects ***or actions***

Parity: What counts for the speaker must count for the listener (**Mirror Property**)

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

Temporal Ordering: Temporal activity coding hierarchical structures "of the mind"

Beyond the Here-and-Now: Verb tenses (language) 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

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

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*?

Stage 3: An Imitation System for Grasping

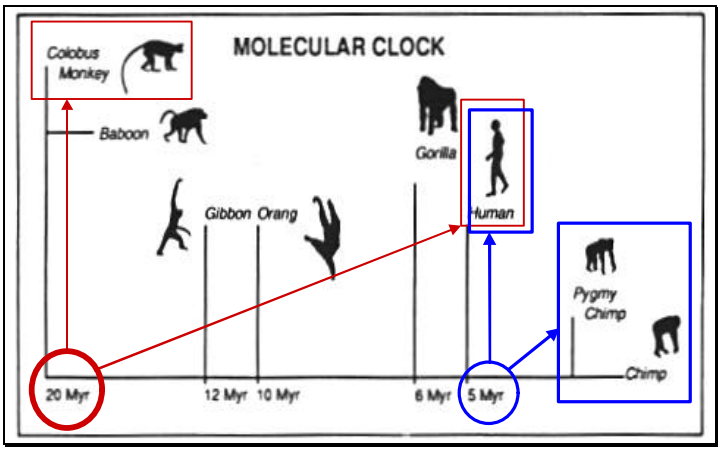


Figure 4. The timetable for hominid evolution inferred from the molecular clock. (Adapted from Clive Gamble: *Timewalkers* Figure 4.2.)

Figure 4 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. Our quest to explore the hypothesis that the mirror system provided the basis for the evolution of human language leads us to two subsidiary questions:

How have the mirror systems of monkey and human diverged from that of their common ancestor?

How have the mirror systems of chimp and human diverged from that of their common ancestor?

What were the properties shared by brains of human, chimp and the common ancestor relevant to language?

It is also clear that mirror neurons may well be fundamental to *imitation*, but it seems that imitation is little developed in the monkey relative to the chimpanzee, so that the utility of the mirror system in the common ancestor of human and monkey presumably resides in functions other than imitation -- we suggest that it functions both in the infant's learning how to observe it's own motor behavior, and in learning how to relate its own actions to those of others.

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; monkeys do not. Chimps can learn a sequence quickly, whereas the monkey cannot. This leads to the following hypothesis: *Extension of the mirror system from single actions to 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.

Important evidence for this imitation 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 (1981, 1984) 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. However, the form of imitation involved here is a long and laborious process compared to the rapidity with which humans can acquire novel sequences if the sequences are not too long and the components are relatively familiar. 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).

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 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 thus 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.

Figure 5 focuses on the generation and observation of a single hand action. We will need to "reflect this up" in a later section to look at its extension to handle the imitation of compound sequences in a way

that meets criteria abstracted from the dance class example. There is a crucial distinction between the monkey's slow conditioning to a particular sequence from the human recognition of the (*parameterizable*) concept of sequence.

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 $\rightarrow$ Q, and an output function β: Q $\rightarrow$ 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.

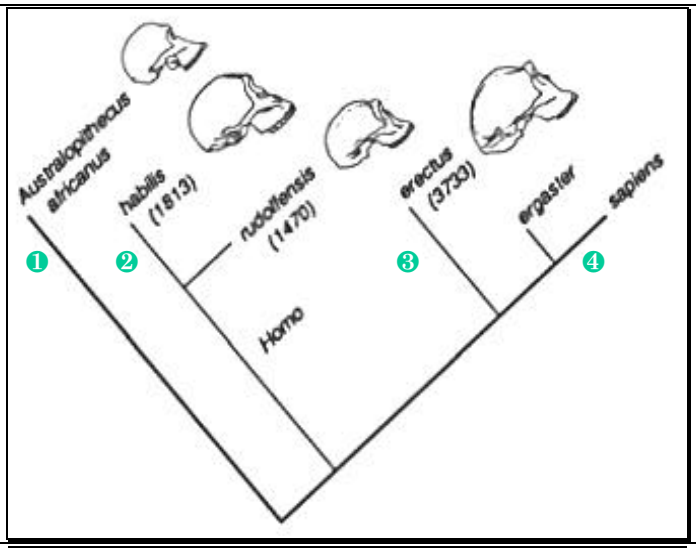
At a basic level, then, we might characterize imitation in terms of ability to "infer automata", recognizing the set of relevant outputs Y (the task of the mirror system) and overt transition signals X, and "inferring" a set of states Q and a set of "covert inputs" X' which allow one to mimic the observed 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 For example, 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.

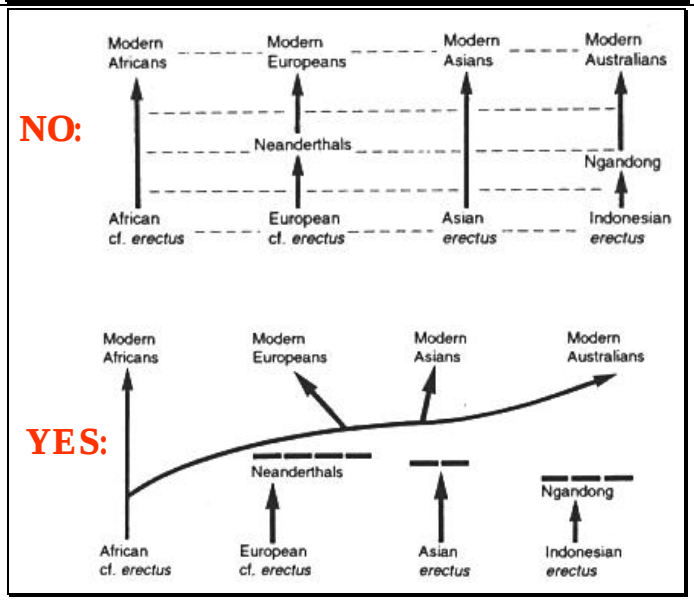
Stage 4: A Manual-Based Communication System

The story of hominid evolution is briefly summarized in Figure 6. Imprints in the cranial cavity of endocasts indicate that "speech areas" were already present in early hominids such as *H. 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.



(6a) A family tree for hominids, with the common ancestor of humans, Homo sapiens (4), and australopithecines (1) preceding that for humans and homo habilis (2) which in turn precedes the common ancestor of humans and homo erectus (3).



(6b) 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.

Figure 6. Five million years of hominid evolution

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 5, 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 for manual gesture is then

i. pragmatic action directed towards a goal object (common ancestor of monkey and human)

ii. imitation of such actions (common ancestor of chimp and human)

iii. pantomime in which similar actions are produced away from the goal object

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.

v. the use of such elements 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) were present in pre-human hominids, that (v) was present in a rather limited form, and 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 and brain even after the skeletal form of *Homo sapiens* was essentially stabilized.

Stage 5: Speech

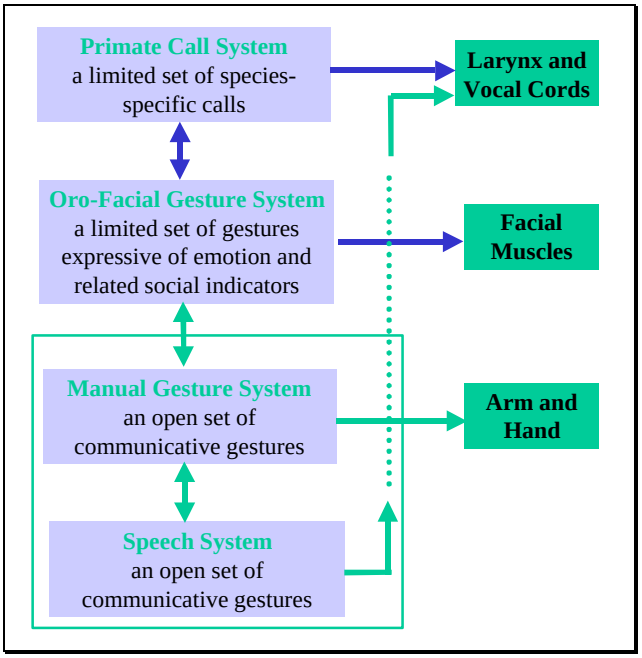


Figure 8. A production view of the evolved speech system of early humans. (Perception systems are not shown.)

We earlier noted that the neural substrate for primate calls is in a region of cingulate cortex distinct from F5, which we have seen to be the monkey 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 in particular, and language in general. Rizzolatti and Arbib (1998) answer this by suggesting three evolutionary stages going beyond the capacities of Figure 4:

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

2. The "speech" area of early hominids (i.e., the area somewhat 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. P. Lieberman views the descent of the larynx seen in *Homo sapiens* as being crucial in enabling the wide articulatory range exploited in human speech. Clearly, some level of language-readiness and *vocal* language preceded this -- a core of proto-speech was needed to provide pressures for larynx evolution.

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

I have schematized the result of the above three evolutionary stages in Figure 8. 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.

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 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 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 that provides 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 imprecations!), 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. 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 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 biological and cultural evolution of language, with advances in the one triggering advances in the other.

The Transition to *Homo sapiens*

The ability for visual scene perception that must underlie the 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. I thus hypothesize that the ability to communicate a fair number of such frames was established in the hominid line prior to the emergence of *Homo sapiens.* Indeed, consideration of the spatial basis for "prepositions" may help show how visuomotor coordination underlies some aspects of language. However, the basic semantic-syntactic correspondences have been overlaid by a multitude of later innovations and borrowings.

The transition to *Homo sapiens* may then have involved "language amplification" through increased speech capability, yielding an increased ability to name actions and objects to create an unlimited 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. How did the needs of human biology and the constraints of the human brain shape these basic "discoveries"?

Butler and Hodos (1996) give a useful account of how vertebrate brains evolve: The course of brain evolution among vertebrates has been determined by (i) Formation of multiple new nuclei through elaboration or duplication; (ii) Regionally specific increases in cell proliferation in different parts of the brain; and (iii) Gain of some new connections and loss of some established connections. These phenomena 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.

In contrasting monkey and human brain, we find

(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) to extend "cognitive comprehension" in space and time

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

(c) Extension of the "reach" of mirror systems

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

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

(f) 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 sees Prefrontal cortex (PFC) as evolving to increase working memory capacity. Petrides argues that we need PFC to go beyond single items to keeping multiple objects or events in order. Note the challenge of embedding the mirror system in a system for handling sequential structure, and hierarchical structure more generally. How does this relate to the role of hippocampus in episodic memory? Note the parallel problem of keeping multiple objects in *spatial relation* in scene perception – and the related syndrome of simultagnosia. 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 speech has long since created the mistaken impression that language is a speech-based system. However, McNeill 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 have only relatively recently discovered (Bellugi and Klima), 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 study of deaf signers suggests that we locate phonology in a *speech-manual-orofacial gesture complex*. In then hypothesize 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 8 by saying that "The vote is in: one box replaces three", as shown in Figure 9. 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 8 nor Figure 9 shows the neuroanatomy of these mechanisms.

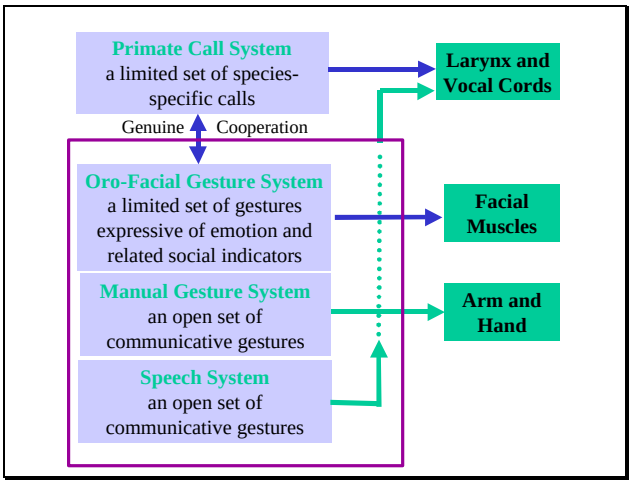


Figure 9. The fruit of evolution: Not three separate communication systems, but a single system operating in at least three motor modalities and at least two sensory modalities.

Language Evolving

Deep Time

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 have already hypothesized that e*xtension of the mirror system from single actions to compound actions* was the key innovation in the brain of the common ancestor of human and chimp (as compared to the monkey-human common ancestor) relevant to language-readiness. Further development of our theory requires us to suggest how this extension of the mirror system was further refined along the hominid evolutionary track. We need to distinguish sequential learning at two levels:

(i) the abstraction of regularities in many sentences to come up with "syntax";

(ii) the ability, given syntactic and semantic knowledge, to extract the sequential or semantic structure of an utterance (parsing) to reflect meaning upward from basic units via constituent structures to larger units.

I have argued that the evolution of *Homo sapiens* (Figure 6) that *biological* evolution 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 P. Lieberman observes, makes humans able to articulate more precisely than other primates, but at the

precise of an increased likelihood of choking. (Colin McLeod quips that "The human vocal tract evolved so that we could cry out "Help, I'm choking!) 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.[1] From this perspective, some of the basic questions are:

1a. What were those features of the human brain that pre-adapted us for human language, i.e., made the human brain “language ready”?

1b. What aspects of human language were already present in the earliest humans of 200,000 years ago?

2a. How can this perspective on language evolution help explain why language looks the way it does today?

2b. What has been the interplay of biological inheritance and cultural evolution since the emergence of *Homo sapiens*?

There is a subsidiary, though crucial, question here: How can we best describe cultural evolution in a way which reflects both its dependence on that biological inheritance and the vast variety behavior exhibits across different cultures?

The biological basis of human evolution includes bipedality, manual dexterity, and a larynx well-suited for vocal production. The Mirror System Hypothesis stresses the ability to relate the actions of others to one’s own actions. However, analysis of human language draws our attention to several key abilities "beyond the mirror", i.e., that involve more than the recognition of single actions. These culminate in the ability to rapidly acquire a vast array of flexible strategies for pragmatic and communicative action, and the ability to generate and comprehend hierarchical structures “on the fly”.

What then of social evolution? Social evolution can result from both biological and cultural (non-genetic) evolution. Clearly, language enables an immense amplification of the second factor. Human evolution saw the co-evolution of increasingly complex social structures and of increasingly complex patterns of behavior and communication to serve those social interactions. Gamble in ***Timewalkers*** views human evolution in terms of preadaptation for global colonization, with language one of many relevant traits in that evolution. He emphasizes the relation of humans to other species in the same environment.

To proceed, we return to the FARS Model of Figures 1 and 2. We saw that the mirror system for grasping in monkey is a subset of the grasp-related machinery in F5. To simplify somewhat, we may

---

[1] 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.

distinguish two main subsystems of F5: The "canonical" system with input from parietal area AIP, and the "mirror" system which has been shown to have input from another parietal area, PF, more tuned to the recognition of motions than of objects. The FARS (Fagg-Arbib-Rizzolatti-Sakata) model provided a computational account of the canonical system, showing how it can account for basic phenomena of grasping. F5 alone is not the "full" mirror system. We want not only the "unit actions" but also sequences and more general patterns. The FARS model sketches how to generate a sequence positing roles for SMA and BG. Our ***proposed*** mirror model (Oztop and Arbib, to appear) must match this with a model of how the units of a sequence *and* their order/interweaving can be recognized. This *new* model requires recognition of a complex behavior on multiple occasions with increasing success in recognizing component actions and in linking them together. [cf. scene analysis in vision.]

The "basic" Mirror System relates observed grasping actions to ones which the observer can himself perform. This raises two data questions:

For monkey: What are the limits on the actions represented in F5? How open is F5 to learning to perform and observe new actions? (cf. the pliers story.)

For human: How does learning enable modern humans to extend the repertoire of recognizable and describable actions well beyond those that can be performed by the speaker/hearer? (consider "the bird is flying").

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

Our starting point for the biological basis of language-readiness is 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). This reflects a crucial understanding gained from Figure 2 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 (PF) 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. (Hypothesis: Inactivation of IT which disturbs neither PF nor AIP will disrupt mirror activity in F5, but not canonical activity in F5.) It is this matching of actions with "object semantics" and "goals", rather than just with affordances, that makes possible the transition to:

3. Naming objects and actions. This involves the creation of symbols for **verb-argument structures** closely linked to specific action-object frames. However, I must stress a subtle point here: the original form of such structures need not have involved separate lexical entries for the verb and the argument. 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.

Stage 3 is then the first steep in the transition to language: abstract symbols are grounded (but more and more indirectly) in action-oriented perception, and 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 noting their use by others. The problem to be solved next probably requires no biological evolution -- the creation of a "symbol toolkit" of meaningless elements from which an open ended class of symbols can be generated. The distinction I have in mind here relates to the earlier discussion of moving beyond pantomime: 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 one integrated system for communication: In this scenario, the ability to create novel sounds to imitate manual gestures in the vocal domain, 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 either case, 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.

Mirror Neurolinguistics: An Early Response to the Challenge

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).

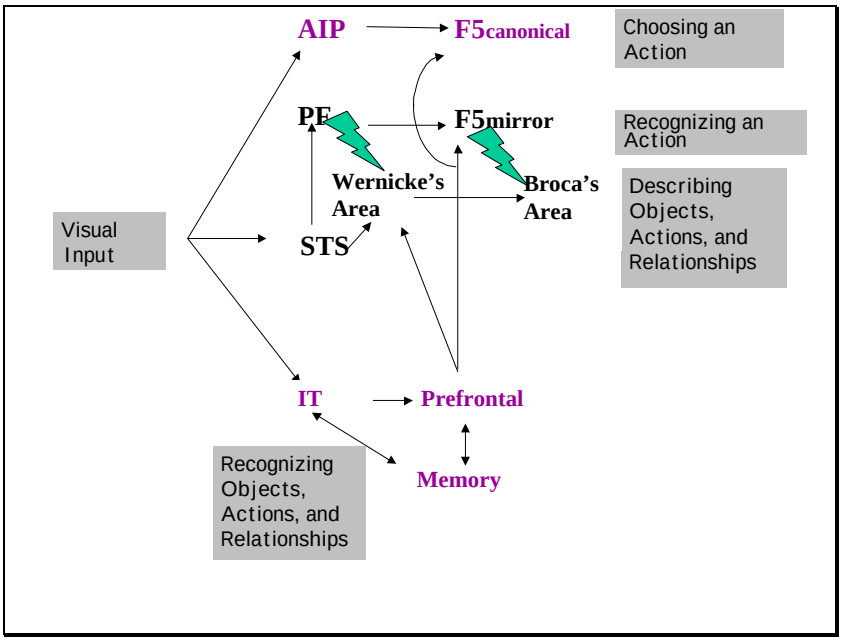


Figure 10. Extending the FARS model to include thee mirror system for grasping and the language system evolved "atop" this.

To resolve this apparent paradox, we must return to the view of the FARS model given in Figure 1, 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 10. 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.

Acknowledgements: I recall with pleasure my many discussions with Giacomo Rizzolatti and Giuseppe Cossu which laid the basis for the present article while I was on sabbatical in Rizzolatti's Institute in Parma, Italy, in May, June and July of 1999. I thank Aude Billard for her constructive comments on earlier drafts.

References

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

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

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

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

Armstrong, D., Stokoe, W., & Wilcox, S. (1995), *Gesture and the Nature of Language,* Cambridge U Press, Cambridge, MA.

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

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

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

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

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

Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G., 1995, Motor facilitation during action observation: a magnetic stimulation study. *J Neurophysiol*., 73: 2608-2611.

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

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

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

Grafton, S.T., Arbib, M.A., Fadiga, L., and Rizzolatti, G., 1996b, Localization of grasp representations in humans by PET: 2. Observation compared with imagination. *Exp Brain Res.* 112:103-111.

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

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

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

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

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

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

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

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

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

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G. & Matelli, M. (l988) 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.

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

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

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

Taira, M., Mine, S., Georgopoulos, A.P., Murata, A., and Sakata, H., Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Exp. Brain Res.*, 83 (1990) 29-36.