BEHAVIORAL AND BRAIN SCIENCES (2005) **28,** 105–167
Printed in the United States of America

From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics

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

Abstract: The article analyzes the neural and functional grounding of language skills as well as their emergence in hominid evolution, hypothesizing stages leading from abilities known to exist in monkeys and apes and presumed to exist in our hominid ancestors right through to modern spoken and signed languages. The starting point is the observation that both premotor area F5 in monkeys and Broca's area in humans contain a "mirror system" active for both execution and observation of manual actions, and that F5 and Broca's area are homologous brain regions. This grounded the mirror system hypothesis of Rizzolatti and Arbib (1998) which offers the mirror system for grasping as a key neural "missing link" between the abilities of our nonhuman ancestors of 20 million years ago and modern human language, with manual gestures rather than a system for vocal communication providing the initial seed for this evolutionary process. The present article, however, goes "beyond the mirror" to offer hypotheses on evolutionary changes within and outside the mirror systems which may have occurred to equip *Homo sapiens* with a language-ready brain. Crucial to the early stages of this progression is the mirror system for grasping and its extension to permit imitation. Imitation is seen as evolving via a so-called simple system such as that found in chimpanzees (which allows imitation of complex "object-oriented" sequences but only as the result of extensive practice) to a so-called complex system found in humans (which allows rapid imitation even of complex sequences, under appropriate conditions) which supports pantomime. This is hypothesized to have provided 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 nonhuman vocalizations), with protosign and protospeech then developing in an expanding spiral. It is argued that these stages involve biological evolution of both brain and body. By contrast, it is argued 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.

Key words: gestures; hominids; language evolution; mirror system; neurolinguistics; primates; protolanguage; sign language; speech; vocalization

1. Action-oriented neurolinguistics and the mirror system hypothesis

1.1. Evolving the language-ready brain

Two definitions:

1. A *protolanguage* is 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 the data were available!), but which is not itself a human language in the modern sense.[1]

2. An infant (of any species) has a *language-ready* brain if it can acquire a full human language when raised in an environment in which the language is used in interaction with the child.

Does the language readiness of human brains require that the richness of syntax and semantics be encoded in the genome, or is language one of those feats – from writing history to building cities to using computers – that played no role in biological evolution but rested on historical developments that created societies that could develop and transmit these skills? My hypothesis is that:

> Language readiness evolved as a multimodal manual/facial/vocal system with protosign (manual-based protolanguage) providing the scaffolding for protospeech (vocal-based protolanguage) to provide "neural critical mass" to allow language to emerge from protolanguage as a result of cultural innovations within the history of *Homo sapiens*.[2]

The theory summarized here makes it understandable why it is as easy for a deaf child to learn a signed language as it is for a hearing child to learn a spoken language.

MICHAEL ANTHONY ARBIB was born in England, grew up in Australia, and received his Ph.D. in Mathematics from MIT. After five years at Stanford, he became chairman of Computer and Information Science at the University of Massachusetts, Amherst, in 1970. He moved to the University of Southern California in 1986, where he is Professor of Computer Science, Neuroscience, Biomedical Engineering, Electrical Engineering, and Psychology. The author or editor of 38 books, Arbib recently co-edited *Who Needs Emotions? The Brain Meets the Robot* (Oxford University Press) with Jean-Marc Fellous. His current research focuses on brain mechanisms of visuomotor behavior, on neuroinformatics, and on the evolution of language.

 0140-525X/05 $12.50

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

1.2. The mirror system hypothesis

Humans, chimps and monkeys share a general physical form and a degree of manual dexterity, but their brains, bodies, and behaviors differ. Moreover, humans can and normally do acquire language, and monkeys and chimps cannot – though chimps and bonobos can be trained to acquire a form of communication that approximates the complexity of the utterances of a 2-year-old human infant. The approach offered here to the evolution of brain mechanisms that support language is anchored in two observations: (1) The system of the monkey brain for visuomotor control of hand movements for grasping has its premotor outpost in an area called F5 which contains a set of neurons, called *mirror neurons*, each of which is active not only when the monkey executes a specific grasp but also when the monkey observes a human or other monkey execute a more or less similar grasp (Rizzolatti et al. 1996a). Thus F5 in monkey contains a *mirror system for grasping* which employs a common neural code for *executed* and *observed* manual actions (sect. 3.2 provides more details). (2) The region of the human brain homologous to F5 is part of Broca's area, traditionally thought of as a speech area but which has been shown by brain imaging studies to be active when humans both execute and observe grasps.

These findings led to the mirror system hypothesis (Arbib & Rizzolatti 1997; Rizzolatti & Arbib 1998, henceforth R&A):

> The *parity requirement* for language in humans – that what counts for the speaker must count approximately the same for the hearer[3] – is met because Broca's area evolved atop the mirror system for grasping, with its capacity to generate and recognize a set of actions.

One of the contributions of this paper will be to stress that the F5 mirror neurons in the monkey are linked to regions of parietal and temporal cortex, and then argue that the evolutionary changes that "lifted" the F5 homologue of the common ancestor of human and monkey to yield the human Broca's area also "lifted" the other regions to yield Wernicke's area and other areas that support language in the human brain.

Many critics have dismissed the mirror system hypothesis, stating correctly that monkeys do not have language and so the mere possession of a mirror system for grasping cannot suffice for language. But the key phrase here is "evolved atop" – and Rizzolatti and Arbib (1998) discuss explicitly how changes in the primate brain might have adapted the use of the hands to support pantomime (intended communication) as well as praxis, and then outlined how further evolutionary changes could support language. The hypothesis provides a neurological basis for the oft-repeated claim that hominids had a (proto)language based primarily on manual gestures before they had a (proto)language based primarily on vocal gestures (e.g., Armstrong et al. 1995; Hewes 1973; Kimura 1993; Stokoe 2001).[4] It could be tempting to hypothesize that certain species-specific vocalizations of monkeys (such as the snake and leopard calls of vervet monkeys) provided the basis for the evolution of human speech, since both are in the vocal domain. However, these primate vocalizations appear to be related to non-cortical regions as well as the anterior cingulate cortex (see, e.g., Jürgens 1997) rather than F5, the homologue of Broca's area. I think it likely (though empirical data are sadly lacking) that the primate cortex contains a mirror system for such species-specific vocalizations, and that a related mirror system persists in humans, but I suggest that it is a complement to, rather than an integral part of, the speech system that includes Broca's area in humans.

The mirror system hypothesis claims that a *specific* mirror system – the primate mirror system for grasping – evolved into a 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 multimodal, its evolution being based on the execution and observation of hand movements. There is no claim that mirroring or imitation 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 (Doupe & Kuhl 1999).[5]

The monkey mirror system for grasping is presumed to allow other monkeys to understand praxic actions and use this understanding as a basis for cooperation, averting a threat, and so on. One might say that this is *implicitly* communicative, as a side effect of conducting an action for non-communicative goals. Similarly, the monkey's orofacial gestures register emotional state, and primate vocalizations can also communicate something of the current priorities of the monkey, but to a first order this might be called "involuntary communication"[6] – these "devices" evolved to signal certain aspects of the monkey's current internal state or situation either through its observable actions or through a fixed species-specific repertoire of facial and vocal gestures. I will develop the hypothesis that the mirror system made possible (but in no sense guaranteed) the evolution of the displacement of hand movements from praxis to gestures that can be controlled "voluntarily."

It is important to be quite clear as to what the mirror system hypothesis does *not* say.

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

2. Having a mirror system for grasping is not in itself sufficient for the copying of actions. It is one thing to recognize an action using the mirror system; it is another thing to use that representation as a basis for repeating the action. Hence, *further evolution of the brain was required for the mirror system for grasping to become an imitation system for grasping.*

3. It does not say that language evolution can be studied in isolation from cognitive evolution more generally.

Arbib (2002) modified and developed the R&A argument to hypothesize seven stages in the evolution of language, with imitation grounding two of the stages.[7] The first three stages are pre-hominid:

S1: Grasping.

S2: A mirror system for grasping shared with the common ancestor of human and monkey.

S3: A simple imitation system for object-directed grasping through much-repeated exposure. This is shared with the common ancestor of human and chimpanzee.

The next three stages then distinguish the hominid line from that of the great apes:

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

S4: A complex imitation system for grasping – the ability to recognize another's performance as a set of familiar actions and then repeat them, or to recognize that such a performance combines novel actions which can be approximated by variants of actions already in the repertoire.[8]

S5: *Protosign*, a manual-based communication system, breaking through the fixed repertoire of primate vocalizations to yield an open repertoire.

S6: *Protospeech*, resulting from the ability of control mechanisms evolved for protosign coming to control the vocal apparatus with increasing flexibility.[9]

The final stage is claimed (controversially!) to involve little if any biological evolution but instead to result from cultural evolution (historical change) in *Homo sapiens*:

S7: *Language*, the change from action-object frames to verb-argument structures to syntax and semantics; the coevolution of cognitive and linguistic complexity.

The Mirror System Hypothesis is simply the assertion that the mechanisms that get us to the role of Broca's area in language depend in a crucial way on the mechanisms established in stage S2. The above seven stages provide just one set of hypotheses on how this dependence may have arisen. The task of this paper is to re-examine this progression, responding to critiques by amplifying the supporting argument in some cases and tweaking the account in others. I believe that the overall framework is robust, but there are many details to be worked out and a continuing stream of new and relevant data and modeling to be taken into account.

The claim for the crucial role of manual communication in language evolution remains controversial. MacNeilage (1998; MacNeilage & Davis, in press b), for example, has argued that language evolved directly as speech. (A companion paper [Arbib 2005] details why I reject MacNeilage's argument. The basic point is to distinguish the evolution of the ability to use gestures that convey meaning from the evolution of syllabification as a way to structure vocal gestures.)

A note to commentators: The arguments for stages S1 through S6 can and should be evaluated quite independently of the claim that the transition to language was cultural rather than biological.

The neurolinguistic approach offered here is part of a performance approach which explicitly analyzes both perception and production (Fig. 1). For production, we have much we could possibly talk about which is 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 that is enriched by linkage to schemas for perceiving and acting upon the world (Arbib 2003; Rolls & Arbib 2003). Finally, the ideas in the semantic form must be expressed in words whose markings and ordering are expressed in phonological form – which may include a wide range of ordered expressive gestures, whether manual, orofacial, or vocal. For perception, the received sentence 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 nonlinguistic *action-object frame* in cognitive form. By contrast, the *verb-argument structure* is an overt linguistic representation in semantic form – in modern human languages, generally the action is named by a verb and the objects are named by nouns or noun phrases (see sect. 7). 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.

In the brain there may be no single grammar serving both production and perception, but rather, a "direct grammar" for production and an "inverse grammar" for perception. Jackendoff (2002) offers a competence theory with a much closer connection with theories of processing than has been common in generative linguistics and suggests (his sect. 9.3) strategies for a two-way dialogue between competence and performance theories. Jackendoff's approach to competence appears to be promising in this regard because it at-

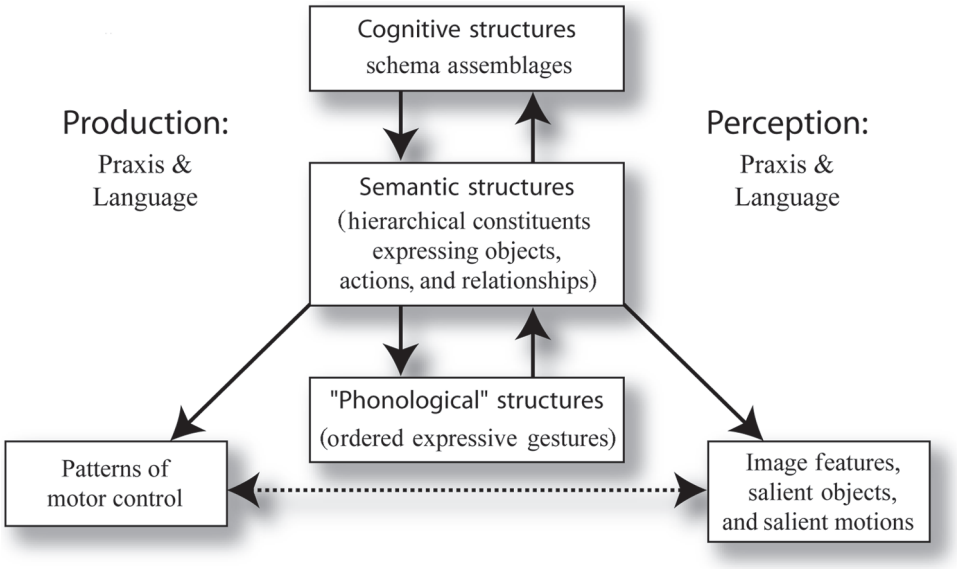

Figure 1. A performance view of the production and perception of language.

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

tends to the interaction of, for example, phonological, syntactic, and semantic representations. There is much, too, to be learned from a variety of approaches to cognitive grammar which relates cognitive form to syntactic structure (see, e.g., Heine 1997; Langacker 1987; 1991; Talmy 2000).

The next section provides a set of criteria for language readiness and further criteria for what must be added to yield language. It concludes (sect. 2.3) with an outline of the argument as it develops in the last six sections of the paper.

2. Language, protolanguage, and language readiness

I earlier defined a protolanguage as any system of utterances which served as a precursor to human language in the modern sense and hypothesized that the first *Homo sapiens* had protolanguage and a "language-ready brain" but did not have language.

Contra Bickerton (see Note 1), I will argue in section 7 that the prelanguage of *Homo erectus* and early *Homo sapiens* was composed mainly of "unitary utterances" that symbolized frequently occurring situations (in a general sense) without being decomposable into distinct words denoting components of the situation or their relationships. Words as we know them then co-evolved culturally with syntax through fractionation. In this view, many ways of expressing relationships that we now take for granted as part of language were the discovery of *Homo sapiens*; for example, adjectives and the fractionation of nouns from verbs may be "post-biological" in origin.

2.1. Criteria for language readiness

Here are properties hypothesized to support protolanguage:

LR1. *Complex imitation*: The ability to recognize another's performance as a set of familiar movements and then repeat them, but also to recognize that such a performance combines novel actions that can be approximated by (i.e., more or less crudely be imitated by) variants of actions already in the repertoire.[10]

The idea is that this capacity – distinct from the simple imitation system for object-directed grasping through much repeated exposure which is shared with chimpanzees – is necessary to support properties LR2 and LR3, including the idea that symbols are potentially arbitrary rather than innate:

LR2. *Symbolization*: The ability to associate symbols with an open class of episodes, objects, or actions.

At first, these symbols may have been unitary utterances, rather than words in the modern sense, and they may have been based on manual and facial gestures rather than being vocalized.

LR3. *Parity (mirror property)*: What counts for the speaker (or producer) must count for the listener (or receiver).

This extends Property LR2 by ensuring that symbols can be shared, and thus is bound up with LR4.

LR4. *Intended communication*: Communication is intended by the utterer to have a particular effect on the recipient rather than being involuntary or a side effect of praxis.

The remainder are more general properties, delimiting cognitive capabilities that underlie a number of the ideas which eventually find their expression in language:

LR5. *From hierarchical structuring to temporal ordering*: Perceiving that objects and actions have subparts; finding the appropriate timing of actions to achieve goals in relation to those hierarchically structured objects.

A basic property of language – translating a hierarchical conceptual structure into a temporally ordered structure of actions – is in fact not unique to language but is apparent whenever an animal takes in the nature of a visual scene and produces appropriate behavior. Animals possess subtle mechanisms of action-oriented perception with no necessary link to the ability to communicate about these components and their relationships. To have such structures does not entail the ability to communicate by using words or articulatory gestures (whether signed or vocalized) in a way that reflects these structures.

Hauser et al. (2002) assert that the faculty of language in the narrow sense (FLN) includes only recursion and is the one uniquely human component of the faculty of language. However, the flow diagram given by Byrne (2003) shows that the processing used by a mountain gorilla when preparing bundles of nettle leaves to eat is clearly recursive. Gorillas (like many other species, and not only mammals) have the working memory to refer their next action not only to sensory data but also to the state of execution of some current plan. Hence, when we refer to the monkey's grasping and ability to recognize similar grasps in others, it is a mistake to treat the individual grasps in isolation – the F5 system is part of a larger system that can direct those grasps as part of a recursively structured plan.

Let me simply list the next two properties here, and then expand upon them in the next section:

LR6. *Beyond the here-and-now 1*: The ability to recall past events or imagine future ones.

LR7. *Paedomorphy and sociality*: Paedomorphy is the prolonged period of infant dependency which is especially pronounced in humans; this combines with social structures for caregiving to provide the conditions for complex social learning.

Where Deacon (1997) makes symbolization central to his account of the coevolution of language and the human brain, the present account will stress the parity property LR3, since it underlies the sharing of meaning, and the capacity for complex imitation. I will also argue that only protolanguage co-evolved with the brain, and that the full development of linguistic complexity was a cultural/historical process that required little or no further change from the brains of early *Homo sapiens*.

Later sections will place LR1 through LR7 in an evolutionary context (see sect. 2.3 for a summary), showing how the coupling of complex imitation to complex communication creates a language-ready brain.

2.2. Criteria for language

I next present four criteria for what must be added to the brain's capabilities for the parity, hierarchical structuring, and temporal ordering of language readiness to yield *language*. Nothing in this list rests on the medium of exchange of the language, applying to spoken language, sign language, or written language, for example. My claim is that a brain that can support properties LR1 through LR7 above can support properties LA1 through LA4 below – as long

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

as its "owner" matures in a society that possesses language in the sense so defined and nurtures the child to acquire it. In other words, I claim that the mechanisms that make LR1 through LR7 possible are supported by the genetic encoding of brain and body and the consequent space of possible social interactions, but that the genome has no additional structures specific to LA1 through LA4. In particular, the genome does not have special features encoding syntax and its linkage to a compositional semantics.[11]

I suggest that "true language" involves the following further properties beyond LR1 through LR7:

LA1. *Symbolization and compositionality*: The symbols become words in the modern sense, interchangeable and composable in the expression of meaning.[12]

LA2. *Syntax, semantics and recursion*: The matching of syntactic to semantic structures coevolves with the fractionation of utterances, with the nesting of substructures making some form of recursion inevitable.

LA1 and LA2 are intertwined. Section 7 will offer candidates for the sorts of discoveries that may have led to progress from "unitary utterances" to more or less structured assemblages of words. Given the view (LR5) that recursion of action (but not of communication) is part of language readiness, the key transition here is the compositionality that allows cognitive structure to be reflected in symbolic structure (the transition from LR2 to LA1), as when perception (not uniquely human) grounds linguistic description (uniquely human) so that, for example, the noun phrase (NP) describing a part of an object may optionally form part of the NP describing the overall object. From this point of view, recursion in language is a corollary of the essentially recursive nature of action and perception *once symbolization becomes compositional, and reflects addition of further detail to, for example, a description when needed to reduce ambiguity in communication*.

The last two principles provide the linguistic complements of two of the conditions for language readiness, LR6 (*Beyond the here-and-now 1*) and LR7 (*Paedomorphy and sociality*), respectively.

LA3. *Beyond the here-and-now 2*: Verb tenses or other circumlocutions express the ability to recall past events or imagine future ones.

There are so many linguistic devices for going beyond the here and now, and beyond the factual, that verb tenses are mentioned to stand in for all the devices languages have developed to communicate about other "possible worlds" that are far removed from the immediacy of, say, the vervet monkey's leopard call.

If one took a human language and removed all reference to time, one might still want to call it a language rather than a protolanguage, even though one would agree that it was thereby greatly impoverished. Similarly, the number system of a language can be seen as a useful, but not definitive, "plug-in." LA3 nonetheless suggests that the ability to talk about past and future is a central part of human languages as we understand them. However, all this would be meaningless (literally) without the underlying cognitive machinery – the substrate for episodic memory provided by the hippocampus (Burgess et al. 1999) and the substrate for planning provided by frontal cortex (Passingham 1993, Ch. 10). It is not part of the mirror system hypothesis to explain the evolution of the brain structures that support LR6; it is an exciting challenge for work "beyond the mirror" to show how such structures could provide the basis for humans to discover the capacities for communication summarized in LA3.

LA4. *Learnability*: To qualify as a human language, much of the syntax and semantics of a human language must be learnable by most human children.

I say "much of" because it is not true that children master all the vocabulary or syntactic subtlety of a language by 5 or 7 years of age. Language acquisition is a process that continues well into the teens as we learn more subtle syntactic expressions and a greater vocabulary to which to apply them (C. Chomsky [1969] traces the changes that occur from ages 5 to 10), allowing us to achieve a richer and richer set of communicative and representational goals.

LR7 and LA4 link a biological condition "orthogonal" to the mirror system hypothesis with a "supplementary" property of human languages. This supplementary property is that languages do not simply exist – they are acquired anew (and may be slightly modified thereby) in each generation (LA4). The biological property is an inherently social one about the nature of the relationship between parent (or other caregiver) and child (LR7) – the prolonged period of infant dependency which is especially pronounced in humans has co-evolved with the social structures for caregiving that provide the conditions for the complex social learning that makes possible the richness of human cultures in general and of human languages in particular (Tomasello 1999b).

2.3. The argument in perspective

The argument unfolds in the remaining six sections as follows:

Section 3. Perspectives on grasping and mirror neurons: This section presents two models of the macaque brain. A key point is that the functions of mirror neurons reflect the impact of experience rather than being pre-wired.

Section 4. Imitation: This section presents the distinction between simple and complex imitation systems for grasping, and argues that monkeys have neither, that chimpanzees have only simple imitation, and that the capacity for complex imitation involved hominid evolution since the separation from our common ancestors, the great apes, including chimpanzees.

Section 5. From imitation to protosign: This section examines the relation between symbolism, intended communication, and parity, and looks at the multiple roles of the mirror system in supporting pantomime and then conventionalized gestures that support a far greater range of intended communication.

Section 6. The emergence of protospeech: This section argues that evolution did not proceed directly from monkey-like primate vocalizations to speech but rather proceeded from vocalization to manual gesture and back to vocalization again.

Section 7. The inventions of languages: This section argues that the transition from action-object frames to verb-argument structures embedded in larger sentences structured by syntax and endowed with a compositional semantics was the effect of the accumulation of a wide range of human discoveries that had little if any impact on the human genome.

Section 8. Toward a neurolinguistics "beyond the mirror": This section extracts a framework for action-oriented linguistics informed by our analysis of the "extended mirror

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

Table 1. *A comparative view of how the following sections relate the criteria LR1–LR for language readiness and LA1–LA2 for language (middle column) to the seven stages, S1–S7, of the extended mirror system hypothesis (right column)*

Section	Criteria	Stages
2.1	LR5: From hierarchical structuring to temporal ordering	This precedes the evolutionary stages charted here.
3.1		S1: Grasping The FARS model.
3.2		S2: Mirror system for grasping Modeling Development of the Mirror System. This supports the conclusion that mirror neurons can be recruited to recognize and encode an expanding set of novel actions.
4	LR1: Complex imitation	S3: Simple imitation This involves properties of the mirror system beyond the monkey's data. S4: Complex imitation This is argued to distinguish humans from other primates.
5	LR2: Symbolization LR4: Intended communication LR3: Parity (mirror property)	S5: Protosign The transition of complex imitation from praxic to communicative use involves two substages: S5a: the ability to engage in pantomime; S5b: the ability to make conventional gestures to disambiguate pantomime.
6.1		S6: Protospeech It is argued that early protosign provided the scaffolding for early protospeech, after which both developed in an expanding spiral until protospeech became dominant for most people.
7	LA1: Symbolization and compositionality LA2: Syntax, semantics, and recursion	S7: Language The transition from action-object frame to verb-argument structure to syntax and semantics.
8		The evolutionary developments of the preceding sections are restructured into synchronic form to provide a framework for further research in neurolinguistics relating the capabilities of the human brain for language, action recognition, and imitation.

system hypothesis" presented in the previous sections. The language-ready brain contains the evolved mirror system as a key component but also includes many other components that lie outside, though they interact with, the mirror system.

Table 1 shows how these sections relate the evolutionary stages S1 through S7, and their substages, to the above criteria for language readiness and language.[13]

3. Perspectives on grasping and mirror neurons

Mirror neurons in F5, which are active both when the monkey performs certain actions and when the monkey observes them performed by others, are to be distinguished from *canonical neurons* in F5, which are active when the monkey performs certain actions but not when the monkey observes actions performed by others. More subtly, canonical neurons fire when they are presented with a graspable object, irrespective of whether the monkey performs the grasp or not – but clearly this must depend on the extra (inferred) condition that the monkey not only sees the object but is aware, in some sense, that it is possible to grasp it. Were it not for the caveat, canonical neurons would also fire when the monkey observed the object being grasped by another.

The "classic" mirror system hypothesis (sect. 1.2) emphasizes the grasp-related neurons of the monkey premotor area F5 and the homology of this region with human Broca's area. However, Broca's area is part of a larger system supporting language, and so we need to enrich the mirror system hypothesis by seeing how the mirror system for grasping in monkey includes a variety of brain regions in addition to F5. I show this by presenting data and models that locate the canonical system of F5 in a systems perspective (the FARS model of sect. 3.1) and then place the mirror system of F5 in a system perspective (the MNS model of sect. 3.2).

3.1. The FARS model

Given our concern with hand use and language, it is striking that the ability to use the size of an object to preshape the hand while grasping it can be dissociated by brain lesions from the ability to consciously recognize and describe that size. Goodale et al. (1991) studied a patient (D.F.) whose cortical damage allowed signals to flow from primary

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

visual cortex (V1) towards posterior parietal cortex (PP) but not from V1 to inferotemporal cortex (IT). When asked to indicate the width of a single block by means of her index finger and thumb, D.F.'s 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, D.F. 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. Jeannerod et al. (1994) reported a study of impairment of grasping in a patient (A.T.) with a bilateral posterior parietal lesion of vascular origin that left IT and the pathway V1 → IT relatively intact, but grossly impaired the pathway V1 → PP. This patient can reach without deficit toward the location of such an object, but cannot preshape appropriately when asked to grasp it.

A corresponding distinction in the role of these pathways in the monkey is crucial to the FARS model (named for Fagg, Arbib, Rizzolatti, and Sakata; see Fagg & Arbib 1998), which embeds F5 canonical neurons in a larger system. Taira et al. (1990) found that anterior intraparietal (AIP) cells (in the anterior intraparietal sulcus of the parietal cortex) extract neural codes for *affordances* for grasping from the visual stream and sends these on to area F5. Affordances (Gibson 1979) are features of the object relevant to action, in this case to grasping, rather than aspects of identifying the object's identity. Turning to human data: Ehrsson et al. (2003) compared the brain activity when humans attempted to lift an immovable test object held between the tips of the right index finger and thumb with the brain activity obtained in two control tasks in which neither the load force task nor the grip force task involved coordinated grip-load forces. They found that the grip-load force task was specifically associated with activation of a section of the right intraparietal cortex. Culham et al. (2003) found greater activity for grasping than for reaching in several regions, including the anterior intraparietal (AIP) cortex. Although the lateral occipital complex (LOC), a ventral stream area believed to play a critical role in object recognition, was activated by the objects presented on both grasping and reaching trials, there was no greater activity for grasping compared to reaching.

The FARS model analyzes how the "canonical system," centered on the AIP → F5 pathway, may account for basic phenomena of grasping. The highlights of the model are shown in Figure 2,[14] which diagrams the crucial role of IT (inferotemporal cortex) and PFC (prefrontal cortex) in modulating F5's selection of an affordance. The *dorsal stream* (from V1 to parietal cortex) carries 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 are then passed on to F5. 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 V1 to IT), by contrast, is able to recognize what the object is. This information is passed to PFC, which can then, on the basis of the current goals of the organism and the recognition of the nature of the object, bias AIP to choose the affordance appropriate to the task at hand. The original FARS model posited connections between PFC and F5. However, there is evidence (reviewed by Rizzolatti & Luppino 2001) that these connections are very limited, whereas rich connections exist between PFC and AIP. Rizzolatti and Luppino (2003) therefore suggested that FARS

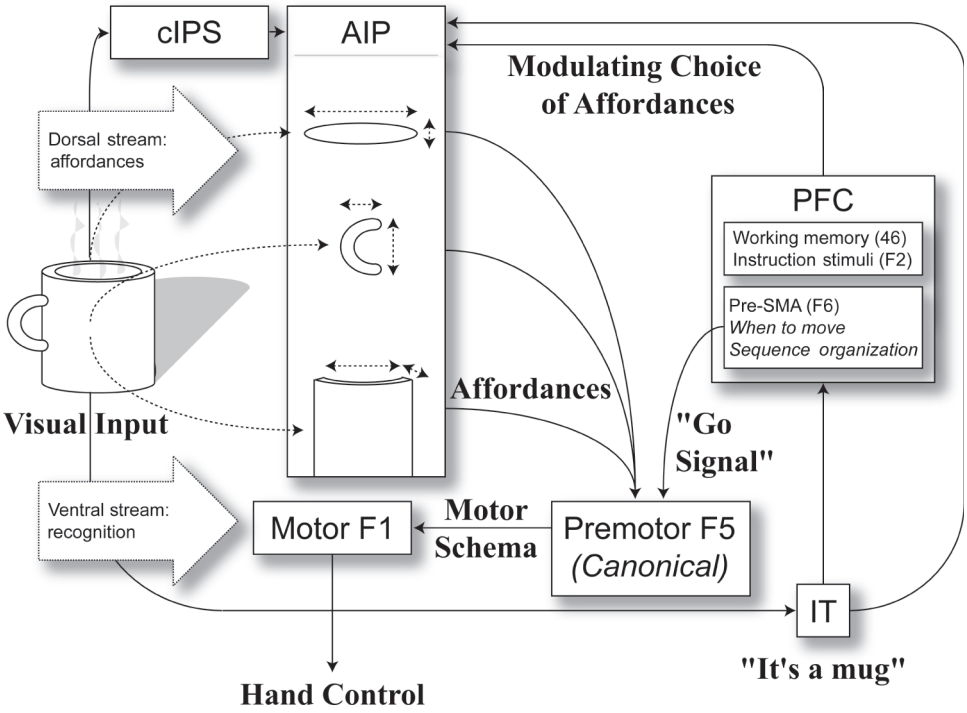

Figure 2. A reconceptualization of the FARS model in which the primary influence of PFC (prefrontal cortex) on the selection of affordances is on parietal cortex (AIP, anterior intraparietal sulcus) rather than premotor cortex (the hand area F5).

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

be modified so that information on object semantics and the goals of the individual influence AIP rather than F5 neurons. I show the modified schematic in Figure 2. The modified figure represents the way in which AIP 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. In other words, AIP provides cues on how to interact with an object, leaving it to IT to categorize the object or determine its identity.

Although the data on cell specificity in F5 and AIP emphasize single actions, these actions are normally part of more complex behaviors – to take a simple example, a monkey who grasps a raisin will, in general, then proceed to eat it. Moreover, a particular action might be part of many learned sequences, and so we do not expect the premotor neurons for one action to prime a single possible consequent action and hence must reject "hard wiring" of the sequence. The generally adopted solution is to segregate the learning of a sequence from the circuitry which encodes the unit actions, the latter being F5 in the present study. Instead, another area (possibly the part of the supplementary motor area called pre-SMA; Rizzolatti et al. 1998) has neurons whose connections encode an "abstract sequence" Q1, Q2, Q3, Q4, with sequence learning then involving learning that the activation of Q1 triggers the F5 neurons for A, Q2 triggers B, Q3 triggers A again, and Q4 triggers C to provide encoding of the sequence A-B-A-C. Other studies suggest that administration of the sequence (inhibiting extraneous actions, while priming imminent actions) is carried out by the basal ganglia on the basis of its interactions with the pre-SMA (Bischoff-Grethe et al. 2003; see Dominey et al. 1995 for an earlier model of the possible role of the basal ganglia in sequence learning).

3.2. Modeling development of the mirror system

The populations of canonical and mirror neurons appear to be spatially segregated in F5 (Rizzolatti & Luppino 2001). Both sectors receive a strong input from the secondary somatosensory area (SII) and parietal area PF. In addition, canonical neurons are the selective target of area AIP. Perrett et al. (1990; cf. Carey et al. 1997) found that STSa, in the rostral part of the superior temporal sulcus (STS), has neurons which discharge when the monkey observes such biological actions as walking, turning the head, bending the torso, and moving the arms. Of most relevance to us is that a few of these neurons discharged when the monkey observed goal-directed hand movements, such as grasping objects (Perrett et al. 1990) – though STSa neurons do not seem to discharge during movement execution as distinct from observation. STSa and F5 may be indirectly connected via the inferior parietal area PF (Brodmann area 7b) (Cavada & Goldman-Rakic 1989; Matelli et al. 1986; Petrides & Pandya 1984; Seltzer & Pandya 1994). About 40% of the visually responsive neurons in PF are active for observation of actions such as holding, placing, reaching, grasping, and bimanual interaction. Moreover, most of these action-observation neurons were also active during the execution of actions similar to those for which they were "observers," and were therefore called PF mirror neurons (Fogassi et al. 1998).

In summary, area F5 and area PF include an observation/execution matching system: When the monkey observes an action that resembles one in its movement repertoire, a subset of the F5 and PF mirror neurons is activated which also discharges when a similar action is executed by the monkey itself.

I next develop the conceptual framework for thinking about the relation between F5, AIP, and PF. Section 6.1 expands the mirror neuron database, reviewing the reports by Kohler et al. (2002) of a subset of mirror neurons responsive to sounds and by Ferrari et al. (2003) of neurons responsive to the observation of orofacial communicative gestures.

Figure 3 provides a glimpse of the schemas (functions) involved in the MNS model (Oztop & Arbib 2002) of the monkey mirror system.[15] First, we look at those elements involved when the monkey itself reaches for an object. Areas IT and cIPS (caudal intraparietal sulcus; part of area 7) provide visual input concerning the nature of the observed object and the position and orientation of the object's surfaces, respectively, to AIP. The job of AIP is then to extract the affordances the object offers for grasping. The upper diagonal in Figure 3 corresponds to the basic pathway AIP → $F5_{canonical}$ → M1 (primary motor cortex) of the FARS model, but Figure 3 does not include the important role of PFC in action selection. The lower-right diagonal (MIP/LIP/VIP → F4) completes the "canonical" portion of the MNS model, since motor cortex must instruct not only the hand muscles how to grasp but also (via various intermediaries) the arm muscles how to reach, transporting the hand to the object. The rest of Figure 3 presents the core elements for the understanding of the mirror system. Mirror neurons do not fire when the monkey sees the hand movement or the object in isolation – it is the sight of the hand moving appropriately to grasp or otherwise manipulate a seen (or recently seen) object (Umiltá et al. 2001) that is required for the mirror neurons attuned to the given action to fire. This requires schemas for the recognition of both the shape of the hand and analysis of its motion (ascribed in the figure to STS), and for analysis of the relation of these hand parameters to the location and affordance of the object (7a and 7b; we identify 7b with PF).

In the MNS model, the *hand state* was accordingly defined as a vector whose components represented the movement of the wrist relative to the location of the object and of the hand shape relative to the affordances of the object. Oztop and Arbib (2002) showed that an artificial neural network corresponding to PF and $F5_{mirror}$ could be trained to recognize the grasp type from the *hand state trajectory*, with correct classification often being achieved well before the hand reached the object. The modeling assumed that the neural equivalent of a grasp being in the monkey's repertoire is that there is a pattern of activity in the F5 canonical neurons which commands that grasp. During training, the output of the F5 canonical neurons, acting as a code for the grasp being executed by the monkey at that time, was used as the training signal for the F5 mirror neurons to enable them to learn which hand-object trajectories corresponded to the canonically encoded grasps. Moreover, the input to the F5 mirror neurons encodes the trajectory of the relation of parts of the hand to the object rather than the visual appearance of the hand in the visual field. As a result of this training, the appropriate mirror neurons come to fire in response to viewing the appropriate trajectories even when the trajectory is not accompanied by F5 canonical firing.

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

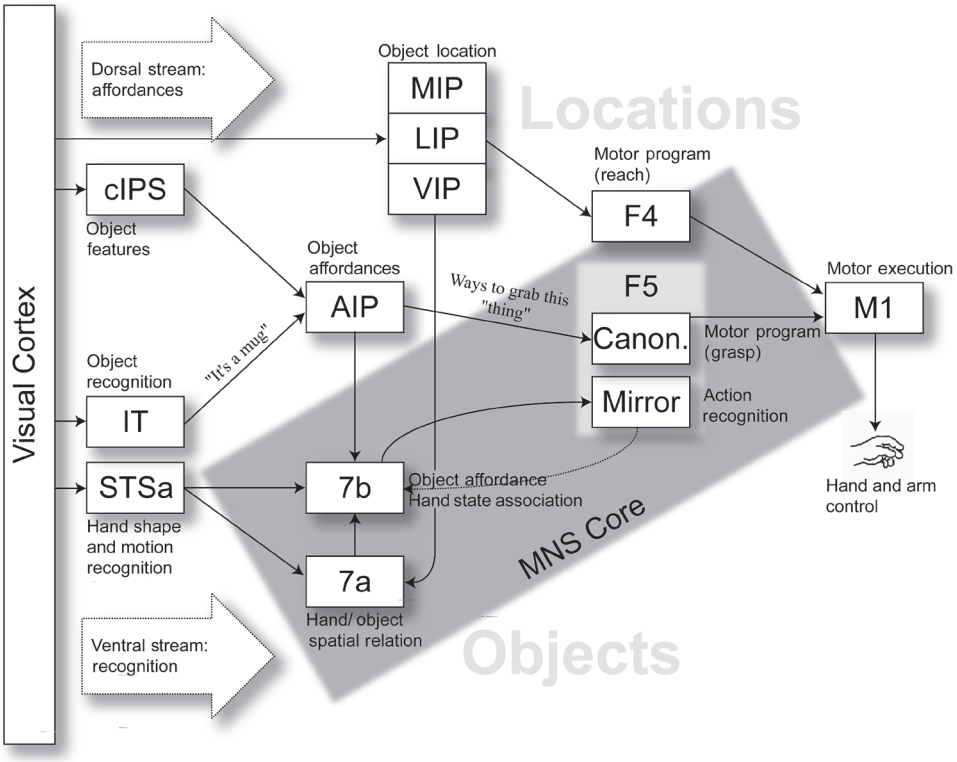

Figure 3. A schematic view of the Mirror Neuron System (MNS) model (Oztop & Arbib 2002).

This training prepares the F5 mirror neurons to respond to hand-object relational trajectories even when the hand is of the "other" rather than the "self," because the hand state is based on the movement of a hand relative to the object, and thus only *indirectly* on the retinal input of seeing hand and object which can differ greatly between observation of self and other. What makes the modeling worthwhile is that the trained network not only responded to hand-state trajectories from the training set, but also exhibited interesting responses to novel hand-object relationships. Despite the use of a non-physiological neural network, simulations with the model revealed a range of putative properties of mirror neurons that suggest new neurophysiological experiments. (See Oztop & Arbib [2002] for examples and detailed analysis.)

Although MNS was constructed as a model of the development of mirror neurons in the monkey, it serves equally well as a model of the development of mirror neurons in the human infant. A major theme for future modeling, then, will be to clarify which aspects of human development are generic for primates and which are specific to the human repertoire. In any case, the MNS model makes the crucial assumption that the grasps that the mirror system comes to recognize are already in the (monkey or human) infant's repertoire. But this raises the question of how grasps entered the repertoire. To simplify somewhat, the answer has two parts: (1) Children explore their environment, and as their initially inept arm and hand movements successfully contact objects, they learn to reproduce the successful grasps reliably, with the repertoire being tuned through further experience. (2) With more or less help from caregivers, infants come to recognize certain novel actions in terms of similarities with and differences from movements already in their repertoires, and on this basis learn to produce some version of these novel actions for themselves. Our Infant Learning to Grasp Model (ILGM; Oztop et al. 2004) strongly supports the hypothesis that grasps are acquired through experience as the infant learns how to conform the biomechanics of its hand to the shapes of the objects it encounters. However, limited space precludes presentation of this model here.

The classic papers on the mirror system for grasping in the monkey focus on a repertoire of grasps – such as the precision pinch and power grasp – that seem so basic that it is tempting to think of them as prewired. The crucial point of this section on modeling is that learning models such as ILGM and MNS, and the data they address, make clear that *mirror neurons are not restricted to recognition of an innate set of actions but can be recruited to recognize and encode an expanding repertoire of novel actions*. I will relate the FARS and MNS models to the development of imitation at the end of section 4.

With this, let us turn to human data. We mentioned in section 1.2 that Broca's area, traditionally thought of as a speech area, has been shown by brain imaging studies to be active when humans both execute and observe grasps. This was first tested by two positron emission tomography (PET) experiments (Grafton et al. 1996; Rizzolatti et al. 1996) which compared brain activation when subjects observed the experimenter grasping an object against activation when subjects simply observed the object. Grasp observation significantly activated the superior temporal sulcus (STS), the inferior parietal lobule, and the inferior frontal gyrus (area 45). All activations were in the left hemisphere. The last area is of especial interest because areas 44 and 45 in the left hemisphere of the human constitute Broca's area. Such data certainly contribute to the growing body of indirect evidence that there is a mirror system for grasping that

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

links Broca's area with regions in the inferior parietal lobule and STS. We have seen that the "minimal mirror system" for grasping in the macaque includes mirror neurons in the parietal area PF (7b) as well as F5, and some not-quite-mirror neurons in the region STSa in the superior temporal sulcus. Hence, in further investigation of the mirror system hypothesis it will be crucial to extend the F5 → Broca's area homology to examine the human homologues of PF and STSa as well. I will return to this issue in section 7 (see Fig. 6) and briefly review some of the relevant data from the rich and rapidly growing literature based on human brain imaging and transcranial magnetic stimulation (TMS) inspired by the effort to probe the human mirror system and relate it to action recognition, imitation, and language.

Returning to the term "language readiness," let me stress that the reliable linkage of brain areas to different aspects of language in normal speaking humans does not imply that language per se is "genetically encoded" in these regions. There is a neurology of writing even though writing was invented only a few thousand years ago. The claim is not that Broca's area, Wernicke's area, and STS are genetically preprogrammed for language, but rather that the development of a human child in a language community normally adapts these brain regions to play a crucial (but not the only) role in language performance.

4. Imitation

We have already discussed the mirror system for grasping as something shared between macaque and human; hence the hypothesis that this set of mechanisms was already in place in the common ancestor of monkey and human some 20 million years ago.[16] In this section we move from stage S2, a mirror system for grasping, to stages S3, a simple imitation system for grasping, and S4, a complex imitation system for grasping. I will argue that chimpanzees possess a capability for *simple* imitation that monkeys lack, but that humans have *complex* imitation whereas other primates do not. The ability to copy *single* actions is just the first step towards complex imitation, which involves parsing a complex movement into more or less familiar pieces and then performing the corresponding composite of (variations on) familiar actions. Arbib and Rizzolatti (1997) asserted that what makes a movement into an action is that it is associated with a goal, and that initiation of the movement is accompanied by the creation of an expectation that the goal will be met. Hence, it is worth stressing that when I speak of imitation here, I speak of the imitation of a movement and its linkage to the goals it is meant to achieve. The action may thus vary from occasion to occasion depending on parametric variations in the goal. This is demonstrated by Byrne's (2003) description, noted earlier, of a mountain gorilla preparing bundles of nettle leaves to eat.

Visalberghi and Fragaszy (2002) review data on attempts to observe imitation in monkeys, including their own studies of capuchin monkeys. They stress the huge difference between the major role that imitation plays in learning by human children, and the very limited role, if any, that imitation plays in social learning in monkeys. There is little evidence for vocal imitation in monkeys or apes (Hauser 1996), but it is generally accepted that chimpanzees are capable of some forms of imitation (Tomasello & Call 1997).

There is not space here to analyze all the relevant distinctions between imitation and other forms of learning, but one example may clarify my view: Voelkl and Huber (2000) had marmosets observe a demonstrator removing the lids from a series of plastic canisters to obtain a mealworm. When subsequently allowed access to the canisters, marmosets that observed a demonstrator using its hands to remove the lids used only their hands. In contrast, marmosets that observed a demonstrator using its mouth also used their mouths to remove the lids. Voelkl and Huber (2000) suggest that this may be a case of true imitation in marmosets, but I would argue that it is a case of *stimulus enhancement*, apparent imitation resulting from directing attention to a particular object or part of the body or environment. This is to be distinguished from *emulation* (observing and attempting to reproduce results of another's actions without paying attention to details of the other's behavior) and *true imitation* which involves copying a novel, otherwise improbable action or some act that is outside the imitator's prior repertoire.

Myowa-Yamakoshi and Matsuzawa (1999) observed in a laboratory setting that chimpanzees typically took 12 trials to learn to "imitate" a behavior and in doing so paid more attention to where the manipulated object was being directed than to the actual movements of the demonstrator. This involves the ability to learn novel actions which may require using one or both hands to bring two objects into relationship, or to bring an object into relationship with the body.

Chimpanzees do use and make tools in the wild, with different tool traditions found in geographically separated groups of chimpanzees: Boesch and Boesch (1983) have observed chimpanzees in Tai National Park, Ivory Coast, using stone tools to crack nuts open, although Goodall has never seen chimpanzees do this in the Gombe in Tanzania. 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.[17] The nut-cracking technique is not mastered until adulthood. Tomasello (1999b) comments that, over many years of observation, Boesch observed only two possible instances in which the mother *appeared* to be actively attempting to instruct her child, and that even in these cases it is unclear whether the mother had the goal of helping the young chimp learn to use the tool. We may contrast the long and laborious process of acquiring the nut-cracking technique with the rapidity with which human adults can acquire novel sequences, and the crucial role of caregivers in the development of this capacity for complex imitation. Meanwhile, reports abound of imitation in many species, including dolphins and orangutans, and even tool use in crows (Hunt & Gray 2002). Consequently, I accept that the demarcation between the capability for imitation of humans and nonhumans is problematic. Nonetheless, I still think it is fair to claim that humans can master feats of imitation beyond those possible for other primates.

The ability to imitate has clear adaptive advantage in allowing creatures to transfer skills to their offspring, and therefore could be selected for quite independently of any adaptation related to the later emergence of protolanguage. By the same token, the ability for complex imitation could provide further selective advantage unrelated to language. However, complex imitation is central to human infants

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

both in their increasing mastery of the physical and social world and in the close coupling of this mastery to the acquisition of language (cf. Donald 1998; Arbib et al., in press). The child must go beyond simple imitation to acquire the phonological repertoire, words, and basic "assembly skills" of its language community, and this is one of the ways in which brain mechanisms supporting imitation were crucial to the emergence of language-ready *Homo sapiens*. If I then assume (1) that the common ancestor of monkeys and apes had no greater imitative ability than present-day monkeys (who possess, I suggest, stimulus enhancement rather than simple imitation), and (2) that the ability for simple imitation shared by chimps and humans was also possessed by their common ancestor, but (3) that only humans possess a talent for "complex" imitation, then I have established a case for the hypothesis that extension of the mirror system from *recognizing* single actions to *being able to copy* compound actions was the key innovation in the brains of our hominid ancestors that was relevant to language. And, more specifically, we have the hypotheses:

Stage S3 hypothesis: Brain mechanisms supporting a simple imitation system – imitation of short, novel sequences of object-directed actions through repeated exposure – for grasping developed in the 15-million-year evolution from the common ancestor of monkeys and apes to the common ancestor of apes and humans; and

Stage S4 hypothesis: Brain mechanisms supporting a complex imitation system – acquiring (longer) novel sequences of more abstract actions in a single trial – developed in the 5-million-year evolution from the common ancestor of apes and humans along the hominid line that led, in particular, to *Homo sapiens*.[18]

Now that we have introduced imitation, we can put the models of section 3.2 in perspective by postulating the following stages prior to, during, and building on the development of the mirror system for grasping in the infant:

A. The child refines a crude map (superior colliculus) to make unstructured reach and "swipe" movements at objects; the grasp reflex occasionally yields a successful grasp.

B. The child develops a set of grasps which succeed by kinesthetic, somatosensory criteria (ILGM).

C. AIP develops as affordances of objects become learned in association with successful grasps. Grasping becomes visually guided; the grasp reflex disappears.

D. The (grasp) mirror neuron system develops driven by visual stimuli relating hand and object generated by the actions (grasps) performed by the infant himself (MNS).

E. The child gains the ability to map other individual's actions into his internal motor representation.

F. Then the child acquires the ability to imitate, creating (internal) representations for novel actions that have been observed and developing an action prediction capability.

I suggest that stages A through D are much the same in monkey and human, but that stages E and F are rudimentary at best in monkeys, somewhat developed in chimps, and well-developed in human children (but not in infants). In terms of Figure 3, we might say that if MNS were augmented to have a population of mirror neurons that could acquire population codes for observed actions not yet in the repertoire of self-actions, then in stage E the mirror neurons would provide training for the canonical neurons, reversing the information flow seen in the MNS model. Note that this raises the further possibility that the human infant may come to recognize movements that not only are not within the repertoire but which never come to be within the repertoire. In this case, the cumulative development of action recognition may proceed to increase the breadth and subtlety of the range of actions that are recognizable but cannot be performed by children.

5. From imitation to protosign

The next posited transition, from stage S4, a complex imitation system for grasping, to stage S5, protosign, a manual-based communication system, takes us from imitation for the sake of instrumental goals to imitation for the sake of communication. Each stage builds on, yet is not simply reducible to, the previous stage.

I argue that the combination of the abilities (S5a) to engage in pantomime and (S5b) to make conventional gestures to disambiguate pantomime yielded a brain which could (S5) support "protosign," a manual-based communication system that broke through the fixed repertoire of primate vocalizations to yield an open repertoire of communicative gestures.

It is important to stress that communication is about far more than grasping. To pantomime the flight of a bird, you might move your hand up and down in a way that indicates the flapping of a wing. Your pantomime uses movements of the hand (and arm and body) to imitate movement other than hand movements. You can pantomime an object either by miming a typical action by or with the object, or by tracing out the characteristic shape of the object.

The transition to pantomime does seem to involve a genuine neurological change. Mirror neurons for grasping in the monkey will fire only if the monkey sees *both* the hand movement and the object to which it is directed (Umiltá et al. 2001). A grasping movement that is not made in the presence of a suitable object, or is not directed toward that object, will not elicit mirror neuron firing. By contrast, in pantomime, the observer sees the movement in isolation and *infers* (1) what non-hand movement is being mimicked by the hand movement, and (2) the goal or object of the action. This is an evolutionary change of key relevance to language readiness. Imitation is the generic attempt to reproduce movements performed by another, whether to master a skill or simply as part of a social interaction. By contrast, pantomime is performed with the intention of getting the observer to think of a specific action, object, or event. It is essentially communicative in its nature. The imitator observes; the pantomimic intends to be observed.

As Stokoe (2001) and others emphasize, the power of pantomime is that it provides open-ended communication that works without prior instruction or convention. However (and I shall return to this issue at the end of this section), even signs of modern signed language which resemble pantomimes are conventionalized and are, thus, distinct from pantomimes. Pantomime per se is not a form of protolanguage; rather it provides a rich scaffolding for the emergence of protosign.

All this assumes rather than provides an explanation for LR4, the transition from making praxic movement – for example, those involved in the immediate satisfaction of some appetitive or aversive goal – to those intended by the utterer to have a particular effect on the recipient. I tentatively offer:

The intended communication hypothesis: The ability to

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

imitate combines with the ability to observe the effect of such imitation on conspecifics to support a migration of closed species-specific gestures supported by other brain regions to become the core of an open class of communicative gestures.

Darwin (1872/1965) observed long ago, across a far wider range of mammalian species than just the primates, that the facial expressions of conspecifics provide valuable cues to their likely reaction to certain courses of behavior (a rich complex summarized as "emotional state"). Moreover, the F5 region contains orofacial cells as well as manual cells. This suggests a progression from control of emotional expression by systems that *exclude* F5 to the extension of F5's mirror capacity for orofacial as well as manual movement (discussed below), via its posited capacity (achieved by stage S3) for simple imitation, to support the imitation of emotional expressions. This would then provide the ability to affect the behavior of others by, for example, *appearing* angry. This would in turn provide the evolutionary opportunity to generalize the ability of F5 activity to affect the behavior of conspecifics from species-specific vocalizations to a general ability to use the imitation of behavior (as distinct from praxic behavior itself) as a means to influence others. This in turn makes possible reciprocity by a process of backward chaining where the influence is not so much on the praxis of the other as on the exchange of information. With this, the transition described by LR4 (intended communication) has been achieved in tandem with the achievement and increasing sophistication of LR2 (symbolization).

A further critical change (labeled 5b above) emerges from the fact that in pantomime it might be hard to distinguish, for example, a movement signifying "bird" from one meaning "flying." This inability to adequately convey shades of meaning using "natural" pantomime would favor the invention of gestures that could in some way disambiguate which of their associated meanings was intended. Note that whereas a pantomime can freely use any movement that might evoke the intended observation in the mind of the observer, a disambiguating gesture must be conventionalized.[19] This use of non-pantomimic gestures requires extending the use of the mirror system to attend to an entirely new class of hand movements. However, this does not seem to require a biological change beyond that limned above for pantomime.

As pantomime begins to use hand movements to mime different degrees of freedom (as in miming the flying of a bird), a dissociation begins to emerge. The mirror system for the pantomime (based on movements of face, hand, etc.) is now different from the recognition system for the action that is pantomimed, and – as in the case of flying – the action may not even be in the human action repertoire. However, the system is still able to exploit the praxic recognition system because an animal or hominid must observe much about the environment that is relevant to its actions but is not in its own action repertoire. Nonetheless, this dissociation now underwrites the emergence of protosign – an open system of actions that are defined only by their communicative impact, not by their direct relation to praxic goals.

Protosign may lose the ability of the original pantomime to elicit a response from someone who has not seen it before. However, the price is worth paying in that the simplified form, once agreed upon by the community, allows more rapid communication with less neural effort. One may see analogies in the history of Chinese characters. The character 山 (san) may not seem particularly pictorial, but if (following the "etymology" of Vaccari & Vaccari 1961), we see it as a simplification of a picture of three mountains, 𡴖, via such intermediate forms as 凵, then we have no trouble seeing the simplified character 山 as meaning "mountain."[20] The important point here for our hypothesis is that although such a "picture history" may provide a valuable crutch to some learners, with sufficient practice the crutch is thrown away, and in normal reading and writing, the link between 山 and its meaning is direct, with no need to invoke an intermediate representation of 𡴖.

In the same way, I suggest that pantomime is a valuable crutch for acquiring a modern sign language, but that even signs which resemble pantomimes are conventionalized and are thus distinct from pantomimes.[21] Interestingly, Emmorey (2002, Ch. 9) discusses studies of signers using ASL which show a dissociation between the neural systems involved in sign language and those involved in conventionalized gesture and pantomime. Corina et al. (1992b) reported left-hemisphere dominance for producing ASL signs, but no laterality effect when subjects had to produce symbolic gestures (e.g., waving good-bye or thumbs-up). Other studies report patients with left-hemisphere damage who exhibited sign language impairments but well-preserved conventional gesture and pantomime. Corina et al. (1992a) described patient W.L. with damage to left-hemisphere perisylvian regions. W.L. exhibited poor sign language comprehension and production. Nonetheless, this patient could produce stretches of pantomime and tended to substitute pantomimes for signs, even when the pantomime required more complex movement. Emmorey sees such data as providing neurological evidence that signed languages consist of linguistic gestures and not simply elaborate pantomimes.

Figure 4 is based on a scheme offered by Arbib (2004) in response to Hurford's (2004) critique of the mirror system hypothesis. Hurford makes the crucial point that we must (in the spirit of Saussure) distinguish the "sign" from the "signified." In the figure, we distinguish the "neural representation of the sign" (top row) from the "neural representation of the signified" (bottom row). The top row of the figure makes explicit the result of the progression within the mirror system hypothesis of mirror systems for:

1. Grasping and manual praxic actions.
2. Pantomime of grasping and manual praxic actions.
3. Pantomime of actions outside the pantomimic's own behavioral repertoire (e.g., flapping the arms to mime a flying bird).
4. Conventional gestures used to formalize and disambiguate pantomime (e.g., to distinguish "bird" from "flying").
5. Protosign, comprising conventionalized manual (and related orofacial) communicative gestures.

However, I disagree with Hurford's suggestion that there is a mirror system for all concepts – actions, objects, and more – which links the perception and action related to each concept.[22] In schema theory (Arbib 1981; 2003), I distinguish between *perceptual schemas*, which determine whether a given "domain of interaction" is present in the environment and provide parameters concerning the current relationship of the organism with that domain, and *motor schemas*, which provide the control systems which can be coordinated to effect a wide variety of actions. Recog-

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

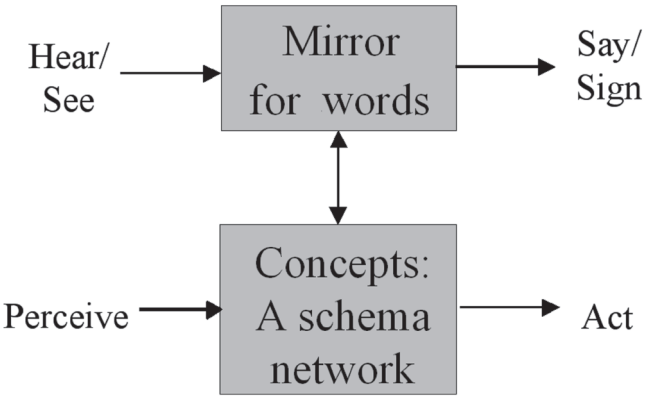

Figure 4. The bidirectional sign relation links words and concepts. The top row concerns Phonological Form, which may relate to signed language as much as to spoken language. The bottom row concerns Cognitive Form and includes the recognition of objects and actions. Phonological Form is present only in humans while Cognitive Form is present in both monkeys and humans. The Mirror System Hypothesis hypothesizes that there is a mirror system for words, but there may not be a mirror system for concepts.

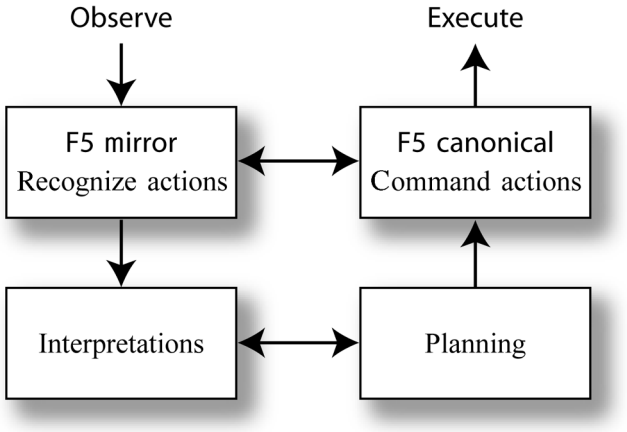

Figure 5. The perceptuomotor coding for both observation and execution contained in the mirror system for manual actions in the monkey is linked to "conceptual systems" for interpretation and planning of such actions. The interpretation and planning systems themselves do not have the mirror property save through their linkage to the actual mirror system.

nizing an object (an apple, say) may be linked to many different courses of action (to place the apple in one's shopping basket; to place the apple in the bowl at home; to peel the apple; to eat the apple; to discard a rotten apple, etc.). In this list, some items are apple-specific, whereas other invoke generic schemas for reaching and grasping. Such considerations led me to separate perceptual and motor schemas – a given action may be invoked in a wide variety of circumstances; a given perception may, as part of a larger assemblage, precede many courses of action. Hence, I reject the notion of a mirror system for concepts. Only rarely (as in the case of certain basic actions such as *grasp* or *run*, or certain expressions of emotion) will the perceptual and motor schemas be integrated into a "mirror schema." I do not see a "concept" as corresponding to one word, but rather to a graded set of activations of the schema network.

But if this is the case, does a mirror system for protosigns (and, later, for the words and utterances of a language) really yield the LR3 form of the mirror property – that what counts for the sender must count for the receiver? Actually, it yields only half of this directly: the recognition that the action of the observed protosigner is his or her version of one of the conventional gestures in the observer's repertoire. The claim, then, is that the LR3 form of the mirror property – that which counts for the sender must count for the receiver – does not result from the evolution of the F5 mirror system *in and of itself* to support communicative gestures as well as praxic actions; rather, this evolution occurs within the neural context that links the execution and observation of an action to the creature's planning of its own actions and interpretations of the actions of others (Fig. 5). These linkages extract more or less coherent patterns from the creature's experience of the effects of its own actions as well as the consequences of actions by others. Similarly, execution and observation of a communicative action must be linked to the creature's planning and interpretations of communication with others in relation to the ongoing behaviors that provide the significance of the communicative gestures involved.

6. The emergence of protospeech

6.1. The path to protospeech is indirect

My claim here is that the path to protospeech is indirect, with early protosign providing a necessary scaffolding for the emergence of protospeech. I thus reject the claim that speech evolved directly as an elaboration of a closed repertoire of alarm calls and other species-specific vocalizations such as exhibited by nonhuman primates. However, I claim neither that protosign attained the status of a full language prior to the emergence of early forms of protospeech, nor even that stage S5 (protosign) was completed before stage S6 (protospeech) began.

Manual gesture certainly appears to be more conducive to iconic representation than oral gesture. The main argument of section 5 was that the use of pantomime made it easy to acquire a core vocabulary, while the discovery of a growing stock of conventional signs (or sign modifiers) to mark important distinctions then created a culture in which the use of arbitrary gestures would increasingly augment and ritualize (without entirely supplanting) the use of pantomime.[23] Once an organism has an iconic gesture, it can both modulate that gesture and/or symbolize it (non-iconically) by "simply" associating a vocalization with it. Once the association had been learned, the "scaffolding" gesture (like the pantomime that supported its conventionalization, or the caricature that supports the initial understanding of some Chinese ideograms) could be dropped to leave a symbol that need have no remaining iconic relation to its referent, even if the indirect associative relationship can be recalled on some occasions. One open question is the extent to which protosign must be in place before this scaffolding can effectively support the development of protospeech. Because there is no direct mapping of sign (with its use of concurrency and signing space) to phoneme sequences, I think that this development is far more of a breakthrough than it may at first sight appear.

I have separated S6, the evolution of protospeech, from S5, the evolution of protosign, to stress the point that the role of F5 in grounding the evolution of a protolanguage

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

system would work just as well if we and all our ancestors had been deaf. However, primates do have a rich auditory system which contributes to species survival in many ways, of which communication is just one (Ghazanfar 2003). The protolanguage perception system could thus build upon the existing auditory mechanisms in the move to derive protospeech. However, it appears that considerable evolution of the vocal-motor system was needed to yield the flexible vocal apparatus that distinguishes humans from other primates. MacNeilage (1998) offers an argument for how the mechanism for producing consonant-vowel alternations en route to a flexible repertoire of syllables might have evolved from the cyclic mandibular alternations of eating, but offers no clue as to what might have linked such a process to the expression of meaning (but see MacNeilage & Davis, in press b). This problem is discussed much further in Arbib (2005) which spells out how protosign (S5) may have provided a scaffolding for protospeech (S6), forming an "expanding spiral" wherein the two interacted with each other in supporting the evolution of brain and body that made *Homo sapiens* "language-ready" in a multi-modal integration of manual, facial and vocal actions.

New data on mirror neurons for grasping that exhibit auditory responses, and on mirror-like properties of orofacial neurons in F5, add to the subtlety of the argument. Kohler et al. (2002) studied mirror neurons for actions which are accompanied by characteristic sounds, and found that a subset of these neurons are activated by the sound of the action (e.g., breaking a peanut in half) as well as sight of the action. Does this suggest that protospeech mediated by the F5 homologue in the hominid brain could have evolved without the scaffolding provided by protosign? My answer is negative for two reasons: (1) I have argued that imitation is crucial to grounding pantomime in which a movement is performed in the absence of the object for which such a movement would constitute part of a praxic action. However, the sounds studied by Kohler et al. (2002) cannot be created in the absence of the object, and there is no evidence that monkeys can use their vocal apparatus to mimic the sounds they have heard. I would further argue that the limited number and congruence of these "auditory mirror neurons" is more consistent with the view that manual gesture is primary in the early stages of the evolution of language readiness, with audiomotor neurons laying the basis for later extension of protosign to protospeech.

Complementing earlier studies on hand neurons in macaque F5, Ferrari et al. (2003) studied mouth motor neurons in F5 and showed that about one-third of them also discharge when the monkey observes another individual performing mouth actions. The majority of these "mouth mirror neurons" become active during the execution and observation of mouth actions related to ingestive functions such as grasping, sucking, or breaking food. Another population of mouth mirror neurons also discharges during the execution of ingestive actions, but the most effective visual stimuli in triggering them are communicative mouth gestures (e.g., lip-smacking) – one action becomes associated with a whole performance of which one part involves similar movements. This fits with the hypothesis that neurons learn to associate patterns of neural firing rather than being committed to learn specifically pigeonholed categories of data. Thus, a potential mirror neuron is in no way committed to become a mirror neuron in the strict sense, even though it may be more likely to do so than otherwise. The observed communicative actions (with the effective executed action for different "mirror neurons" in parentheses) include lip-smacking (sucking and lip-smacking); lips protrusion (grasping with lips, lips protrusion, lip-smacking, grasping, and chewing); tongue protrusion (reaching with tongue); teeth-chatter (grasping); and lips/tongue protrusion (grasping with lips and reaching with tongue; grasping). We therefore see that the communicative gestures and their associated effective observed actions are a long way from the sort of vocalizations that occur in speech (see Fogassi & Ferrari [in press] for further discussion).

Rizzolatti and Arbib (1998) stated that "This new use of vocalization [in speech] 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 orofacial gestures, which made possible a process of collateralization whereby this "proto" Broca's area gained primitive control of the vocal machinery, thus yielding increased skill and openness in vocalization, moving from the fixed repertoire of primate vocalizations to the unlimited (open) range of vocalizations exploited in speech. Speech apparatus and brain regions could then coevolve to yield the configuration seen in modern *Homo sapiens*.

Corballis (2003b) argues that there may have been a single-gene mutation producing a "dextral" allele, which created a strong bias toward right-handedness and left-cerebral dominance for language at some point in hominid evolution.[24] He then suggests that the "speciation event" that distinguished *Homo sapiens* from other large-brained hominids may have been a switch from a predominantly gestural to a predominantly vocal form of language. By contrast, I would argue that there was no one distinctive speciation event, and that the process whereby communication for most humans became predominantly vocal was not a switch but was "cultural" and cumulative.

7. The inventions of languages

The divergence of the Romance languages from Latin took about one thousand years. The divergence of the Indo-European languages to form the immense diversity of Hindi, German, Italian, English, and so on took about 6,000 years (Dixon 1997). How can we imagine what has changed since the emergence of *Homo sapiens* some 200,000 years ago? Or in 5,000,000 years of prior hominid evolution? I claim that the first *Homo sapiens* were language-ready but did not have language in the modern sense. Rather, my hypothesis is that stage S7, the transition from protolanguage to language, is the culmination of manifold discoveries in the history of mankind:

In section 2, I asserted that in much of protolanguage, a complete communicative act involved a unitary utterance, the use of a single symbol formed as a sequence of gestures, whose component gestures – whether manual or vocal – had no independent meaning. Unitary utterances such as "grooflook" or "koomzash" might have encoded quite complex descriptions such as "The alpha male has killed a meat animal and now the tribe has a chance to feast together. Yum, yum!" or commands such as "Take your spear and go

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

around the other side of that animal and we will have a better chance together of being able to kill it." On this view, "protolanguage" grew by adding arbitrary novel unitary utterances to convey *complex but frequently important* situations, and it was a major later discovery en route to language as we now understand it that one could gain expressive power by *fractionating* such utterances into shorter utterances conveying components of the scene or command (cf. Wray 1998; 2000). Put differently, the utterances of prelanguage were more akin to the "calls" of modern primates – such as the "leopard call" of the vervet monkey, which is emitted by a monkey who has seen a leopard and which triggers the appropriate escape behavior in other monkeys – than to sentences as defined in a language like English, but they differed *crucially* from the primate calls in that new utterances could be invented and acquired through learning within a community, rather than emerging only through biological evolution. Thus, the set of such unitary utterances was open, whereas the set of calls was closed.

The following hypothetical but instructive example is similar to examples offered at greater length by Wray (1998; 2000) to suggest how the fractionation of unitary utterances might occur (and see Kirby [2000] for a related computer simulation): Imagine that a tribe has two unitary utterances concerning fire which, by chance, contain similar substrings which become regularized so that for the first time there is a sign for "fire." Now the two original utterances are modified by replacing the similar substrings by the new regularized substring. Eventually, some tribe members regularize the complementary gestures in the first string to get a sign for "burns"; later, others regularize the complementary gestures in the second string to get a sign for "cooks meat." However, because of the arbitrary origin of the sign for "fire," the placement of the gestures that have come to denote "burns" relative to "fire" differs greatly from those for "cooks meat" relative to "fire." It therefore requires a further invention to regularize the placement of the gestures in both utterances – and in the process, words are crystallized at the same time as the protosyntax that combines them. Clearly, such fractionation could apply to protosign as well as to protospeech.

However, fractionation is not the only mechanism that could produce composite structures. For example, a tribe might over the generations develop different signs for "sour apple," "ripe apple," "sour plum," "ripe plum," and so on, but not have signs for "sour" and "ripe" even though the distinction is behaviorally important. Hence, $2n$ signs are needed to name n kinds of fruit. Occasionally someone will eat a piece of sour fruit by mistake and make a characteristic face and intake of breath when doing so. Eventually, some genius pioneers the innovation of getting a conventionalized variant of this gesture accepted as the sign for "sour" by the community, to be used as a warning before eating the fruit, thus extending the protolanguage.[25] A step towards language is taken when another genius gets people to use the sign for "sour" plus the sign for "ripe X" to replace the sign for "sour X" for each kind X of fruit. This innovation allows new users of the protolanguage to simplify learning fruit names, since now only $n + 1$ names are required for the basic vocabulary, rather than $2n$ as before. More to the point, if a new fruit is discovered, only one name need be invented rather than two. I stress that the invention of "sour" is a great discovery in and of itself. It might take hundreds of such discoveries distributed across centuries or more before someone could recognize the commonality across all these constructions and thus invent the precursor of what we would now call adjectives.[26]

The latter example is meant to indicate how a sign for "sour" could be added to the protolanguage vocabulary with no appeal to an underlying "adjective mechanism." Instead, one would posit that the features of language emerged by bricolage (tinkering) which added many features as "patches" to a protolanguage, with general "rules" emerging both consciously and unconsciously only as generalizations could be imposed upon, or discerned in, a population of ad hoc mechanisms. Such generalizations amplified the power of groups of inventions by unifying them to provide expressive tools of greatly extended range. According to this account, there was no sudden transition from unitary utterances to an elaborate language with a rich syntax and compositional semantics; no point at which one could say of a tribe "Until now they used protolanguage but henceforth they use language."

To proceed further, I need to distinguish two "readings" of a case frame like Grasp(Leo, raisin), as an action-object frame and as a verb-argument structure. I chart the transition as follows:

(1) As an *action-object frame*, Grasp(Leo, raisin) represents the perception that Leo is grasping a raisin. Here the action "grasp" involves two "objects," one the "grasper" Leo and the other the "graspee," the "raisin." Clearly the monkey has the perceptual capability to recognize such a situation[27] and enter a brain state that represents it, with that representation distributed across a number of brain regions. Indeed, in introducing principle LR5 (from hierarchical structuring to temporal ordering) I noted that the ability to translate a hierarchical conceptual structure into a temporally ordered structure of actions is apparent whenever an animal takes in the nature of a visual scene and produces appropriate behavior. *But to have such a capability does not entail the ability to communicate in a way that reflects these structures*. It is also crucial to note here the importance of recognition not only of the action (mediated by F5) but also of the object (mediated by IT). Indeed, Figure 2 (the FARS model) showed that the canonical activity of F5 already exhibits a choice between the affordances of an object (mediated by the dorsal stream) that involves the nature of the object (as recognized by IT and elaborated upon in PFC in a process of "action-oriented perception"). In the same way, the activity of mirror neurons does not rest solely upon the parietal recognition (in PF, Fig. 3) of the hand motion and the object's affordances (AIP) but also on the "semantics" of the object as extracted by IT. In the spirit of Figure 2, I suggest that this semantics is relayed via PFC and thence through AIP and PF to F5 to affect there the mirror neurons as well as the canonical neurons.

(2) My suggestion is that at least the immediate hominid precursors of *Homo sapiens* would have been able to perceive a large variety of action-object frames and, for many of these, to form a distinctive gesture or vocalization to appropriately direct the attention of another tribe member, but that the vocalization used would be in general a unitary utterance which need not have involved separate lexical entries for the action or the objects. However, the ability to symbolize more and more situations would have required the creation of a "symbol tool kit" of meaningless elements[28] from which an open-ended class of symbols could be generated.

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

(3) As a verb-argument structure, Grasp(Leo, raisin) is expressed in English in a sentence such as "Leo grasps the raisin," with "grasps" the verb, and "Leo" and "raisin" the arguments. I hypothesize that stage S7 was grounded in the development of precursors to verb-argument structure using vocalizations that were decomposable into "something like a verb" and two somethings that would be "something like nouns." This is the crucial step in the transition from protolanguage to human language as we know it. Abstract symbols are grounded (but more and more indirectly) in action-oriented perception; members of a community may acquire the use of these new symbols (the crucial distinction here is with the fixed repertoire of primate calls) by imitating their use by others; and, crucially, these symbols can be compounded in novel combinations to communicate about novel situations for which no agreed-upon unitary communicative symbol exists.

Having stressed above that adjectives are not a "natural category," I hasten to add that I do not regard verbs or nouns as natural categories either. What I do assert is that every human language must find a way to express the content of action-object frames. The vast variety of these frames can yield many different forms of expression across human languages. I view linguistic universals as being based on universals of communication that take into account the processing loads of perception and production rather than as universals of autonomous syntax. Hence, in emphasizing verb-argument structures in the form familiar from English, I am opting for economy of exposition rather than further illustration of the diversities of human language. To continue with the bricolage theme, much of "protosyntax" would have developed at first on an ad hoc basis, with variations on a few basic themes, rather than being grounded from the start in broad categories like "noun" or "verb" with general rule-like procedures to combine them in the phonological expression of cognitive form. 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. As a language emerged, it would come to include mechanisms to express kinship structures and technologies of the tribes, and these cultural products would themselves be expanded by the increased effectiveness of transmission from generation to generation that the growing power of language made possible. Evans (2003) supports this view by surveying a series of linguistic structures in which some syntactic rules must refer to features of the kinship system which are common in Australian aboriginal tribes but are unknown elsewhere. On this basis, we see such linguistic structures as historical products reflecting the impact of various processes of "cultural selection" on emerging structure.

If one starts with unitary utterances, then symbols that correspond to statements like "Take your spear and go around the other side of that animal and we will have a better chance together of being able to kill it" must each be important enough, or occur often enough, for the tribe to agree on a symbol (e.g., arbitrary string of phonemes) and for each one to replace an elaborate pantomime with a conventionalized utterance of protosign or protospeech. Discovering that separate names could be assigned to each actor, object, and action would require many words instead of one to express such an utterance. However, once the number of utterances with overlap reaches a critical level, economies of word learning would accrue from building utterances from "reusable" components (cf. the Wray-Kirby and "sour fruit" scenarios above). Separating verbs from nouns lets one learn m + n + p words (or less if the same noun can fill two roles) to be able to form m*n*p of the most basic utterances. Of course, not all of these combinations will be useful, but the advantage is that new utterances can now be coined "on the fly," rather than each novel event acquiring group mastery of a novel utterance.

Nowak et al. (2000) analyzed conditions under which a population that had two genes – one for unitary utterances and one for fractionated utterances – would converge into a situation in which one gene or the other (and therefore one type of language or the other) would predominate. But I feel that this misses the whole point: (1) It assumes that there is a genetic basis for this alternative, whereas I believe the basis is historical, without requiring genetic change. (2) It postulates that the alternatives already exist. I believe it is necessary to offer a serious analysis of how both unitary and fractionated utterances came to exist, and of the *gradual process* of accumulating changes that led from the predominance of the former to the predominance of the latter. (3) Moreover, it is not a matter of either/or – modern languages have a predominance of fractionated utterances but make wide use of unitary utterances as well.

The spread of these innovations rested on the ability of other humans not only to imitate the new actions and compounds of actions demonstrated by the innovators, but also to do so in a way that related increasingly general classes of symbolic behavior to the classes, events, behaviors, and relationships that they were to represent. Indeed, consideration of the spatial basis for "prepositions" may help show how visuomotor coordination underlies some aspects of language (cf. Talmy 2000), whereas the immense variation in the use of corresponding prepositions even in closely related languages like English and Spanish shows how the basic functionally grounded semantic-syntactic correspondences have been overlaid by a multitude of later innovations and borrowings.

The transition to *Homo sapiens* thus may have involved "language amplification" through increased speech ability coupled with the ability to name certain actions and objects separately, followed by the ability to create a potentially unlimited set of verb-argument structures and the ability to compound those structures in diverse ways. Recognition of hierarchical structure rather than mere sequencing provided the bridge to constituent analysis in language.

8. Towards a neurolinguistics "beyond the mirror"

Most of the stages of our evolutionary story are not to be seen so much as replacing "old" capabilities of the ancestral brain with new ones, but rather, as extending those capabilities by embedding them in an enriched system. I now build on our account of the evolution of the language-ready brain to offer a synchronic account of the "layered capabilities" of the modern adult human brain.

Aboitiz and García (1997) offer a neuroanatomical perspective on the evolutionary origin of the language areas in the human brain by analyzing possible homologies between language areas of the human brain and areas of the monkey

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

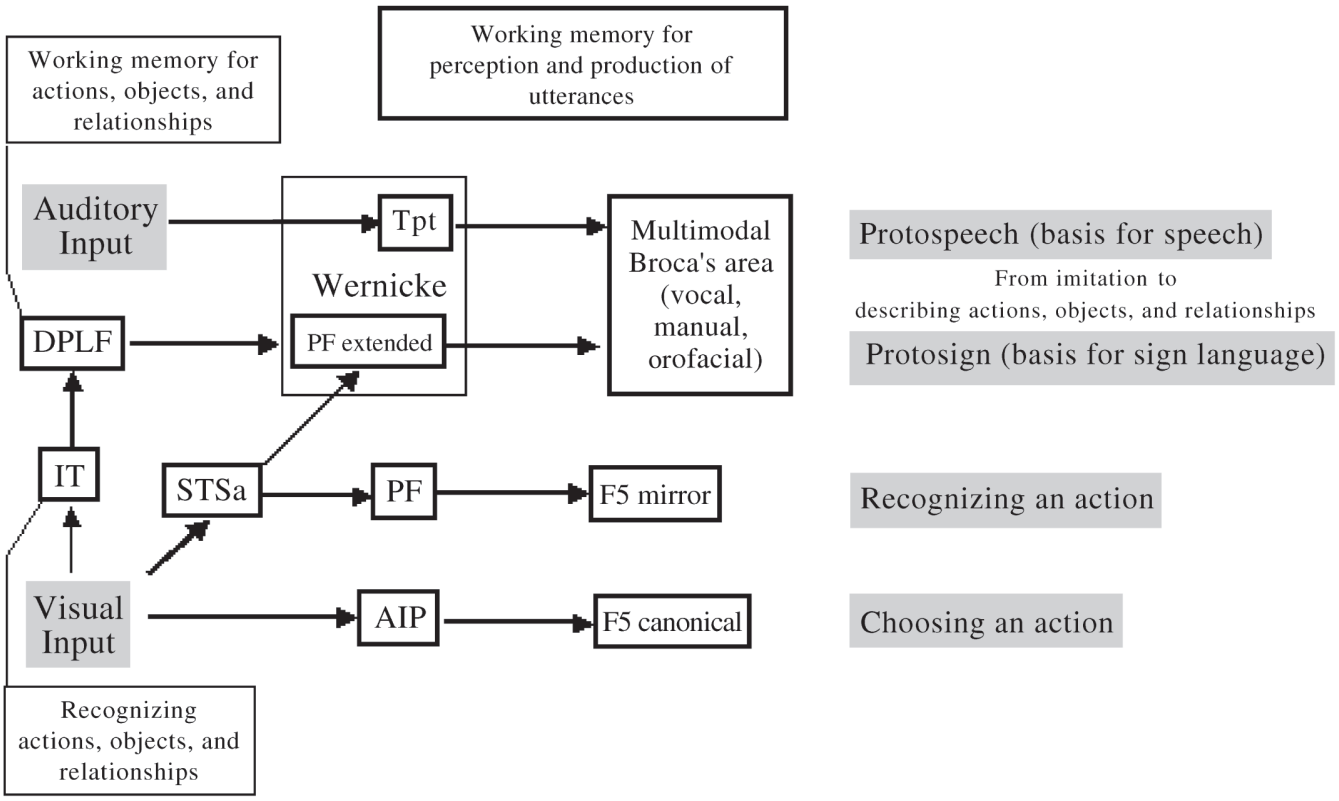

Figure 6. Extending the FARS model to include the mirror system for grasping and the language system evolved "atop" this. Note that this simple figure neither asserts nor denies that the extended mirror system for grasping and the language-supporting system are anatomically separable, nor does it address issues of lateralization. (From Arbib & Bota 2003.)

brain that may offer clues as to the structures of the brains of our ancestors of 20 million years ago. Arbib and Bota (2003) summarize the Aboitiz-García and mirror system hypotheses and summarize other relevant data on homologies between different cortical areas in macaque and human to ground further work on an evolutionary account of the readiness of the human brain for language.

Figure 6 is the diagram Arbib and Bota (2003) used to synthesize lessons about the language mechanisms of the human brain, extending a sketch for a "mirror neurolinguistics" (Arbib 2001b). This figure was designed to *elicit* further modeling; it does not have the status of fully implemented models, such as the FARS and MNS models, whose relation to, and prediction of, empirical results has been probed through computer simulation.

To start our analysis of Figure 6, note that an over-simple analysis of praxis, action understanding, and language production might focus on the following parallel parieto-frontal interactions:

I.	object → AIP → $F5_{canonical}$	praxis
II.	action → PF → $F5_{mirror}$	action understanding
III.	scene → Wernicke's → Broca's	language production

The data on patients A.T. and D.F. reviewed in section 3.1 showed a dissociation between the praxic use of size information (parietal) and the "declaration" of that information either verbally or through pantomime (inferotemporal). D.F. had a lesion allowing signals to flow from V1 towards posterior parietal cortex (PP) but not from V1 to inferotemporal cortex (IT). D.F. could preshape accurately when reaching to grasp an object, even though she was unable to declare, either verbally or in pantomime, the visual parameters that guided the preshape. By contrast, A.T. had a bilateral posterior parietal lesion. A.T. could use her hand to pantomime the size of a cylinder, but could not preshape appropriately when asked to grasp it. This suggests the following scheme:

IV. Parietal "affordances" → preshape
V. IT "perception of object" → pantomime or verbally describe size

That is, one cannot pantomime or verbalize an affordance; but rather one needs a "recognition of the object" (IT) to which attributes can be attributed before one can express them. Recall now the path shown in Figure 2 from IT to AIP, both directly and via PFC. I postulate that similar pathways link IT and PF. I show neither of these pathways in Figure 6, but rather show how this pathway might in the human brain not only take the form needed for praxic actions but also be "reflected" into a pathway that supports the recognition of communicative manual actions. We would then see the "extended PF" of this pathway as functionally integrated with the posterior part of Brodmann's area 22, or area Tpt (temporo-parietal) as defined by Galaburda and Sanides (1980). Indeed, lesion-based views of Wernicke's area may include not only the posterior part of Tpt but also (in whole or in part) areas in the human cortex that correspond to macaque PF (see Arbib & Bota [2003] for further details). In this way, we see Wernicke's area as combining capabilities for recognizing protosign and protospeech to support a language-ready brain that is capable of learning signed languages as readily as spoken languages.

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

Finally, we note that Arbib and Bota (2003) responded to the analysis of Aboitiz and García (1997) by including a number of working memories crucial to the linkage of visual scene perception, motor planning, and the production and recognition of language. However, they did not provide data on the integration of these diverse working memory systems into their anatomical scheme.

When building upon Figure 6 in future work in neurolinguistics, we need to bear in mind the definition of "complex imitation" as the ability to recognize another's performance as a set of familiar movements and then repeat them, but also to recognize when such a performance combines novel actions that can be approximated by (i.e., more or less crudely be imitated by) variants of actions already in the repertoire. Moreover, in discussing the FARS model in section 3.1, I noted that the interactions shown in Figure 2 are supplemented in the computer implementation of the model by code representing the role of the basal ganglia in administering sequences of actions, and that Bischoff-Grethe et al. (2003) model the possible role of the basal ganglia in interactions with the pre-SMA in sequence learning. Therefore, I agree with Visalberghi and Fragaszy's (2002, p. 495) suggestion that "[mirror] neurons provide a neural substrate for segmenting a stream of action into discrete elements matching those in the observer's repertoire, as Byrne (1999) has suggested in connection with his string-parsing theory of imitation," while adding that the success of complex imitation requires that the appropriate motor system be linked to appropriate working memories (as in Fig. 6) as well as to pre-SMA and basal ganglia (not shown in Fig. 6) to extract and execute the overall structure of the compound action (which may be sequential, or a more general coordinated control program [Arbib 2003]). Lieberman (2002) emphasizes that the roles of Broca's and Wernicke's areas must be seen in relation to larger neocortical and subcortical circuits. He cites data from studies of Broca's aphasia, Parkinson's disease, focal brain damage, and so on, to demonstrate the importance of the basal ganglia in sequencing the elements that constitute a complete motor act, syntactic process, or thought process. Hanakawa et al. (2002) investigated numerical, verbal, and spatial types of nonmotor mental-operation tasks. Parts of the posterior frontal cortex, consistent with the pre-supplementary motor area (pre-SMA) and the rostral part of the dorsolateral premotor cortex (PMdr), were active during all three tasks. They also observed activity in the posterior parietal cortex and cerebellar hemispheres during all three tasks. An fMRI study showed that PMdr activity during the mental-operation tasks was localized in the depths of the superior precentral sulcus, which substantially overlapped the region active during complex finger movements and was located dorsomedial to the presumptive frontal eye fields.

Such papers are part of the rapidly growing literature that relates human brain mechanisms for action recognition, imitation, and language. A full review of such literature is beyond the scope of the target article, but let me first list a number of key articles – Binkofski et al. (1999), Decety et al. (1997), Fadiga et al. (2002), Grezes et al. (1998), Grezes and Decety (2001; 2002), Heiser et al. (2003), Hickok et al. (1998), Iacoboni et al. (1999; 2001), and Floel et al. (2003) – and then briefly describe a few others:

Koski et al. (2002) used fMRI to assess the effect of explicit action goals on neural activity during imitation. Their results support the hypothesis that areas relevant to motor preparation and motor execution are tuned to coding goal-oriented actions and are in keeping with single-cell recordings revealing that neurons in area F5 of the monkey brain represent goal-directed aspects of actions. Grezes et al. (2003) used event-related fMRI to investigate where in the human brain activation can be found that reflects both canonical and mirror neuronal activity. They found activation in the intraparietal and ventral limbs of the precentral sulcus when subjects observed objects and when they executed movements in response to the objects ("canonical neurons"); and activation in the dorsal premotor cortex, the intraparietal cortex, the parietal operculum (SII), and the superior temporal sulcus when subjects observed gestures ("mirror neurons"). Finally, activations in the ventral premotor cortex and inferior frontal gyrus (Brodmann area [BA] 44) were found when subjects imitated gestures and executed movements in response to objects. These results suggest that in the human brain, the ventral limb of the precentral sulcus may form part of the area designated F5 in the macaque monkey. It is possible that area 44 forms an anterior part of F5, though anatomical studies suggest that it may be a transitional area between the premotor and prefrontal cortices.

Manthey et al. (2003) used fMRI to investigate whether paying attention to objects versus movements modulates premotor activation during the observation of actions. Participants were asked to classify presented movies as showing correct actions, erroneous actions, or senseless movements. Erroneous actions were incorrect either with regard to employed objects, or to performed movements. The ventrolateral premotor cortex (vPMC) and the anterior part of the intraparietal sulcus (aIPS) were strongly activated during the observation of actions in humans. Premotor activation was dominantly located within BA 6, and sometimes extended into BA 44. The presentation of object errors and movement errors showed that left premotor areas were more involved in the analysis of objects, whereas right premotor areas were dominant in the analysis of movements. (Since lateralization is not analyzed in this article, such data may be a useful springboard for commentaries.)

To test the hypothesis that action recognition and language production share a common system, Hamzei et al. (2003) combined an action recognition task with a language production task and a grasping movement task. Action recognition-related fMRI activation was observed in the left inferior frontal gyrus and on the border between the inferior frontal gyrus (IFG) and precentral gyrus (PG), the ventral occipito-temporal junction, the superior and inferior parietal cortex, and in the intraparietal sulcus in the left hemisphere. An overlap of activations due to language production, movement execution, and action recognition was found in the parietal cortex, the left inferior frontal gyrus, and the IFG-PG border. The activation peaks of action recognition and verb generation were always different in single subjects, but no consistent spatial relationship was detected, presumably suggesting that action recognition and language production share a common functional architecture, with functional specialization reflecting developmental happenstance.

Several studies provide behavioral evidence supporting the hypothesis that the system involved in observation and preparation of grasp movements partially shares the cortical areas involved in speech production. Gentilucci (2003a) had subjects pronounce either the syllable *ba* or *ga* while

https://doi.org/10.1017/S0140525X05000038 Published online by Cambridge University Press

observing motor acts of hand grasp directed to objects of two sizes, and found that both lip aperture and voice peak amplitude were greater when the observed hand grasp was directed to the large object. Conversely, Glover and Dixon (2002; see Glover et al. 2004 for related results) presented subjects with objects on which were printed either the word *large* or *small*. An effect of the words on grip aperture was found early in the reach, but this effect declined continuously as the hand approached the target, presumably due to the effect of visual feedback. Gerlach et al. (2002) showed that the left ventral premotor cortex is activated during categorization not only for tools but also for fruits and vegetables and articles of clothing, relative to animals and nonmanipulable man-made objects. Such findings support the notion that certain lexical categories may evolve from action-based knowledge but are difficult to account for should knowledge representations in the brain be truly categorically organized.

Several insights have been gleaned from the study of signed language. Corina et al. (2003) used PET to examine deaf users of ASL as they generated verb signs independently with their right dominant and left nondominant hands (compared to the repetition of noun signs). Nearly identical patterns of left inferior frontal and right cerebellum activity were observed, and these were consistent with patterns that have been reported for spoken languages. Thus, lexical-semantic processing in production relies upon left-hemisphere regions regardless of the modality in which a language is realized, and, in signing, no matter which hand is used. Horwitz et al. (2003) studied the activation of Broca's area during the production of spoken and signed language. They showed that BA45, not BA44, was activated by both speech and signing during the production of language narratives in bilingual subjects (fluent from early childhood in both ASL and English) with the generation of complex movements and sounds as control. Conversely, BA44, not BA45, was activated by the generation of complex articulatory movements of oral-laryngeal or limb musculature. Horwitz et al. therefore conclude that BA45 is the part of Broca's area that is fundamental to the modality-independent aspects of language generation.

Gelfand and Bookheimer (2003), using fMRI, found that the posterior portion of Broca's area responded specifically to sequence manipulation tasks, whereas the left supramarginal gyrus was somewhat more specific to sequencing phoneme segments. These results suggest that the left posterior inferior frontal gyrus responds not to the sound structure of language but rather to sequential operations that may underlie the ability to form words out of dissociable elements.

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

ACKNOWLEDGMENTS

The early stages of building upon "Language within Our Grasp" (Rizzolatti & Arbib 1998) were conducted during my sabbatical visits in 1999 to the University of Western Australia and the Institute of Human Physiology in Parma, Italy, and my conversations there with Robyn Owens, E. J. Holden, Giacomo Rizzolatti, Morten Christiansen, Giuseppe Cossu, Giuseppe Luppino, Massimo Matelli, Vittorio Gallese, and other colleagues. So many people have offered perceptive comments on various results of that effort (as published in, e.g., Arbib 2001a; 2001b; 2002) that the following list is woefully incomplete – Shannon Casey, Chris Code, Bob Damper, Kerstin Dautenhahn, Barry Gordon, Jim Hurford, Bipin Indurkhya, Chrystopher Nehaniv, and Chris Westbury – but I do hope that all these people (and the *BBS* referees), whether named or not, will realize how much I value their thoughtful comments and that they will see how their suggestions and comments have helped me clarify, correct, and extend my earlier analyses.

Preparation of the present paper was supported in part by a fellowship from the Center for Interdisciplinary Research of the University of Southern California. In particular, this fellowship allowed me to initiate a faculty seminar in September of 2002 at which my ideas have been exposed to intense though friendly scrutiny and placed in the context of the range of fascinating work by the members of the seminar – Amit Almor, Elaine Andersen, Aude Billard, Mihail Bota, Dani Byrd, Vincent Chen, Karen Emmorey, Andrew Gordon, James Gordon, Jack Hawkins, Jerry R. Hobbs, Laurent Itti, Toby Mintz, Stefan Schaal, Craig Stanford, Jean-Roger Vergnaud, Christoph von der Malsburg, Carolee Winstein, Michail Zak, Patricia Zukow-Goldring, and Kie Zuraw.

NOTES

1. Bickerton (1995) views infant language, pidgins, and the "language" taught to apes as *protolanguages* in the sense of a form of communication whose users can only string together a small handful of words at a time with little if any syntax. Bickerton hypothesizes that the protolanguage (in my sense) of *Homo erectus* was a protolanguage in his sense, in which a few words much like those of today's language are uttered a few at a time to convey meaning without the aid of syntax. I do not assume (or agree with) this hypothesis.

2. Today's signed languages are fully expressive human languages with a rich syntax and semantics, and are not to be confused with the posited systems of protosign communication. By the same taken, protospeech is a primitive form of communication based on vocal gestures but without the richness of modern human spoken languages.

3. Since we will be concerned in what follows with sign language as well as spoken language, the "speaker" and "hearer" may be using hand and face gestures rather than vocal gestures for communication.

4. However, I shall offer below the view that early forms of protosign provided a scaffolding for the initial development of protospeech, rather than holding that protosign was "completed" before protospeech was "initiated."

5. I would welcome commentaries on "language-like" aspects of communication in nonprimates, but the present article is purely about changes within the primates that led to the human language-ready brain.

6. It could be objected that monkey calls are not "involuntary communication" because, for example, vervet alarm calls are given usually in the presence of conspecifics who would react to them. However, I would still call this involuntary – this just shows that two conditions, rather than one, are required to trigger the call. This is distinct from the human use of language to conduct a conversation that may have little or no connection to the current situation.

7. When I speak of a "stage" in phylogeny, I do not have in mind an all-or-none switch in the genotype that yields a discontinuous change in the phenotype, but rather the coalescence of a variety of changes that can be characterized as forming a global pattern that may emerge over the course of tens or even hundreds of millennia.

8. Let me stress that complex imitation involves both the

recognition of an action as a certain combination of actions and the ability to replicate (something like) that combination. Both skills play a role in the human child's acquisition of language; the latter remains important in the adult's language comprehension.

9. But see note 4 above.

10. The attainment of complex imitation was seen as a crucial stage of the evolution of language readiness in Arbib (2002), but was not listed there as a condition for language readiness. I now see this as a mistake.

11. Unfortunately, space does not permit development of an argument for this controversial claim. Commentaries pro or con the hypothesis will be most welcome.

12. I wonder at times whether properties LR1 through LR7 do indeed support LA1 or whether LA1 should itself be seen as part of the biological equipment of language readiness. I would welcome commentaries in support of either of these alternatives. However, I remain convinced that LR1 through LR7 coupled with LA1 provide all that is needed for a brain to support LA2, LA3, and LA4.

13. The pairs (LR6: Beyond the here-and-now 1; LA3: Beyond the here-and-now 2) and (LR7: Paedomorphy and sociality; LA4: Learnability) do not appear in Table 1 because the rest of the paper will not add to their brief treatment in section 2.2.

14. Figure 2 provides only a partial overview of the model. The full model (see Fagg & Arbib 1998 for more details) includes a number of brain regions, offering schematic models for some and detailed neural-network models for others. The model has been implemented on the computer so that simulations can demonstrate how the activities of different populations vary to explain the linkage between visual affordance and manual grasp.

15. To keep the exposition compact, in what follows I will use without further explanation the abbreviations for the brain regions not yet discussed. The reader wanting to see the abbreviations spelled out, as well as a brief exposition of data related to the hypothesized linkage of schemas to brain structures, is referred to Oztop and Arbib (2002).

16. Estimates for the timetable for hominid evolution (I use here those given by Gamble 1994, see his Fig. 4.2) are 20 million years ago for the divergence of monkeys from the line that led to humans and apes, and 5 million years ago for the divergence of the hominid line from the line that led to modern apes.

17. For more on "chimpanzee culture," see Whiten et al. (2001) and the Chimpanzee Cultures Web site: http://culture.st-and.ac.uk:16080/chimp/, which gives access to an online database that describes the cultural variations in chimpanzee behavior and shows behavior distributions across the sites in Africa where long-term studies of chimpanzees have been conducted in the wild.

18. Recall the observation (Note 8) that both the recognition of an action as a certain combination of actions and the ability to replicate (something like) that combination play a role in the human child's acquisition of language, while the former remains important in the adult's language comprehension. But note, too, that stage S4 only takes us to complex imitation of *praxic* actions; Sections 5 and 6 address the transition to an open system of *communicative* actions.

19. As ahistorical support for this, note that *airplane* is signed in American Sign Language (ASL) with tiny repeated movements of a specific handshape, whereas *fly* is signed by moving the same handshape along an extended trajectory (Supalla & Newport 1978). I say "ahistorical" because such signs are part of a modern human language rather than holdovers from protosign. Nonetheless, they exemplify the mixture of iconicity and convention that, I claim, distinguishes protosign from pantomime.

20. Of course, relatively few Chinese characters are so pictographic in origin. For a fuller account of the integration of semantic and phonetic elements in Chinese characters (and a comparison with Sumerian logograms) see Chapter 3 of Coulmas 2003.

21. Of course, those signs that most clearly resemble pantomimes will be easier for the nonsigner to recognize, just as certain Chinese characters are easier for the novice to recognize. Shannon Casey (personal communication) notes that moving the hands in space to represent actions involving people interacting with people, animals, or other objects is found in signed languages in verbs called "spatial verbs" or "verbs of motion and location." These verbs can be used with handshapes to represent people or objects called "semantic classifiers" and "size and shape specifiers" (Supalla 1986; see p. 196 for a description of these classifiers and p. 211 for figures of them). Hence, to describe giving someone a cup, the ASL signer may either use the standard *give* handshape (palm up with fingertips and thumb-tip touching) or use an open, curved handshape with the fingertips and thumb-tip apart and the palm to the side (as if holding a cup). Similarly, to describe giving someone a thick book, the signer can use a handshape with the palm facing up, fingertips pointing outward and thumb also pointing outward with about an inch of space between the thumb and fingertips (as if holding a book). In her own research Casey (2003) has found that hearing subjects with no knowledge of a signed language do produce gestures resembling classifiers. Stokoe (2001, pp. 188–91) relates the use of shape classifiers in ASL to the use of shape classifiers in spoken Native American languages.

22. Added in proof: Hurford notes that this suggestion was made and discarded prior to publication of Hurford (2004).

23. Such developments and inventions may have occurred very slowly over the course of many (perhaps even thousands) of generations during which expansion of the proto-vocabulary was piecemeal; it may then have been a major turning point in human history when it was realized that symbols could be created *ad libitum* and this realization was passed on to future generations. See also Note 25.

24. Where Corballis focuses on the *FOXP2* gene, Crow (2002a) links lateralization and human speciation to a key mutation which may have speciated on a change in a homologous region of the X and Y chromosomes.

25. I use the word "genius" advisedly. I believe that much work on language evolution has been crippled by the inability to imagine that things we take for granted were in no way a priori obvious, or to see that current generalities were by no means easy to discern in the particularities that they embrace. Consider, for example, that Archimedes (c. 287–212 BCE) had the essential idea of the integral calculus, but it took almost 2,000 years before Newton (1642–1727) and Leibniz (1646–1716) found notations that could express the generality implicit in his specific examples and hence unleash an explosion of mathematical innovation. I contend that language, like mathematics, has evolved culturally by such fits and starts. Note 23.

26. Indeed, adjectives are not the "natural category" they may appear to be. As Dixon (1997, pp. 142 et seq.) observes, there are two kinds of adjective classes across human languages: (1) an open class with hundreds of members (as in English); (2) a small closed class. Languages with small adjective classes are found in every continent except Europe. Igbo, from west Africa, has just eight adjectives: *large* and *small*; *black/dark* and *white/light*; *new* and *old*; and *good* and *bad*. Concepts that refer to physical properties tend to be placed in the verb class (e.g., "the stone heavies") and words referring to human propensities tend to be nouns (e.g., "she has cleverness").

27. Leaving aside the fact that the monkey probably does not know that Leo's name is "Leo."

28. Not all the symbols need be meaningless; some signs of a signed language can be recognized as conventionalized pantomime, and some Chinese characters can be recognized as conventionalized pictures. But we have already noted that relatively few Chinese characters are pictographic in origin. Similarly, many signs have no link to pantomime. As Coulmas (2003) shows us in analyzing writing systems – but the point holds equally well for speech and sign – the mixture of economy of expression and increasing range of expression leads to more and more of a symbol being built up from meaningless components.