The Origin and Evolution of Language:  
A Plausible, Strong-AI Account  

Jerry R. Hobbs  
USC Information Sciences Institute  
Marina del Rey, California

Abstract  
A large part of the mystery of the origin of language is the difficulty we experience in trying to imagine what the intermediate stages along the way to language could have been. An elegant, detailed, formal account of how discourse interpretation works in terms of a mode of inference called abduction, or inference to the best explanation, enables us to spell out with some precision a quite plausible sequence of such stages. In this chapter I outline plausible sequences for two of the key features of language—Gricean nonnatural meaning and syntax. I then speculate on the time in the evolution of modern humans each of these steps may have occurred.

1 Framework  
In this chapter I show in outline how human language as we know it could have evolved incrementally from mental capacities it is reasonable to attribute to lower primates and other mammals. I do so within the framework of a formal computational theory of language understanding (Hobbs et al., 1993). In the first section I describe some of the key elements in the theory, especially as it relates to the evolution of linguistic capabilities. In the next two sections I describe plausible incremental paths to two key aspects of language—meaning and syntax. In the final section I discuss various considerations of the time course of these processes.

1.1 Strong AI  
It is desirable for psychology to provide a reduction in principle of intelligent, or intentional, behavior to neurophysiology. Because of the extreme
complexity of the human brain, more than the skethest account is not likely to be possible in the near future. Nevertheless, the central metaphor of cognitive science, “The brain is a computer”, gives us hope. Prior to the computer metaphor, we had no idea of what could possibly be the bridge between beliefs and ion transport. Now we have an idea. In the long history of inquiry into the nature of mind, the computer metaphor gives us, for the first time, the promise of linking the entities and processes of intentional psychology to the underlying biological processes of neurons, and hence to physical processes. We could say that the computer metaphor is the first, best hope of materialism.

The jump between neurophysiology and intentional psychology is a huge one. We are more likely to succeed in linking the two if we can identify some intermediate levels. A view that is popular these days identifies two intermediate levels—the symbolic and the connectionist.

\[
\text{Intentional Level} \quad | \quad \text{Symbolic Level} \quad | \quad \text{Connectionist Level} \quad | \quad \text{Neurophysiological Level}
\]

The intentional level is implemented in the symbolic level, which is implemented in the connectionist level, which is implemented in the neurophysiological level.\(^1\) From the “strong AI” perspective, the aim of cognitive science is to show how entities and processes at each level emerge from the entities and processes of the level below.

The reasons for this strategy are clear. We can observe intelligent activity and we can observe the firing of neurons, but there is no obvious way of linking these two together. So we decompose the problem into three smaller problems. We can formulate theories at the symbolic level that can, at least in a small way so far, explain some aspects of intelligent behavior; here we work from intelligent activity down. We can formulate theories at the connectionist level in terms of elements that behave very much like what we know of the neuron’s behavior; here we work from the neuron up. Finally, efforts are being made to implement the key elements of symbolic processing in connectionist architecture. If each of these three efforts were to succeed, we would have the whole picture.

\(^1\)Variations on this view dispense with the symbolic or with the connectionist level.
In my view, this picture looks very promising indeed. Mainstream AI and cognitive science have taken it to be their task to show how intentional phenomena can be implemented by symbolic processes. The elements in a connectionist network are modeled closely on certain properties of neurons. The principal problems in linking the symbolic and connectionist levels are representing predicate-argument relations in connectionist networks, implementing variable-binding or universal instantiation in connectionist networks, and defining the right notion of “defeasibility” in logic to reflect the “soft corners” that make connectionist models so attractive. Progress is being made on all these problems.

Although we do not know how each of these levels is implemented in the level below, nor indeed whether it is, we know that it could be, and that at least is something.

1.2 Logic as the Language of Thought

A very large body of work in AI begins with the assumptions that information and knowledge should be represented in first-order logic and that reasoning is theorem-proving. On the face of it, this seems implausible as a model for people. It certainly doesn’t seem as if we are using logic when we are thinking, and if we are, why are so many of our thoughts and actions so illogical? In fact, there are psychological experiments that purport to show that people do not use logic in thinking about a problem (e.g., Wason and Johnson-Laird, 1972).

I believe that the claim that logic is the language of thought comes to less than one might think, however, and that thus it is more controversial than it ought to be. It is the claim that a broad range of cognitive processes are amenable to a high-level description in which six key features are present. The first three of these features characterize propositional logic and the next two first-order logic. I will express them in terms of “concepts”, but one can just as easily substitute propositions, neural elements, or a number of other terms.

- **Conjunction:** There is an additive effect \((P \land Q)\) of two distinct concepts \((P\) and \(Q)\) being activated at the same time.

- **Modus Ponens:** The activation of one concept \((P)\) triggers the activation of another concept \((Q)\) because of the existence of some structural relation between them \((P \supset Q)\).
Recognition of Obvious Contradictions: The recognition of contradictions in general is undecidable, but we have no trouble with the easy ones, for example, that cats aren’t dogs.

Predicate-Argument Relations: Concepts can be related to other concepts in several different ways. We can distinguish between a dog biting a man ($\text{bite}(D,M)$) and a man biting a dog ($\text{bite}(M,D)$).

Universal Instantiation (or Variable Binding): We can keep separate our knowledge of general (universal) principles (“All men are mortal”) and our knowledge of their instatiations for particular individuals (“Socrates is a man” and “Socrates is mortal”).

Any plausible proposal for a language of thought must have at least these features, and once you have these features you have first-order logic.

Note that in this list there are no complex rules for double negations or for contrapositives (if $P$ implies $Q$ then not $Q$ implies not $P$). In fact, most of the psychological experiments purporting to show that people don’t use logic really show that they don’t use the contrapositive rule or that they don’t handle double negations well. If the tasks in those experiments were recast into problems involving the use of modus ponens, no one would think to do the experiments because it is obvious that people would have no trouble with the task.

There is one further property we need of the logic if we are to use it for representing and reasoning about commonsense world knowledge—defeasibility or nonmonotonicity. Our knowledge is not certain. Different proofs may have different consequences, and one proof can be “better” than another.

The mode of defeasible reasoning used here is “abduction”, or inference to the best explanation. Briefly, one tries to prove something, but where there is insufficient knowledge, one can make assumptions. One proof is better than another if it makes fewer, more plausible assumptions, and if the knowledge it uses is more plausible and more salient. This is spelled out in detail in Hobbs et al. (1993).

1.3 Discourse Interpretation: Examples of Definite Reference

In the “Interpretation as Abduction” framework, world knowledge is expressed as defeasible logical axioms. To interpret the content of a discourse
is to find the best explanation for it, that is, to find a minimal-cost abductive proof of its logical form. To interpret a sentence is to deduce its syntactic structure and hence its logical form, and simultaneously to prove that logical form abductively. To interpret suprasentential discourse is to interpret individual segments, down to the sentential level, and to abduce relations among them.

Consider as an example the problem of resolving definite references. The following four examples are sometimes taken to illustrate four different kinds of definite reference.

I bought a new car last week. *The car* is already giving me trouble.

I bought a new car last week. *The vehicle* is already giving me trouble.

I bought a new car last week. *The engine* is already giving me trouble.

*The engine* of my new car is already giving me trouble.

In the first example, the same word is used in the definite noun phrase as in its antecedent. In the second example, a hyponym is used. In the third example, the reference is not to the “antecedent” but to an object that is related to it, requiring what Clark (1975) called a “bridging inference”. The fourth example is a determinative definite noun phrase, rather than an anaphoric one; all the information required for its resolution is found in the noun phrase itself.

These distinctions are insignificant in the abductive approach. In each case we need to prove the existence of the definite entity. In the first example it is immediate. In the second, we use the axiom

$$(\forall x) car(x) \supset vehicle(x)$$

In the third example, we use the axiom

$$(\forall x) car(x) \supset (\exists y) engine(y, x)$$

that is, cars have engines. In the fourth example, we use the same axiom, but after assuming the existence of the speaker’s new car.
1.4 Syntax in the Abduction Framework

Syntax can be integrated into this framework in a thorough fashion, as described at length in Hobbs (1998). In this treatment, the predication

\[(1) \quad \text{Syn}(w, e, \ldots)\]

says that the string \(w\) is a grammatical, interpretable string of words describing the situation or entity \(e\). The arguments of \(\text{Syn}\) indicated by the dots include information about complements and various agreement features.

Composition is effected by axioms of the form

\[(2) \quad \text{Syn}(w_1, e, \ldots, y, \ldots) \land \text{Syn}(w_2, y, \ldots) \supset \text{Syn}(w_1w_2, e, \ldots)\]

A string \(w_1\) whose head describes the eventuality \(e\) and which is missing an argument \(y\) can be concatenated with a string \(w_2\) describing \(y\), yielding a string describing \(e\). The interface between syntax and world knowledge is effected by “lexical axioms” of a form illustrated by

\[(3) \quad \text{read}'(e, x, y) \land \text{text}(y) \supset \text{Syn}(\text{"read"}, e, \ldots, x, \ldots, y, \ldots)\]

If \(e\) is the eventuality of \(x\) reading \(y\) (the logical form fragment supplied by the word “read”), where \(y\) is a text (the selectional constraint imposed by “read”), then \(e\) can be described by a phrase headed by the word “read” provided it picks up, as subject and object, phrases of the right sort describing \(x\) and \(y\).

To interpret a sentence \(w\), one seeks to show it is a grammatical, interpretable string of words by proving there in an eventuality \(e\) that it describes, that is, by proving (1). One does so by decomposing it via composition axioms like (2) and bottoming out in lexical axioms like (3). This yields the logical form of the sentence, which then must be proved abductively, the characterization of interpretation we gave above.

1.5 Discourse Structure

When confronting an entire discourse, one must break it into interpretable segments and show that those segments themselves are coherently related. That is, one must use a rule like

\[\text{Segment}(w_1, e_1) \land \text{Segment}(w_2, e_2) \land \text{rel}(e, e_1, e_2) \supset \text{Segment}(w_1w_2, e)\]
That is, if \( w_1 \) and \( w_2 \) are interpretable segments describing situations \( e_1 \) and \( e_2 \) respectively, and \( e_1 \) and \( e_2 \) stand in some relation \( rel \) to each other, then the concatenation of \( w_1 \) and \( w_2 \) constitute an interpretable segment, describing a situation \( e \) that is determined by the relation. More about the possible relations in Section 4.

This rule applies recursively and bottoms out in sentences.

\[
\text{Syn}(w,e,\ldots) \supset \text{Segment}(w,e)
\]

A grammatical, interpretable sentence \( w \) describing eventuality \( e \) is a coherent segment of discourse describing \( e \). This axiom effects the interface between syntax and discourse structure.

To interpret a discourse, we break it into coherently related successively smaller segments until we reach the level of sentences. Then we do a syntactic analysis of the sentences, bottoming out in their logical form, which we then prove abductively.\(^2\)

1.6 Discourse as a Purposeful Activity

This view of discourse interpretation is embedded in a view of interpretation in general in which an agent, to interpret the environment, must find the best explanation for the observables in that environment.

An intelligent agent is embedded in the world and must, at each instant, understand the current situation. The agent does so by finding an explanation for what is perceived. Put differently, the agent must explain why the complete set of observables encountered constitutes a coherent situation. Other agents in the environment are viewed as intentional, that is, as planning mechanisms, and this means that the best explanation of their observable actions is most likely to be that the actions are steps in a coherent plan. Thus, making sense of an environment that includes other agents entails making sense of the other agents’ actions in terms of what they are intended to achieve. When those actions are utterances, the utterances must be understood as actions in a plan the agents are trying to effect. The speaker’s plan must be recognized.

Generally, when a speaker says something it is with the goal that the hearer believe the content of the utterance, or think about it, or consider it, or take some other cognitive stance toward it. Let us subsume all these

\(^2\)This is an idealized, after-the-fact picture of the result of the process. In fact, interpretation, or the building up of this structure, proceeds word-by-word as we hear or read the discourse.
mental terms under the term “cognize”. We can then say that to interpret a speaker $A$’s utterance to $B$ of some content, we must explain the following:

$$\text{goal}(A, \text{cognize}(B, \text{content-of-discourse}))$$

Interpreting the content of the discourse is what we described above. In addition to this, one must explain in what way it serves the goals of the speaker to change the mental state of the hearer to include some mental stance toward the content of the discourse. We must fit the act of uttering that content into the speaker’s presumed plan.

The (defeasible) axiom that encapsulates this is

$$(\forall s, h, e_1, e, w) \text{goal}(s, e_1) \land \text{cognize}'(e_1, h, e) \land \text{Segment}(w, e) \supset \text{utter}(s, h, w)$$

That is, normally if a speaker $s$ has a goal of the hearer $h$ cognizing a situation $e$ and $w$ is a string of words that conveys $e$, then $s$ will utter $w$ to $h$. We appeal to this axiom to interpret the utterance as an intentional communicative act. That is, if $A$ utters to $B$ a string of words $W$, then to explain this observable event, we have to prove

$$\text{utter}(A, B, W)$$

and we begin to do so by backchaining on the above axiom. Reasoning about the speaker’s plan is a matter of establishing the first two propositions in the antecedent of the axiom. Determining the informational content of the utterance is a matter of establishing the third. The two sides of the proof influence each other since they share variables and since a minimal proof will result when both are explained and when their explanations use much of the same knowledge.

### 1.7 A Structured Connectionist Realization of Abduction

Because of its elegance and very broad coverage, the abduction model is very appealing on the symbolic level. But to be a plausible candidate for how people understand language, there must be an account of how it could be implemented in neurons. In fact, the abduction framework can be realized in a structured connectionist model called SHRUTI developed by Lokendra Shastri (Shastri and Ajjanagadde, 1993). Substantial work must be done in neurophysics to determine whether this kind of model is what actually exists in the human brain, although there is suggestive evidence. But by
linking the symbolic and connectionist levels one at least provides a proof of possibility for the abductive framework.

Traditional connectionist models have been very good at implementing defeasible propositional logic. Indeed, nearly all the applications to natural language processing in this tradition begin by setting up the problem so that it is a problem in propositional logic. But this is not adequate for natural language understanding in general. For example, the coreference problem requires the expressivity of first-order logic even to state. We need a way of expressing predicate-argument relations and a way of expressing different instantiations of the same general principle. We need a mechanism for universal instantiation. In the connectionist literature, this has gone under the name of the variable-binding problem.

The essential idea behind the shruti architecture is simple and elegant. A predication is represented as an assemblage or cluster of nodes, and axioms representing general knowledge are realized as connections among these clusters. Inference is accomplished by means of spreading activation through these structures.

In the cluster representing predications, two nodes, a collector node and an enabler node, correspond to the predicate and fire asynchronously. The level of activation on the enabler node keeps track of the “utility” of this predication in the proof that is being searched for. That is, the activation is higher the greater the need to find a proof for this predication, and thus the more expensive it is to assume. The level of activation on the collector node is higher the greater the plausibility that this predication is part of the desired proof. We can think of the activations on the enabler nodes as prioritizing goal expressions, whereas the activations on the collector nodes indicate degree of belief in the predications, or more properly, degree of belief in the current relevance of the predications. The connections between nodes of different predication clusters have a strength of activation, or link weight, that corresponds to strength of association between the two concepts. The proof process then consists of activation spreading through enabler nodes, as we backchain through axioms, and then spreading back through collector nodes after bottoming out in something known or assumed.

In addition, in the predication cluster, there are argument nodes, one for each argument of the predication. These fire synchronously with the argument nodes in other predication clusters to which they are connected. Thus, if the clusters for \( p(x, y) \) and \( q(z, x) \) are connected, with the two \( x \) nodes linked to each other, then the two \( x \) nodes will fire in synchrony, and the \( y \) and \( z \) nodes will fire at an offset with the \( x \) nodes and with each
other. This synchronous firing indicates that the two \( x \) nodes represent variables bound to the same value. This constitutes the solution to the variable-binding problem.

Proofs are searched for in parallel, and winner-takes-all circuitry suppresses all but the one whose collector nodes have the highest level of activation.

There are complications in this model for such things as managing different predications with the same predicate but different arguments. But the essential idea is as described. In brief, the view of relational information processing implied by shruti is one where reasoning is a transient but systematic propagation of rhythmic activity over structured cell-ensembles, each active entity is a phase in the rhythmic activity, dynamic bindings are represented by the synchronous firing of appropriate nodes, long-term facts are circuits that detect coincidences in the ongoing flux of rhythmic activity, and rules are high-efficacy links that cause the propagation of rhythmic activity between cell-ensembles. Reasoning is the spontaneous outcome of a shruti network.

In the abduction framework, the typical axiom in the knowledge base is of the form

\[
(\forall x, y)[p_1(x, y) \land p_2(x, y) \supset (\exists z)[q_1(x, z) \land q_2(x, z)]]
\]

That is, the top level logical connective will be implication. There may be multiple predications in the antecedent and in the consequent. There may be variables \( x \) that occur in both the antecedent and the consequent, variables \( y \) that occur only in the antecedent, and variables \( z \) that occur only in the consequent. Abduction backchains from predications in consequents of axioms to predications in antecedents. Every step in the search for a proof can be considered an abductive proof where all unproved predications are assumed for a cost. The best proof is the least cost proof.

The implementation of this axiom in shruti requires predication clusters of nodes and axiom clusters of nodes (see Figure 1). A predication cluster has one collector node and one enabler node, both firing asynchronously, corresponding to the predicate and one synchronously firing node for each argument. An axiom cluster has one collector node and one enabler node, both firing asynchronously, recording the plausibility and the utility, respectively, of this axiom participating in the best proof. It also has one synchronously firing node for each variable in the axiom—in our example, nodes for \( x, y \) and \( z \).
The axiom is then encoded in a structure like that shown in Figure 1. There is a predication cluster for each of the predications in the axiom and one axiom cluster that links the predications of the consequent and antecedent. In general, the predication clusters will occur in many axioms; this is why their linkage in a particular axiom must be mediated by an axiom cluster.

Figure 1: Axiom encoded in SHRUTI.

Suppose the proof process is backchaining from the predication $q_1(x, z)$. The activation on the enabler node (?) of the cluster for $q_1(x, z)$ induces an activation on the enabler node for the axiom cluster. This in turn induces activation on the predication nodes for $p_1(x, y)$ and $p_2(x, y)$. Meanwhile the firing of the $x$ node in the $q_1$ cluster induces the $x$ node of the axiom cluster to fire in synchrony with it, which in turn causes the $x$ nodes of the $p_1$ and $p_2$ clusters to fire in synchrony as well. In addition, a link (not shown) from the enabler node of the axiom cluster to the $y$ argument node of the same cluster causes the $y$ argument node to fire, while links (not shown) from the $x$ and $z$ nodes cause that firing to be out of phase with the firing of the $x$ and $z$ nodes. This firing of the $y$ node of the axiom cluster induces synchronous firing in the $y$ nodes of the $p_1$ and $p_2$ clusters.

By this means we have backchained over axiom (4) while keeping distinct
the variables that are bound to different values. We are then ready to
backchain over axioms in which $p_1$ and $p_2$ are in the consequent.

As mentioned above, the $q_1$ cluster is linked to other axioms as well, and
in the course of backchaining, it induces activation in those axioms’ clusters
too. In this way, the search for a proof proceeds in parallel. Inhibitory links
will eventually force a winner-takes-all outcome.

1.8 Incremental Changes to Axioms
In this framework, incremental increases in linguistic competence, and other
knowledge as well, can be achieved by means of a small set of simple oper-
ations on the axioms in the knowledge base.

- The introduction of a new predicate, where the utility of that predicate
can be argued for for cognition in general, independent of language.

- The introduction of a new predicate $p$ specializing an old predicate $q$:

$$\forall x \, p(x) \supset q(x)$$

For example, we learn that a beagle is a type of dog.

$$\forall x \, \text{beagle}(x) \supset \text{dog}(x)$$

- The introduction of a new predicate $p$ generalizing one or more old
predicates $q_i$:

$$\forall x \, q_1(x) \supset p(x), \forall x \, q_2(x) \supset p(x), \ldots$$

For example, we learn that dogs and cats are both mammals.

$$\forall x \, \text{dog}(x) \supset \text{mammal}(x), \forall x \, \text{cat}(x) \supset \text{mammal}(x)$$

- Increasing the arity of a predicate to allow more arguments.

$$p(x) \Rightarrow p(x, y)$$

For example, we learn that “mother” is not a property but a relation.

$$\text{mother}(x) \Rightarrow \text{mother}(x, y)$$

- Adding a literal to the antecedent of an axiom.

$$p_1(x) \supset q(x) \Rightarrow p_1(x) \land p_2(x) \supset q(x)$$
For example, we might first believe that a seat is a chair, then learn that a seat with a back is a chair.

\[ \text{seat}(x) \supset \text{chair}(x) \Rightarrow \text{seat}(x) \land \text{back}(y, x) \supset \text{chair}(x) \]

- Adding a literal to the consequent of an axiom.

\[ p(x) \supset q_1(x) \Rightarrow p(x) \supset q_1(x) \land q_2(x) \]

For example, a child might see snow for the first time and see that it’s white; he then goes outside and realizes it’s also cold.

\[ \text{snow}(x) \supset \text{white}(x) \Rightarrow \text{snow}(x) \supset \text{white}(x) \land \text{cold}(x) \]

All of these incremental changes can be implemented in the SHRUTI model via relatively simple means involving the recruitment of nodes, by strengthening latent connections as a response to frequent simultaneous activations.

These incremental operations can be seen as constituting a plausible mechanism for both the development and the evolution of cognitive capacities. In this paper, I will show how the principal features of language could have resulted from a sequence of such incremental steps, starting from the cognitive capacity one could expect of ordinary primates.

### 1.9 Summary of Background

To summarize, the framework assumed in this chapter has the following features:

- A detailed, plausible, computational model for a large range of linguistic behavior.
- A possible implementation in a connectionist model.
- An incremental model of learning, development, and evolution.
- An implementation of that in terms of node recruitment.

In the remainder of the paper it is shown how two of the principal features of language—Gricean non-natural meaning and syntax—could have arisen from nonlinguistic cognition via

- incremental changes to axioms
• folk theories required independent of language
• compilation of proofs into axioms

These two features are, in a sense, the two key features of language. Gricean meaning tells how single words convey meaning in discourse. Syntax tells how multiple words combine to convey complex meanings.

2 The Evolution of Gricean Meaning

In Gricean non-natural meaning, what is conveyed is not merely the content of the utterance, but also the intention of the speaker to convey that meaning, and the intention of the speaker to convey that meaning by means of that specific utterance. When A shouts “Fire!” to B, A expects that

1. B will believe there is a fire
2. B will believe A wants B to believe there is fire
3. 1 will happen because of 2

Five steps take us from natural meaning, as in “Smoke means fire,” to Gricean meaning (Grice, 1948). Each step depends on certain background theories being in place, theories that are motivated even in the absence of language. Each new step in the progression introduces a new element of defeasibility. The steps are as follows:

1. Smoke means fire
2. “Fire!” means fire
3. Mediation by belief
4. Mediation by intention
5. Full Gricean meaning

2.1 Smoke Means Fire

The first required folk theory is a theory of causality (or rather, a number of theories with causality). There will be no definition of the predicate cause, that is, no set of necessary and sufficient conditions.
Rather there will be a number of domain-dependent theories saying what sorts of things cause what other sorts of things. There will be lots of necessary conditions

\[ \text{cause}(e_1, e_2) \equiv \ldots \]

and lots of sufficient conditions

\[ \ldots \supset \text{cause}(e_1, e_2) \]

An example of the latter type of rule is

\[ \text{smoke}(y) \supset (\exists x)[\text{fire}(x) \land \text{cause}(x, y)] \]

That is, if there’s smoke, there’s fire (that caused it).

This kind of causal knowledge enables prediction, and is required for the most rudimentary intelligent behavior.

Now suppose an agent B sees smoke. In the abductive account of intelligent behavior, an agent interprets the environment by telling the most plausible causal story. Here the story is that since fire causes smoke, there is a fire. B’s seeing smoke causes B to believe there is fire, because B knows fire causes smoke.

2.2 “Fire!” Means Fire

Suppose seeing fire causes another agent A to emit a particular sound, say, “Fire!” and B knows this. Then we are in exactly the same situation as in Step 1. B’s perceiving A making the sound “Fire!” causes B to believe there is a fire. B requires one new axiom about what causes what, but otherwise no new cognitive capabilities.

In this sense, sneezing means pollen, and “Ouch!” means pain.

It is often been stated that one of the true innovations of language is its arbitrariness. The word “fire” is in no way iconic; its relation to fire is arbitrary and purely a matter of convention. The arbitrariness does not seem to me especially remarkable, however. A dog that has been trained to salivate when it hears a bell is responding to an association just as arbitrary as the relation between “fire” and fire.
2.3 Mediation by Belief

For the next step we require a folk theory of belief, that is, a set of axioms explicating, though not necessarily defining, the predicate believe. The principal elements of a folk theory of belief are the following:

- a. An event occurring in an agent’s presence causes the agent to perceive the event.

\[ \text{cause}(at(x, y, t), perceive(x, y, t)) \]

This is only defeasible. Sometimes an individual doesn’t know what’s going on around him.

- b. Perceiving an event causes the agent to believe the event occurred.

\[ \text{cause}(perceive(x, y, t), believe(x, y, t)) \]

(Seeing is believing.)

- c. Beliefs persist.

\[ t_1 < t_2 \supset \text{cause}(\text{believe}(x, y, t_1), \text{believe}(x, y, t_2)) \]

- d. Certain beliefs of an agent can cause certain actions by the agent.

\[ \text{cause}(\text{believe}(x, P, t), \text{ACT}(x, t)) \]

(This is an axiom schema, that can be instantiated in many ways.)

For example, an individual may have the rule that an agent’s believing there is fire causes the agent to utter “Fire!”

\[ \text{fire}(f) \supset \text{cause}(\text{believe}(x, f, t), \text{utter}(x, \text{“Fire!”}, t)) \]

Such a theory would be useful to an agent even in the absence of language, for it provides an explanation of how agents can transmit causality, that is, how an event can happen at one place and time and cause an action that happens at another place and time. It enables an individual to draw inferences about unseen events from the behavior of another individual. Belief functions as a carrier of information.

Such a theory of belief allows a more sophisticated interpretation, or explanation, of an agent A’s utterance, “Fire!” A fire occurred in A’s presence. Thus, A believed there was a fire. Thus, A uttered “Fire!” The link between the event and the utterance is mediated by belief.

In particular, the observable event that needs to be explained is that an agent A uttered “Fire!” and the explanation is as follows:
There may well be other causes of a belief besides seeing. For example, communication with others might cause belief. Thus the above proof could have branched another way below the third line. This fact means that with this innovation, there is the possibility of “language” being cut loose from direct reference.

If an ostrich had a theory of belief and believed that the dominant ostrich’s observation that all was safe caused it to yawn, it would conclude that all was safe (Gordon, this volume). (However, I do not think an ostrich has a theory of belief.)

Note that this theory of belief could in principle be a theory of other individuals, and not a theory of one’s self. There is no need in this analysis that the interpreter even have a concept of self.

2.4 Near-Gricean Non-Natural Meaning
The next step is a close approximation of Gricean meaning. It requires a much richer cognitive model. In particular, three more background folk theories are needed, each again motivated independently of language.

The first is a theory of goals, or intentionality. By adopting a theory that attributes agents’ actions to their goals, one’s ability to predict the actions of other agents is greatly enhanced. The principal elements of a theory of goals are the following:

a. If an agent $x$ has an action by $x$ as a goal, it will cause $x$ to perform this action. This is an axiom schema, instantiated for many different actions.

\[
\text{cause}(\text{goal}(x, \text{ACT}(x)), \text{ACT}(x))
\]
That is, wanting to do something causes an agent to do it. Using this rule in reverse amounts to the attribution of intention. We see someone doing something and we assume they did it because they wanted to do it.

b. If an agent \( x \) has a goal \( g_1 \) and \( g_2 \) tends to cause \( g_1 \), then \( x \) may have \( g_2 \) as a goal.

\[
(6) \quad \text{cause}(g_2, g_1) \supset \text{cause}(\text{goal}(x, g_1), \text{goal}(x, g_2))
\]

This is only a defeasible rule. There may be other ways to achieve the goal \( g_1 \), other than \( g_2 \). This rule corresponds to the body of a STRIPS planning operator as used in AI (Fikes and Nilsson, 1971). When we use this rule in the reverse direction, we are inferring an agent’s ends from the means.

c. If an agent \( A \) has a goal \( g_1 \) and \( g_2 \) enables \( g_1 \), then \( A \) has \( g_2 \) as a goal.

\[
(7) \quad \text{enable}(g_2, g_1) \supset \text{cause}(\text{goal}(x, g_1), \text{goal}(x, g_2))
\]

This rule corresponds to the prerequisites in the STRIPS planning operators.

Many actions are enabled by the agent knowing something. These are knowledge prerequisites. The form of these rules is

\[
\text{enable} (\text{believe}(x, P), \text{ACT}(x))
\]

The structure of goals linked in these ways constitutes a plan. To achieve a goal, one must make all the enabling conditions true and then find an action that will cause the goal to be true, and then do that.

The second required theory is a theory of joint action or collective intentionality. This is the same as a theory of individual intentionality, except that collectives of individuals can have goals and beliefs and can carry out actions. In addition, collective plans must bottom out in individual action. In particular, a group believes a proposition if every member of the group believes it.
Agents can have as goals events that involve other agents. Thus, they can have in their plans knowledge prerequisites for other agents. A can have as a goal that B believe some fact. Communication is the satisfaction of such a goal.

The third theory is a theory of how agents understand. The essential content of this theory is that agents try to fit events into causal chains. The first rule is a kind of causal modus ponens. If an agent believes $e_2$ and believes $e_2$ causes $e_3$, that will cause the agent to believe $e_3$.

$$\text{cause}(\text{believe}(x, e_2) \land \text{believe}(x, \text{cause}(e_2, e_3)), \text{believe}(x, e_3))$$

The second rule allows us to infer that agents backchain on enabling conditions. If an agent believes $e_2$ and believes $e_1$ enables $e_2$, then the agent will believe $e_1$.

$$\text{cause}(\text{believe}(x, e_2) \land \text{believe}(x, \text{enable}(e_1, e_2)), \text{believe}(x, e_1))$$

The third rule allows us to infer that agents do causal abduction. That is, they look for causes of events that they know about. If an agent believes $e_2$ and believes $e_1$ causes $e_2$, then the agent may come to believe $e_1$.

$$\text{cause}(\text{believe}(x, e_2) \land \text{believe}(x, \text{cause}(e_1, e_2)), \text{believe}(x, e_1))$$

This is defeasible since the agent may have beliefs about other possible causes of $e_2$.

The final element of the folk theory of cognition is that all folk theories, including this one, are believed by every individual in the group. It is a corollary of this that A’s uttering “Fire!” may cause B to believe there is a fire.

Now the near-Gricean explanation for the utterance is this: A uttered “Fire!” because A had the goal of uttering “Fire!”, because A had as a goal that B believe there is a fire, because B’s belief is a knowledge prerequisite in some joint action that A has as a goal (perhaps merely joint survival) and because A believes there is a fire, because there was a fire in A’s presence.

## 2.5 Full Gricean Non-Natural Meaning

Only one more step is needed for full Gricean meaning. It must be a part of B’s explanation of A’s utterance not only that A had as a goal that B believe there is a fire and that caused A to have the goal of uttering “Fire!”, but also that A had as a goal that A’s uttering “Fire!” would cause B to believe there is a fire. To accomplish this we must split the planning axiom (6) into two:
b1. If an agent A has a goal \( g_1 \) and \( g_2 \) tends to cause \( g_1 \), then A may have as a goal that \( g_2 \) cause \( g_1 \).

b2. If an agent A has as a goal that \( g_2 \) cause \( g_1 \), then A has the goal \( g_2 \).

The planning axioms (5), (6), and (7) implement means-end analysis. This elaboration captures the intentionality of the means-end relation.

The capacity for language evolved over a long period of time, after and at the same time as a number of other cognitive capacities were evolving. Among the other capacities were theories of causality, belief, intention, understanding, joint action, and (nonlinguistic) communication. As the relevant elements of each of these capacities were acquired, they would have enabled the further development of language as well.

3 The Evolution of Syntax

3.1 The Two-Word Stage

When agents encounter two objects in the world that are adjacent, they need to explain this adjacency by finding a relation between the objects. Usually, the explanation for why something is where it is is that that is its normal place. It is normal to see a chair at a desk, and we don’t ask for further explanation. But if something is out of place, we do. If we walk into a room and see a chair on a table, or we walk into a lecture hall and see a dog in the aisle, we wonder why.

Similarly, when agents hear two adjacent utterances, they need to explain the adjacency by finding a relation between them. A variety of relations are possible. “Mommy sock” might mean “This is Mommy’s sock” and it might mean “Mommy, put my sock on”.

In general, the problem facing the agent can be characterized by the following pattern:

\[
(\forall w_1, w_2, x, y, z) B(w_1, y) \land C(w_2, z) \land \text{rel}(x, y, z) \supset A(w_1 w_2, x)
\]

That is, to recognize two adjacent words or strings of words \( w_1 \) and \( w_2 \) as a composite utterance of type \( A \) meaning \( x \), one must recognize \( w_1 \) as an object of type \( B \) meaning \( y \), recognize \( w_2 \) as an object of type \( C \) meaning \( z \), and find some relation between \( y \) and \( z \), where \( x \) is determined by the relation that is found.
This is the characterization of what Bickerton calls proto-language. One utters meaningful elements sequentially and the interpretation of the combination is determined by context. The utterance “Lion. Tree.” could mean there’s a lion behind the tree or there’s a lion nearby so let’s climb that tree, or numerous other things. Bickerton gives several examples, including the language of children in the two-word phase and the language of apes. I’ll offer another example: the language of panic. If a man runs out of his office shouting, “Help! Heart attack! John! My office! CPR! Just sitting there! 911! Help! Floor! Heart attack!” we don’t need syntax to tell us that he was just sitting in his office with John when John had a heart attack, and John is now on the floor, and the man wants someone to call 911 and someone to apply CPR.

Most if not all rules of grammar can be seen as specializations and elaborations of pattern (8).

The simplest example in English is compound nominals. To understand “turpentine jar” one must understand “turpentine” and “jar” and find the most plausible relation (in context) between turpentine and jars. In fact, compound nominals can be viewed as a relic of protolanguage in full-blown language.

Often with compound nominals the most plausible relation is a predicate-argument relation, where the head noun supplies the predicate and the prenominal noun supplies an argument. In “chemistry teacher”, a teacher is a teacher of something, and the word “chemistry” tells us what that something is. In “language origin”, something is originating, and the word “language” tells us what that something is.

The two-word utterance “Men work” can be viewed in the same way. We must find a relation between the two words to explain their adjacency. The relation we find is the predicate-argument relation, where “work” is the predicate and “men” is the argument.

The phrase structure rules

\[ S \rightarrow NP \ VP; \ VP \rightarrow V \ NP \]

can be written in the abductive framework (Hobbs et al., 1993) as

\[ (\forall w_1, w_2, x, e) \ Syn(w_1, x) \land \ Syn(w_2, e) \land \ Lsubj(x, e) \supset \ Syn(w_1w_2, e) \]

\[ (\forall w_3, w_4, y, e) \ Syn(w_3, e) \land \ Syn(w_4, y) \land \ Lobj(y, e) \supset \ Syn(w_3w_4, e) \]

21
In the first rule, if \( w_1 \) is string of words describing an entity \( x \) and \( w_2 \) is a string of words describing the eventualty \( e \) and \( x \) is the logical subject of \( e \), then the concatenation \( w_1w_2 \) of the two strings can be used to describe \( e \). This means that to interpret \( w_1w_2 \) as describing some eventualty \( e \), segment it into a string \( w_1 \) describing the logical subject of \( e \) and a string \( w_2 \) providing the rest of the information about \( e \). The second rule is similar.

These axioms instantiate pattern (8). The predicate \( \text{Syn} \), which relates strings of words to the entities and situations they describe, plays the role of \( A \) in pattern (8), and the relation \( \text{rel} \) in pattern (8) is instantiated by the \( \text{Lsubj} \) and \( \text{Lobj} \) relations.

Syntax in general can be viewed as a set of constraints on the interpretation of adjacency, specifically, as predicate-argument relations.

Rule (9) is not sufficiently constrained, since \( w_2 \) could already contain the subject. We can prevent this by adding to the arity of \( \text{Syn} \), one of the incremental evolutionary modifications in rules, and giving \( \text{Syn} \) a further argument indicating that something is missing.

\[
(\forall w_1, w_2, x, e)\text{Syn}(w_1, x, - , -) \land \text{Syn}(w_2, e, x, -) \land \text{Lsubj}(x, e)
\]

\[
\supset \text{Syn}(w_1w_2, e, - , -)
\]

\[
(\forall w_3, w_4, y, e)\text{Syn}(w_3, e, x, y) \land \text{Syn}(w_4, y, - , -) \land \text{Lobj}(y, e)
\]

\[
\supset \text{Syn}(w_3w_4, e, x, -)
\]

Now the expression \( \text{Syn}(w_3, e, x, y) \) says something like “String \( w_3 \) would describe situation \( e \) if strings of words describing \( x \) and \( y \) can be found in the right places.”

But when we restructure the axioms like this, the \( \text{Lsubj} \) and \( \text{Lobj} \) are no longer needed where they are, because the \( x \) and \( y \) arguments are now available at the lexical level. We can add axioms linking predicates in the knowledge base with words in the language. We then have following rules, where the lexical axiom is illustrative.

\[
(\forall w_1, w_2, x, e)\text{Syn}(w_1, x, - , -) \land \text{Syn}(w_2, e, x, -)
\]

\[
\supset \text{Syn}(w_1w_2, e, - , -)
\]

\[
(\forall w_3, w_4, y, e)\text{Syn}(w_3, e, x, y) \land \text{Syn}(w_4, y, - , -)
\]

\[
\supset \text{Syn}(w_3w_4, e, x, -)
\]

\[
\text{read}(e, x, y) \land \text{text}(y) \supset \text{Syn}(\text{"read"}, e, x, y)
\]

This is the form of the rules shown in Section 1.

We can add three more arguments to incorporate part-of-speech, agreement, and subcategorization constraints. A rather extensive account of English syntax in this framework, similar to that in Pollard and Sag (1994), is given in Hobbs (1998).
This account must be modified somewhat. Metonymy is a pervasive characteristic of discourse. When we say

I’ve read Shakespeare.

we coerce “Shakespeare” into something that can be read, namely, the writings of Shakespeare. So syntax is a set of constraints on the interpretation of adjacency as predicate-argument relations plus metonymy. This is probably not a recent development in the evolution of language. Rather it is the most natural starting point for syntax. In many protolanguage utterances, the relation found between adjacent elements involves just such a metonymic coercion.

In multiword discourse, when a relation is found to link two words or larger segments into a composite unit, it too can be related to adjacent segments in various ways. The tree structure of sentences arises out of this recursion.

The competitive advantage this development confers is clear. There is less ambiguity in utterances and therefore more precision, and therefore more complex messages can be constructed. People can thereby engage in more complex joint action.

3.2 Signalling Predication and Modification

The languages of the world signal predication primarily by means of position and particles (or affixes). They signal modification primarily by means of adjacency and various concord phenomena. Signalling predication by postpositions, as does Japanese, can be captured in axioms, specializing and elaborating pattern (8) and similar to (11), as follows:

\[
\begin{align*}
\text{Syn}(w_1, x, n, -) \land \text{Syn}(w_2, e, p, x) & \supset \text{Syn}(w_1 w_2, e, p, -) \\
\text{Syn}(w_3, e, p, -) \land \text{Syn}(w_4, e, v, -) & \supset \text{Syn}(w_1 w_2, e, v, -) \\
\text{from}(e, x) & \supset \text{Syn}(\text{“kara”}, e, p, x) \\
\text{go}(e) & \supset \text{Syn}(\text{“iki”}, e, v, -)
\end{align*}
\]

The first rule combines a noun phrase and a particle into a particle phrase. The second rule combines a particle phrase and a verb into a clause, and permits multiple particle phrases to be combined with the verb. The two lexical axioms link Japanese words with underlying world-knowledge predicates.\(^3\) The fourth rule generates a logical form for the verb specifying the type

---

\(^3\)Apologies for using English as the language of thought in this example.
of event it describes. The third rule links that event with the arguments described by the noun phrases via the relation specified by the particle.

The other means of signalling predication and modification can be represented similarly.

3.3 Discontinuous Elements

Consider

John is likely to go.

To interpret this, an agent must find a relation between “John” and “is likely”. Syntax says that it should be a predicate-argument relation plus metonymy. The predicate “is likely” requires a proposition as its argument, so we must coerce “John” into a proposition. The next phrase “to go” provides the required coercion function. John’s going is likely.

Formally, this link between the subject and the infinitive can be mediated by modifying the VP rule slightly, enabling it to pass information between the subject and the infinitival complement. The VP rule for such “raising” constructions in the framework presented here is

\[
\text{Syn}(w_1, e, e_1, -) \land \text{Syn}(w_2, e_1, x, -) \supset \text{Syn}(w_1w_2, e, x, -)
\]

That is, if a string \(w_1\) (“is likely”) describing a situation \(e\) and looking for a logical subject referring to \(e_1\) (John’s going) is concatenated with a string \(w_2\) (“to go”) describing \(e_1\) and looking for a subject \(x\) (John), then the result describes the situation \(e\) provided we can find a logical subject describing \(x\).

In the approach taken here, as well as in unification grammar approaches such as that of Pollard and Sag (1994), information is passed between distant parts of a sentence by means of variables, which our structured connectionist model accommodates naturally, rather than by a “move α” transformation, as in government and binding approaches, where a possible underlying mechanism is quite mysterious.

3.4 Long-Distance Dependencies

One of the most “advanced” and probably one of the latest universal phenomena of language is long-distance dependencies, as illustrated by relative clauses and wh-questions. They are called long-distance dependencies because in principal the head noun can be an argument of a predications that is embedded arbitrarily deeply. In the noun phrase
the man John believes Mary said Bill saw

the man is the logical object of the seeing event, at the third level of embedding.

In accounting for the evolution of long-distance dependencies, we will take our cue from the Japanese. It has been argued that the Japanese relative clause is as free as the English compound nominal in its interpretation. All that is required is that there be some relation between the situation described by the relative clause and the entity described by the head noun (Akmajian and Kitagawa, 1974; Kameyama, 1994). They cite the following noun phrase as an example.

Hanako ga iede shita Taroo

Hanako Subj run-away-from-home did Taroo

Taroo such that Hanako ran away from home

Here it is up to the interpreter to find some plausible relation between Taroo and Hanako’s running away from home.

We may take Japanese as an example of the basic case. Any relation will explain the adjacency of the relative clause and the noun. In English, a further constraint is added, analogous to the constraint between subject and verb. The relation must be the predicate-argument relation, where the head noun is the argument and the predicate is provided, roughly, by the top-level assertion in the relative clause and its successive clausal complements. Thus, in “the man who John saw”, the relation between the man and the seeing event is the predicate-argument relation—the man is the logical object of the seeing. It is thus a specialization of pattern (8), and a constraint on the interpretation of the relation. The constraints in French relative clauses lie somewhere between those of English and Japanese.

The English case can be incorporated into the grammar by increasing the arity of the Syn predicate, relating strings of words to their meaning. Before we had arguments for the string, the entity or situation it described, and the missing logical subject and object. We will increase the arity by one, and add an argument for the entity that will fill the gap in relative clause. The rules for relative clauses then becomes

\[ \text{Syn}(w_1, e_1, x, y, -) \land \text{Syn}(\text{"\text{who}"}, y, - , - , - ) \supset \text{Syn}(w_1 w_2, e_1, x, - , y) \]
\[ \text{Syn}(w_1, x, - , - , - ) \land \text{Syn}(w_2, e, - , - , x) \supset \text{Syn}(w_1 w_2, x, - , - , - ) \]

The first rule introduces the gap. It says an eventuality \(e_1\) looking for its logical object \(y\) can concatenate with the empty string provided the gap is
eventually matched with a head describing $y$. The second rule says, roughly, that a head noun $w_1$ describing $x$ can concatenate with a relative clause $w_2$ describing $e$ but having a gap $x$ to form a string $w_1w_2$ that describes $x$. The rare reader interested in seeing the details of this treatment should consult Hobbs (1998).

Sportscaster English, where “which” is used as a subordinate conjunction, can be seen as a relaxation of this constraint back to the protolanguage pattern of composition.

Seeking a relation between adjacent or proximate words or larger segments in an utterance is simply an instance of seeking explanations for the observables in our environment, specifically, observable relations. Syntax can be seen largely as a set of constraints on such interpretations, primarily constraining the relation to predicate-argument relations. The changes are of three kinds, the first two of which we have discussed.

- Specializing predicates that characterize strings of words, as the predicate $Syn$ specializes the predicates in pattern (8).
- Increasing the arity of the $Syn$ predicate, i.e., adding arguments, to transmit arguments from one part of a sentence to another.
- Adding predications to antecedents of rules to capture agreement and subcategorization constraints.

The acquisition of syntax, whether in evolution or in development, can be seen as the accumulation of such constraints.

4 Remarks on the Course of the Evolution of Language

Relevant dates in the time course of the evolution of language and language readiness are as follows:

1. Mammalian dominance: c65-50M years ago
2. Common ancestor of monkeys and great apes: c15M years ago
3. Common ancestor of hominids and chimpanzees: c5M years ago
4. Appearance of Homo erectus: c1.5M years ago
5. Appearance of Homo sapiens sapiens: c200-150K years ago
6. African/non-African split: c90K years ago
7. Appearance of preserved symbolic artifacts: c70-40K years ago
8. Time depth of language reconstruction: c10K years ago
9. Historical evidence: c5K years ago

In this section I will speculate about the times at which various components of language, as explicated in this chapter, evolved. I will then discuss three issues that have some prominence in this volume:

1. Which came first—gesture or voice? This is a question about hominid evolution, somewhere between the last common ancestor with chimpanzees and the appearance of Homo sapiens.

2. Was there a holophrastic stage before full-blown language? This is a question, probably, about the period just before the evolution of Homo sapiens sapiens.

3. When Homo sapiens sapiens evolved, did they have full-blown language or merely language readiness? This is a question about the period between the evolution of Homo sapiens sapiens and the appearance of preserved material culture.

4.1 Evolution of the Components of Language

Language is generally thought of as having three parts: phonology, syntax, and semantics. Language maps between sound (phonology) and meaning (semantics), and syntax provides the means for composing elementary mappings into complex mappings.

The evolution of the components of language are illustrated in Figure 2. According to Arbib (this volume), gestural communication led to vocal communication, which is phonology. Also according to Arbib, the ability to understand composite event structure was a necessary precursor to protolanguage (as was, of course, vocal communication). Protolanguage led to syntax. According to the account in Section 2 of this chapter, the use of causal associations was the first requirement for the development of semantics. Causal association is possible in a brain that does the equivalent of propositional logic (such as most current neural models), but before one can
have a theory of mind, one must have the equivalent of first-order logic. A
creature must be able to distinguish different tokens of the same type. The
last requirement is the development of a theory of mind, including models
of belief, mutual belief, goals, and plans.

Arbib lists a number of features of language that have to evolve before we
can say that full-blown language has evolved. It is useful to point out what
elements in Figure 2 are necessary to support each of these features. Naming
requires only causal association. A causal link must be set up between a
sound and a physical entity in the world. Parity between comprehension and
production requires a theory of the mind of others. The “Beyond Here and
Now” feature also requires a theory of mind; one function of belief, it was
pointed out, is to transmit causality to other times and places. Hierarchical
structure first appears with composite event structure. Once there is pro-
tolanguage, there is a lexicon in the true sense. The significance of temporal
order of elements in a message begins somewhere between the development
of protolanguage and real syntax. Learnability is there from start to finish,
in every stage.

Causal associations are possible from at least the earliest stages of mul-
ticellular life. A leech that moves up a heat gradient and attempts to bite
when it encounters an object is responding to a causal regularity in the world. Of course, it does not know that it is responding the causal regularities; that would require a theory of mind. But the causal associations themselves are very early. The naming that this capability enables is quite within the capability of parrots, for example. Thus, in Figure 2, we can say that causal association is pre-mammalian.

At what point are animals aware of different types of the same token? That is, at what point are they first-order? My purely speculative guess would be that it happens early in mammalian evolution. Reptiles and birds have an automaton-like quality associated with propositional representations, but most mammals that I am at all familiar with, across a wide range of genera, exhibit a flexibility of behavior that would require different responses to different tokens of the same type.

One reason to be excited about the discovery of the mirror neuron system (Rizzolatti and Arbib, 1998) is that it is evidence of an internal representation “language” that abstracts away from a concept’s role in perception or action, and thus is possibly the earliest solid indication of “first-order” features in the evolution of the brain.

Gestural communication, composite event structure, and a theory of mind probably appear somewhere between the separation of great apes and monkeys, and the first hominids, between 15 and 5 million years ago. Arbib discusses the recognition and repetition of composite events. There are numerous studies of the gestural communication that the great apes can perform. The evolution of the theory of mind is very controversial (e.g., Heyes, 1998), but it has certainly been argued that chimpanzees have it. These three features can thus probably be assigned to the pre-hominid era.

My, again purely speculative, guess would be that vocal communication emerged with Homo erectus, and I would furthermore guess that they were capable of protolanguage—that is, stringing together a few words or signals to convey novel though not very complex messages. It’s impossible to speculate how large their lexicon would have been.

Finally, full-blown language probably emerged simultaneously with Homo sapiens sapiens, and is what gave us a competitive advantage over our hominid cousins. We were able to construct more complex messages and therefore were able to carry out more complex joint action. As Dunbar (1996) has argued, full-blown language would have allowed us to maintain much larger social groups, a distinct evolutionary advantage.
4.2 Gesture Before Vocal Communication?

There are a number of false questions that have been posed in speculations on the evolution of language that have resulted from granularity mismatches. For example, I have heard people argue about whether language for hunting or language for social networks came first, and provided the impetus for language evolution. (We can think of these positions as the Mars and Venus theories.) But language capabilities in a species evolves over hundreds or thousands of generations, whereas hunting and currying social networks are daily activities. It is ludicrous to think that there was a time when some form of language precursor was used for hunting but not for social networking, or vice versa. The obvious truth is that language is for establishing mutual belief, enabling joint action, and that would be a distinct advantage for both hunting and for building social networks.

A similar false question is “Which came first—gesture or voice?” Monkeys have cries and they have gestures. Each modality, as Byrd (this volume) points out, has its advantages over the other. In some situations gestural communication would have been the most appropriate, and in others vocal communication. It is ludicrous to imagine that there was a stage in hominid evolution when individuals sat quietly and communicated to each other by gesture, or a stage when they sat with their arms inert at their sides and chattered with each other. Gesture and voice have always both been there.

As Arbib points out, language developed in the region of the brain that had originally been associated with gesture and hand manipulations. In that sense gesture has a certain primacy. The most likely scenario is that there was a stage when manual gestures were the more expressive system. Articulatory gestures co-opted that region of the brain, and eventually became a more expressive system than the manual gestures, for all the reasons Byrd points out.

4.3 A Holophrastic Stage?

Wray (1998) and Zurow (this volume) argue that there was a holophrastic stage in the evolution of language, and Arbib in his chapter mentions it as well. First there were utterances of one word. These words became more and more complex as the lexicon expanded, and they described more and more complex situations. This is the holophrastic stage. Then these words were analysed into parts, which became the constituents of phrases. It’s as though they had a words “lionattack” and “liondead”, and then decided
that there should be a word “lion” describing the common part of these two situations.

A holophrastic stage has sometimes been hypothesized in child language. Children goes through a one-word stage followed by a two-word stage. The holophrastic stage would be between these two. The evidence is from “words” like “allgone”, “whazzat”, and “gimme”. An alternative explanation for these holophrases is that the child has failed to segment the string, due to insufficient segmentation ability, insufficient vocabulary, and so on. For a holophrastic phrase to exist, we would have to show that such holophrases don’t occur in the one-word stage, and it is completely implausible that such misanalyses don’t occur then.

In any case, one has to be careful about inference from child language to language evolution. There are a couple of major differences. First, adult hominids always had to make a living. So for example, it is extremely implausible that there was a babbling stage in language evolution. Babbling is a child’s way of practicing the sounds of language, but there was never an era in human evolution when that, by itself, would have conferred any advantage.

The other difference is that children have models whose language is substantially in advance of their own. That was never the case in language evolution. Holophrasis in child language is a misanalysis. There was nothing for holophrasis in language evolution to be a misanalysis of.

The coherence structure of discourse provides a more compelling account of the evolution of the sentence. Discourse and interaction precede language. Events in the world and in discourse cohere because they stand in coherence relations with each other. Among the relations are causality,

“Smoke. Fire.”
novel, or similarity,

I signal that I go around to the right. I signal that you go around to the left.
ground-figure,

“Bushes. Tiger.”

occasion, or the next step in the process,

You hand me grain. I grind it.

“Youhand me grain. I grind it.”

“Approach antelope. Throw spear.”
and the predicate-argument or argument-predicate relation,

“Sock. On.”
“Antelope. Kill.”
I point to myself. I point to the right.

It is much more likely that the road to syntax was via coherence relations between successive one-word utterances, as described in Section 3, rather than via holophrasis. The holophrastic account requires some mysterious implicit derivational morphology that just happened to underlie lexical items. The coherence account requires no new mechanisms. It is just a matter of adding constraints on the interpretation of temporal order as indicating predicate-argument relations. Construction is more plausible than deconstruction. It is more plausible that “Lion attack.” would derive from “Lion. Attack.” than from “Lionattack.”

In fact, what does often precede the two-word stage in children is what has been called “vertical constructions” (Scollon, 1979). Children convey a two-concept message by successive one-word utterances, each with sentence intonation, and often with some time and some interaction between them. Hoff (2001, p. 210) quotes a child near the end of the one-word stage saying, “Ow. Eye.” Scollon reports a similar sequence: “Car. Go.”

Jackendoff (1999) points out that one of the important advances leading to language was the analysis of words into individual phonemes. It is much more plausible that this happened by some sort of holophrastic process. We first have unanalyzed utterances “pig” and “pit” and we then analyze them into the sounds of p, i, g, and t, and realize that further words can be built out of these elements. There are two reasons this account is more plausible than the holophrastic account of the evolution of sentences. The system is much simpler, having many fewer elements, and phonemes have no semantics to overconstrain decompositions, as words do.

4.4 Language or Language Readiness?

Arbib (this volume) expresses his belief that the first physically modern Homo sapiens sapiens did not have language, only language readiness. This is a not uncommon opinion. In most such accounts, language is a cultural development that happened with the appearance of preserved symbolic artifacts, and the date one most often hears is around thirty-five to fifty thousand years ago. If we had had language before that, we would have left symbolic relics before that.
One problem with this argument is that it is one bit of graffiti away from refutation. And in fact in June 2002 a symbolic relic was found in South Africa, a rock with cross-hatched markings on it with no apparent nonsymbolic utility, and it was dated to 70,000 years ago. Since then several other symbolic artifacts of similar age have been discovered.

But Arbib, in his concept of language readiness, has given us a way to understand what the real situation was. Symbolic culture may not have happened until 70,000 years ago, but from the beginning we had culture readiness. Language is not identical to symbolic culture. Rather it is a component of culture readiness.

Another argument against Arbib’s position comes from language universals. In some scholarly communities it is fashionable to emphasize how few language universals there are; in others the opposite is the case. But it is certainly true that the following are universal:

- All languages encode predicate-argument relations and assertion-modification distinctions by means of word order and/or particles/inflection.

- All languages have verbs, nouns, and other words.

- Many words have associated, grammatically realized nuances of meaning, like tense, aspect, number, and gender, and in every language verbs are the most highly developed in this regard, followed by nouns, followed by the other words.

- All languages have relative clauses.

- All languages have anaphoric expressions.

Modern humans first left Africa between 100,000 and 90,000 years ago, and began to spread over Eurasia. The universal features of language may seem to us to be inevitable, but we know from formal language theory and logic that information can be conveyed in a very wide variety of ways. It is unlikely that distant groups not in contact would have evolved language in precisely the same way. That means that the language universals were almost surely characteristic of the language of early Homo sapiens sapiens, before the African/non-African split.

There are some features of language that may indeed be a cultural development. These are features that, though widespread, are not universal, and tend to exhibit aereal patterns. Among these features, I would include such things as gender, shape classifiers, plurals, and definiteness.
they have evolved among different distant groups, each area tends to have a distinct character. There are phenomena in the languages of East Asia and of West Africa that are called shape classifiers, but these two phenomena look quite different.

There are also areas of language that are quite clearly relatively recent cultural inventions. These include the grammar of numbers, of clock and calendar terms, and of personal names, and the language of mathematics. These tend to have a very different character than we see in the older parts of language; they tend to be of a simpler, more regular structure.

If language were more recent than the African/non-African split, we would expect to see a great many features that only African languages have and a great many features that only non-African languages have. If, for example, only African languages had relative clauses, or if all African languages were VSO while all non-African languages were SVO, then we could argue that they must have evolved separately, and more recently than 90,000 years ago. But in fact nothing of the sort is the case. There are very few phenomena that occur only in African languages, and they are not widespread even in Africa, and are rather peripheral features of language; among these very few features are clicks in the phonology and logophoric pronouns. There are also very few features that occur only in non-African languages. Object-initial word order is one of them. These features are also not very widespread.

Finally, if language were a cultural achievement within the last 50,000 years, rather than a biological achievement, we would expect to see significant developments in language in the era that we have more immediate access to, the last five or ten thousand years. For example, it might be that languages were becoming more efficient, more learnable, or more expressive in historical times. As a native English speaker, I might cite a trend from inflection and case markings to encode predicate-argument relations to word order for the same purpose. But in fact linguists detect no such trend. Within the last century there have been numerous discoveries of relatively isolated groups with a more primitive culture. There have been no discoveries of isolated groups with a more primitive language.

All of these arguments should lead us to reject the possibility that language is a recent cultural acquisition. It has very likely been, more than anything else, what made us human right from the beginning of the history of our species.
Acknowledgments

This chapter is an expansion of a talk I gave at the Meeting of the Language Origins Society in Berkeley, California, in July 1994. The original key ideas arose out of discussions I had with Jon Oberlander, Mark Johnson, Megumi Kameyama, and Ivan Sag. I have profitted more recently from discussions with Lokendra Shastri, Chris Culy, Cynthia Hagstrom, and Srin Narayanan, and with Michael Arbib, Dani Byrd, Andrew Gordon, and the other members of Michael Arbib’s language evolution study group. None of these people would necessarily agree with anything I have said.

References


