Part I

Two perspectives

1

The Mirror System Hypothesis on the linkage of action and languages

Michael A. Arbib

1.1 Introduction

Our progress towards an understanding of how the human brain evolved to be ready for language starts with the mirror neurons for grasping in the brain of the macaque monkey. Area F5 of the macaque brain is part of premotor cortex, i.e., F5 is part of the area of cerebral cortex just in front of the primary motor cortex shown as F1 in Fig. 1.1 (left). Different parts of F5 contain neurons active during manual and orofacial actions. Crucially for us, an anatomically segregated subset of these neurons are *mirror neurons*. Each such mirror neuron is active not only when the monkey performs actions of a certain kind (e.g., a precision pinch or a power grasp) but also when the monkey observes a human or another monkey perform a more or less similar action. In humans, we cannot measure the activity of single neurons (save when needed for testing during neurosurgery) but we can gather comparatively crude data on the relative blood flow through (and thus, presumably, the neural activity of) a brain region when the human performs one task or another. We may then ask whether the human brain also contains a "mirror system for grasping" in the sense of a region active for both execution and observation of manual actions as compared to some baseline task like simply observing an object. Remarkably, such sites were found in frontal, parietal, and temporal cortex of the human brain. Most significantly for this book, the frontal activation was found in or near Broca's area (Fig. 1.1 right), a region which in most humans lies in the left hemisphere and is traditionally associated with speech production. Moreover, macaque F5 and human Broca's area are considered to be (at least in part) homologous brain regions, in the sense that they are considered to have evolved from the same brain region of the common ancestor of monkeys and humans (see Section 1.2 as well as Arbib and Bota, this volume).

But why should a neural system for language be intimately related with a mirror system for grasping? Pondering this question led Giacomo Rizzolatti and myself to formulate the Mirror System Hypothesis (MSH) (Arbib and Rizzolatti, 1997; Rizzolatti and Arbib, 1998) which will be presented more fully in Section 1.3.1. We view the mirror system for

Action to Language via the Mirror Neuron System, ed. Michael A. Arbib. Published by Cambridge University Press.

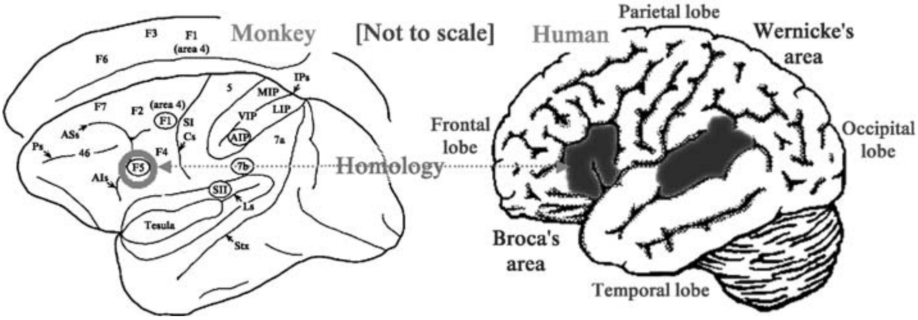

Figure 1.1 A comparative side view of the monkey brain (left) and human brain (right), not to scale. The view of the monkey brain emphasizes area F5 of the frontal lobe of the monkey; the view of the human brain emphasizes two of the regions of cerebral cortex, Broca's area and Wernicke's area, considered crucial for language processing. The homology between monkey F5 and human Broca's area is a key component of the present account.

grasping as a key neural "missing link" between the abilities of our non-human ancestors of 20 million years ago and the modern human capability for language, with manual gestures rather than a system for vocal communication providing the initial seed for this evolutionary process. The present chapter, however, goes "beyond the mirror" to offer hypotheses on evolutionary changes within and outside the mirror system which may have occurred to equip *Homo sapiens* with a language-ready brain, a brain which supports language as a behavior in which people communicate "symbolically" through integrated patterns of vocal, manual and facial movements. Language is essentially *multi-modal*, not just a set of sequences of words which can be completely captured by marks on the printed page.

1.1.1 Protolanguage defined

In historical linguistics (e.g., Dixon, 1997), a *protolanguage* for a family of extant human languages is the posited ancestral language – such as proto-Indo-European for the family of Indo-European languages – from which all languages of the family are hypothesized to descend historically with various modifications. In this chapter, however, we reserve the term *protolanguage* for a system of utterances used by a particular hominid species (possibly including *Homo sapiens*) which we would recognize as a precursor to human language (if only we had the data!), but which is not itself a human language in the modern sense.

Bickerton (1995) hypothesized that the protolanguage of *Homo erectus* was a system whose utterances took the form of a string of a few words much like those of today's language, but which conveyed meaning without the aid of syntax. On this view, language just "added syntax" through the evolution of Universal Grammar. Contrary to Bickerton's hypothesis, I argue (Section 1.5) that the protolanguage of *Homo erectus* and early *Homo* *sapiens* was “holophrastic”, i.e., composed mainly of “holophrases” or “unitary utterances” (phrases in the form of semantic wholes with no division into meaningful subparts) which symbolized frequently occurring situations and that words as we know them then co-evolved culturally with syntax through fractionation. My view is shared, e.g., by Wray (2002) and opposed by Hobbs (this volume). Bickerton (2005) and Wray (2005) advance the debate within the context of the Mirror System Hypothesis.

1.1.2 Relating language to the vocalizations of non-human primates

Humans, chimpanzees, and monkeys share a general physical form and a degree of manual dexterity, but their brains, bodies, and behaviors differ. Humans have abilities for bipedal locomotion and learnable, flexible vocalization that are not shared by other primates. Monkeys exhibit a primate call system (a limited set of species-specific calls) and an orofacial (mouth and face) gesture system (a limited set of gestures expressive of emotion and related social indicators). Note the linkage between the two systems: communication is inherently multi-modal. This communication system is *closed* in the sense that it is restricted to a specific repertoire. This is to be contrasted with the open nature of human languages which can form endlessly many novel sentences from the current word stock and add new words to that stock. Admittedly, chimpanzees and bonobos (great apes, not monkeys) can be trained to acquire a form of communication – based either on the use of hand signs or objects that each have a symbolic meaning – that approximates the complexity of the utterances of a 2-year-old human child, in which a “message” generally comprises one or two “lexemes.” However, there is no evidence that an ape can reach the linguistic ability of a 3-year-old human. Moreover, such “ape languages”[1] are based on hand–eye coordination rather than vocalization whereas a crucial aspect of human biological evolution has been the emergence of a vocal apparatus and control system that can support speech.

It is tempting to hypothesize that certain species-specific vocalizations of monkeys (such as the “snake call” and “leopard call” of vervet monkeys) provided the basis for the evolution of human speech, since both are in the vocal domain (see Seyfarth (2005) for a summary of arguments supporting this view[2]). However, combinatorial properties for the openness of communication are virtually absent in basic primate calls and orofacial communication, though individual calls may be graded. Moreover, Jürgens (1997, 2002) found that voluntary control over the initiation and suppression of such vocalizations relies on the mediofrontal cortex including anterior cingulate gyrus (see Arbib and Bota, this volume). MSH explains why it is F5, rather than the cingulate area involved in macaque vocalization, that is homologous to the human’s frontal substrate for language by asserting that a *specific* mirror system – the primate mirror system for grasping – evolved into a

[1] I use the quotes because these are not languages in the sense in which I distinguish languages from protolanguages.

[2] For some sense of the debate between those who argue that protosign was essential for the evolution of the language-ready brain and those who take a “speech only” approach, see Fogassi and Ferrari (2004), Arbib (2005b), and MacNeilage and Davis (2005).

key component of the mechanisms that render the human brain language-ready. It is this specificity that will allow us to explain below why language is multi-modal, its evolution being rooted in the execution and observation of hand movements and extended into speech.

Note that the claim is not that Broca's area is genetically preprogrammed for language, but rather that the development of a human child in a language community normally adapts this brain region to play a crucial role in language performance. Zukow-Goldring (this volume) takes a mirror-system-oriented view of cognitive development in the child, showing how caregivers help the child learn the effectivities of her own body and the affordances of the world around her. As we shall see later in this chapter, the notion of *affordance* offered by Gibson (1979) – that of information in the sensory stream concerning opportunities for action in the environment – has played a critical role in the development of MSH, with special reference to neural mechanisms which extract affordances for grasping actions. Not so much has been made of *effectivities*, the range of possible deployments of the organism's degrees of freedom (Turvey *et al*., 1981), but Oztop *et al*. (this volume) offer an explicit model (the Infant Learning to Grasp Model, ILGM) of how effectivities for grasping may develop as the child engages in new activities. Clearly, the development of novel effectivities creates opportunities for the recognition of new affordances, and vice versa.

1.1.3 Language in an action-oriented framework

Neurolinguistics should emphasize performance, explicitly analyzing both perception and production. The framework sketched in Fig. 1.2 leads us to ask: to what extent do language mechanisms exploit existing brain mechanisms and to what extent do they involve biological specializations specific to humans? And, in the latter case, did the emergence of language drive the evolution of such mechanisms or exploit them?

Arbib (2002, 2005a) has extended the original formulation of MSH to hypothesize a number of (possibly overlapping) stages in the evolution of the language-ready brain. As we shall see in more detail in Section 1.3.2, crucial early stages extend the mirror system for grasping to support imitation, first so-called "simple" imitation such as that found in chimpanzees (which allows imitation of "object-oriented" sequences only as the result of extensive practice) and then so-called "complex" imitation found in humans which combines the perceptual ability to recognize that a novel action was in fact composed of (approximations to) known actions with the ability to use this analysis to guide the more or less successful reproduction of an observed action of moderate complexity.[3] Complex imitation was a crucial evolutionary innovation in its own right, increasing the ability to learn and transmit novel skills, but also provides a basis for the later emergence of the language-ready brain – it is crucial not only to the child's ability to acquire

[3] See the discussion of research by Byrne (2003) on "program-level imitation" in gorillas (cf. Stanford, this volume). A key issue concerns the divergent time course of acquisition of new skills in human versus ape.

Cambridge University Press
978-0-521-84755-1 - Action to Language via the Mirror Neuron System
Edited by Michael A. Arbib
Excerpt
More information

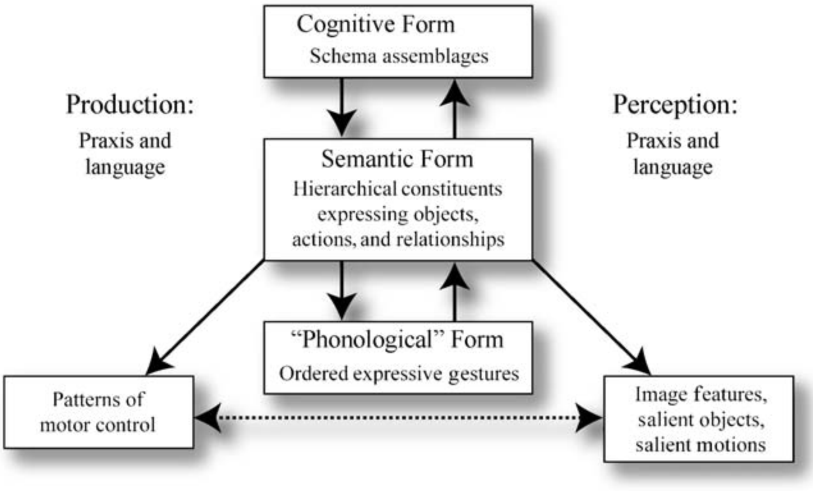
Cognitive Form
Schema assemblages
Production:
Praxis and language
Perception:
Praxis and language
Semantic Form
Hierarchical constituents expressing objects, actions, and relationships
"Phonological" Form
Ordered expressive gestures
Patterns of motor control
Image features, salient objects, salient motions

Figure 1.2 The figure places production and perception of language within a framework of action and perception considered more generically. Language production and perception are viewed as the linkage of Cognitive Form, Semantic Form, and Phonological Form, where the "phonology" may involve vocal or manual gestures, or just one of these, with or without the accompaniment of facial gestures.

language and social skills, but the perceptual ability within it is essential to the adult use of language.

But it requires a further evolutionary breakthrough for the complex imitation of action to yield pantomime, as the purpose comes to be to communicate rather than to manipulate objects. This is hypothesized to provide the substrate for the development of protosign, a combinatorially open repertoire of manual gestures, which then provides the scaffolding for the emergence of protospeech (which thus owes little to non-human vocalizations), with protosign and protospeech then developing in an expanding spiral. Here, I am using "protosign" and "protospeech" for the manual and vocal components of a protolanguage, in the sense defined earlier. I argue that these stages involve biological evolution but that the progression from protosign and protospeech to languages with full-blown syntax and compositional semantics was a historical phenomenon in the development of *Homo sapiens*, involving few if any further biological changes.

For *production*, the notion is that at any time we have much that we could possibly talk about which might be represented as cognitive structures (Cognitive Form; schema assemblages) from which some aspects are selected for possible expression. Further selection and transformation yields semantic structures (hierarchical constituents expressing objects, actions and relationships) which constitute a Semantic Form enriched by linkage to schemas for perceiving and acting upon the world. Finally, the ideas in the Semantic Form must be expressed in words whose markings and ordering reflect the relationships within Semantic Form. These words must be conveyed as "phonological" structures – where here I extend phonological form to embrace a wide range of ordered expressive gestures which may include speech, sign, and orofacial expressions (and even writing and typing).

For *perception*, the received utterance must be interpreted semantically with the result updating the "hearer's" cognitive structures. For example, perception of a visual scene may reveal "Who is doing what and to whom/which" as part of a non-linguistic *action–object frame* in cognitive form. By contrast, the *verb–argument structure* is an overt linguistic representation in semantic form – in most human languages, the action is named by a verb and the objects are named by nouns or noun phrases. A production grammar for a language is then a specific mechanism (whether explicit or implicit) for converting verb–argument structures into strings of words (and hierarchical compounds of verb–argument structures into complex sentences) and vice versa for perception.

These notions of "forward" and "inverse" grammars may be compared with the forward and inverse models discussed by Oztop *et al*. (this volume) and by Skipper *et al*. (this volume).

Emmorey (this volume) discusses the neural basis for parallels between praxis (including spatial behavior), pantomime, sign production, and speech production, establishing what is modality-general for brain mechanisms of language and what is modality-specific. Here I would suggest a crucial distinction between signed language and speech. In both speech and signing, I claim, we recognize a novel utterance as in fact composed of (approximations to) known actions, namely uttering words, and – just as crucially – the stock of words is open-ended. However, signed language achieves this by a very different approach to speech. Signing exploits the fact that the signer has a very rich repertoire of arm, hand, and face movements, and thus builds up vocabulary by variations on this multidimensional theme (move a hand shape (or two) along a trajectory to a particular position while making appropriate facial gestures). By contrast, speech employs a system of articulators that are specialized for speech – there is no rich behavioral repertoire of non-speech movements to build upon.[4] Instead evolution "went particulate" (Studdert-Kennedy, 2000; Goldstein *et al*., this volume), so that the spoken word is built (to a first approximation) from a language-specific stock of phonemes (actions defined by the coordinated movement of several articulators, but with only the goal of "sounding right" rather than conveying meaning in themselves).

But if single gestures are the equivalent of phonemes in speech or words in sign, what levels of motor organization correspond to derived words, compound words, phrases, sentences, and discourse? Getting to derived words seems simple enough. In speech, we play variations on a word by various morphological changes which may modify internal phonemes or add new ones. In sign, "words" can be modified by changing the source and

[4] However, it is interesting to speculate on the possible relevance of the oral dexterity of an omnivore – the rapid adaptation of the coordination of chewing with lip, tongue, and swallowing movements to the contents of the mouth – to the evolution of these articulators. Such oral dexterity goes well beyond the "mandibular cyclicities" posited by MacNeilage (1998) to form the basis for the evolution of the ability to produce syllables as consonant–vowel pairs.

origin, and by various modifications to the path between. For everything else, it seems enough – for both action and language – that we can create hierarchical structures subject to a set of transformations from those already in the repertoire. For this, the brain must provide a computational medium in which already available elements can be composed to form new ones, irrespective of the "level" at which these elements were themselves defined. When we start with words as the elements, we may end up with compound words or phrases, other operations build from both words and phrases to yield new phrases or sentences, etc., and so on recursively. Similarly, we may learn arbitrarily many new motor skills based on those with which we are already familiar.

With this, let me give a couple of examples (Arbib, 2006) which suggest how to view language in a way which better defines its relation to goal-directed action. Consider a conditional, hierarchical motor plan for opening a child-proof aspirin bottle:

> While holding the bottle with the non-dominant hand, grasp the cap, push down and turn the cap with the dominant hand; then repeat (release cap, pull up cap, and turn) until the cap comes loose; then remove the cap.

This hierarchical structure unpacks to different sequences of action on different occasions, with subsequences conditioned on the achievement of goals and subgoals. To ground the search for similarities between action and language, I suggest we view an action such as this one as a "sentence" made up of "words" which are basic actions. A "paragraph" or a "discourse" might then correspond to, e.g., an assembly task which involves a number of such "sentences."

Now consider a sentence like "Serve the handsome old man on the left.", spoken by a restaurant manager to a waiter. From a "conventional" linguistic viewpoint, we would apply syntactic rules to parse this specific string of words. But let us look at the sentence not as a structure to be parsed but rather as the result of the manager's attempt to achieve a *communicative goal*: to get the waiter to serve the intended customer. His *sentence planning strategy* repeats the "loop"

<add adjective or prepositional phrase>

until (he thinks) ambiguity is resolved:

(1) Serve the old man.

Still ambiguous?

(2) Serve the old man on the left.

Still ambiguous?

(3) Serve the handsome old man on the left.

Still ambiguous?

Apparently not. So the manager "executes the plan" and says to the waiter

"Serve the handsome old man on the left."

Here, a noun phrase NP may be expanded by adding a prepositional phrase PP after it (as in expanding (1) to (2) above) or an adjective Adj before it (as in expanding (2) to (3)). The suggestion is that syntactic rules of English which I approximate by NP → NP PP and

NP → Adj NP are abstracted from procedures which serve to reduce ambiguity in reaching a communicative goal. This example concentrates on a noun phrase – and thus exemplifies ways in which reaching a communicative goal (identifying the right person, or more generally, object) may yield an unfolding of word structures in a way that may clarify the history of syntactic structures.

While syntactic constructions can be usefully analyzed and categorized from an abstract viewpoint, the pragmatics of what one is trying to say and to whom one is trying to say it will drive the goal-directed process of producing a sentence. Conversely, the hearer has the inferential task of unfolding multiple meanings from the word stream (with selective attention) and deciding (perhaps unconsciously) which ones to meld into his or her cognitive state and narrative memory.

1.2 Grasping and the mirror system

1.2.1 Brain mechanisms for grasping

Figure 1.3 shows the brain of the macaque (rhesus monkey) with its four lobes: frontal, parietal, occipital, and temporal. The parietal area AIP is near the front (anterior) of a groove in the parietal lobe, shown opened up in the figure, called the intraparietal sulcus – thus AIP (for anterior region of the intraparietal sulcus).[5] Area F5 (the fifth region in a numbering for areas of the macaque frontal lobe) is in what is called ventral premotor cortex. The neuroanatomical coordinates refer to the orientation of the brain and spinal cord in a four-legged vertebrate: "dorsal" is on the upper side (think of the dorsal fin of a shark) while "ventral" refers to structures closer to the belly side of the animal. Together, AIP and F5 anchor the cortical circuit in macaque which transforms visual information on intrinsic properties of an object into hand movements for grasping it. AIP processes visual information to implement perceptual schemas for extracting grasp parameters (affordances) relevant to the control of hand movements and is reciprocally connected with the so-called *canonical neurons* of F5. Discharge in most grasp-related F5 neurons correlates with an action rather than with the individual movements that form it so that one may relate F5 neurons to various *motor schemas* corresponding to the action associated with their discharge. By contrast, primary motor cortex (F1) formulates the neural instructions for lower motor areas and motor neurons.

The FARS (Fagg–Arbib–Rizzolatti–Sakata) model (Fagg and Arbib, 1998) provides a computational account of the system (it has been implemented, and various examples of grasping simulated) centered on this pathway: the dorsal stream via AIP does not know "what" the object is, it can only see the object as a set of possible affordances, whereas the ventral stream from primary visual cortex to inferotemporal cortex (IT), by contrast, is able to recognize what the object is. This information is passed to prefrontal cortex (PFC)

[5] The reader uncomfortable with neuroanatomical terminology can simply use AIP and similar abbreviations as labels in what follows, without worrying about what they mean. On the other hand, the reader who wants to know more about the neuroanatomy should turn to Arbib and Bota (this volume) in due course.

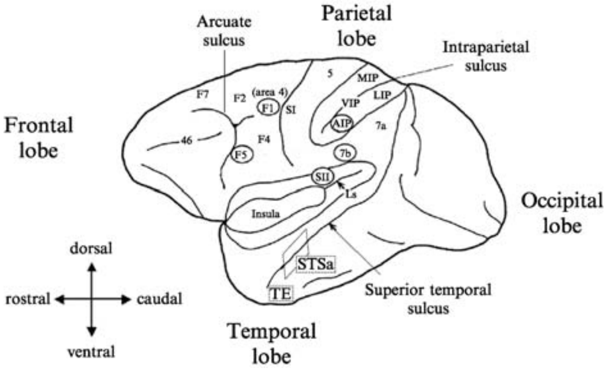

Figure 1.3 A side view of the left hemisphere of the macaque brain. Area 7b is also known as area PF. (Adapted from Jeannerod *et al*., 1995.)

which can then, on the basis of the current goals of the organism and the recognition of the nature of the object, bias the affordance appropriate to the task at hand. The original FARS model suggested that the bias was applied by PFC to F5; subsequent neuroanatomical data (as analyzed by Rizzolatti and Luppino, 2003) suggest that PFC and IT may modulate action selection at the level of parietal cortex rather than premotor cortex.

Figure 1.4 gives a partial view of "FARS Modificato," the FARS model updated to show this modified pathway. AIP may represent several affordances initially, but only one of these is selected to influence F5. This affordance then activates the F5 neurons to command the appropriate grip once it receives a "go signal" from another region, F6, of PFC.

We now turn to those parts of the FARS model which provide mechanisms for sequencing actions: the circuitry encoding a sequence is postulated to lie within the supplementary motor area called pre-SMA (Rizzolatti *et al*., 1998), with administration of the sequence (inhibiting extraneous actions, while priming imminent actions) carried out by the basal ganglia. In Section 1.3, I will argue that the transition to early *Homo* coincided with the transition from a mirror system used only for action recognition and "simple" imitation to more elaborate forms of "complex" imitation: I argue that what sets hominids apart from their common ancestors with the great apes is the ability to rapidly exploit novel sequences as the basis for immediate imitation or for the immediate construction of an appropriate response.

Here, I hypothesize that the macaque brain can generate sequential behavior on the basis of overlearned coordinated control programs (cf. Itti and Arbib, this volume), but that only the human brain can perform the inverse operation of observing a sequential behavior and inferring the structure of a coordinated control program that might have generated it. In the FARS model, Fagg and Arbib (1998) modeled the interaction of AIP and F5 in the sequential task employed by Sakata in his studies of AIP. In this task, the monkey sits with his hand on a key, while in front of him is a manipulandum. A change in an LED signals him to prepare to grasp; a further change then tells him to execute the