THE ROYAL SOCIETY

doi 10.1098/rstb.2000.0723

Evolutionary biology of language

Martin A. Nowak

Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, USA (nowak@ias.edu)

Language is the most important evolutionary invention of the last few million years. It was an adaptation that helped our species to exchange information, make plans, express new ideas and totally change the appearance of the planet. How human language evolved from animal communication is one of the most challenging questions for evolutionary biology. The aim of this paper is to outline the major principles that guided language evolution in terms of mathematical models of evolutionary dynamics and game theory. I will discuss how natural selection can lead to the emergence of arbitrary signs, the formation of words and syntactic communication.

Keywords: evolution; game theory; animal communication; human language

1. INTRODUCTION

Everyone who reads this paper knows of the order of 50 000 words of his primary language. These words are stored in the 'mental lexicon' together with one or several meanings, and some information about how they relate to other words and how they fit into sentences. During the first 16 years of life we learn about one new word every 90 min of waking time: a six year old knows about 13 000 words (Nagy & Anderson 1984; Miller 1991; Pinker 1994, 1999; Nagy *et al.* 1993).

Words are strings of phonemes. Sentences are strings of words. Language makes use of combinatorics on two levels. This is what linguists call 'duality of patterning'. While words have to be learned, virtually every sentence that a person utters is a brand-new combination. The brain contains a programme that can build an unlimited number of sentences out of a finite list of words. This programme is called 'mental grammar' (Jackendoff 1997). Children develop this grammar rapidly and without formal instruction. Experimental observations show that three-year-old children are already on target 90% of times when applying grammatical rules.

The most complicated mechanical motion that the human body can perform is the activity of speaking. While generating the sounds of spoken language, the various parts of the vocal tract perform movements that have to be accurate within millimetres and synchronized to within a few hundredths of a second (Miller 1981).

Speech perception is another biological miracle of our language faculty. The auditory system is so well adapted to speech that we can understand 10–15 phonemes per second during casual speech and up to 50 phonemes per second in artificially speeded-up speech. These numbers are surprising given the physical limitations of our auditory system: if a sound like a click is repeated at a rate of about $20\,s^{-1}$, we no longer hear it as a sequence of separate sounds, but as a continuous buzz. Hence we apparently do not perceive phonemes as consecutive bits of sound, but each moment of spoken sound must have several phonemes packed into it, and our brain knows how to unzip them (Miller 1967; Liberman *et al.* 1967; Cole & Jakimik 1980).

The preceding paragraphs show that human language is an enormously complex trait. Our language performance relies on precisely coordinated interactions of various parts of our neural and other anatomy, and we are amazingly good at it. We can all speak without thinking. In contrast, we cannot perform basic mathematical operations without concentration. Why is doing mathematics or playing chess painfully difficult for most of us, when the computational tasks necessary for generating or interpreting language are arguably more complicated? A plausible answer is that evolution designed some parts of our brain specifically for dealing with language.

Worldwide there are about 6000 different human languages. Among all of these there is no 'simple' human language. There may have been Stone Age societies, but there was no Stone Age language. As Edward Sapir wrote: 'When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam.' The ubiquity of complex language is for some linguists compelling evidence that language is not a cultural invention, but an innate instinct. Noam Chomsky, who revolutionized linguistic research, argues that all human languages have the same underlying 'universal grammar' and that this universal grammar is the product of an innate 'language organ' (which need not be seen as a localized organ but an ensemble of language-specific neuronal circuitry within the human brain).

For Chomsky the necessity for innate components of language also comes from what he calls 'the poverty of input'. Children learn the grammatical rules of their native language by hearing a small subset of correct sentences. Since the number of possible rule systems is unlimited, how could they deduce the correct rule system without any preformed expectation that restricts the number of possibilities? According to Chomsky (1965, 1972), children have an innate expectation of universal grammar. Their learning of the grammar is given by the

task of acquiring certain parameters of their particular language while the principles are innate.

The observation that only humans but no animals have complex language led Chomsky (1972, 1988) and others to question how language could have arisen by Darwinian evolution (Bickerton 1990). Perspectives on language evolution have ranged from language being the by-product of a big brain, or language being the consequence of a single dramatic mutation, all the way to language being incompatible with Darwinian evolution. Such views are surprising for evolutionary biologists who would argue that gradual adaptation is the only mechanism by which biology can build a trait as complex as human language (Pinker & Bloom 1990; Newmeyer 1991; Brandon & Hornstein 1986; Corballis 1991). The observation that primates, who are our closest living relatives, apparently do not have complex language does not contradict its evolution. Instead, the implication is that complex language must have originated in our ancestral lines after the separation from chimpanzees, that is in the last seven million years. Thus evolution had about 350 000 generations to build our language instinct from material that was already present in our ancestor species at that time.

For Darwin, there was no question that human language originated from animal communication. He was also fascinated by the analogy between human languages and biological species. In his book, *The descent of man* (1871), he writes: 'The formation of different languages and of distinct species, and the proofs that both have developed through a gradual process, are curiously the same.... Dominant languages and dialects spread widely and lead to gradual extinction of other tongues.... A language, like a species, when once extinct never reappears.'

Evolution relies on the transfer of information from one generation to the next. For billions of years this process was limited to the transfer of genetic information. Language facilitates the transfer of non-genetic information and thus leads to a new mode of evolution. Therefore the emergence of language can be seen as a major transition in evolutionary history (Maynard Smith & Szathmáry 1995, 1999), being of comparable importance to the origin of genetic replication, the first cells, or the emergence of multicellular organisms.

Attempts to shed light on the evolution of language have come from many areas, including studies of primate social behaviour (Seyfarth *et al.* 1980; Burling 1993; Cheney & Seyfarth 1990) or animal communication (Von Frisch 1967; Hauser 1996; Smith 1977), the diversity of existing human languages (Greenberg 1971; Cavalli-Sforza & Cavalli-Sforza 1995), the development of language in children (Newport 1990; Bates 1992; Hurford 1991), the genetic and anatomical correlates of language competence (Lieberman 1984; Nobre *et al.* 1994; Aboitiz & Garcia 1997; Hutsler & Gazzaniga 1997; Deacon 1997; Gopnik & Crago 1991), theoretical studies of cultural evolution (Cavalli-Sforza & Feldman 1981; Yasuda *et al.* 1974; Aoki & Feldman 1989; Cavalli-Sforza 1997) and learning theory (Niyogi & Berwick 1996; Niyogi 1998). Our objective here and in several related papers (Nowak & Krakauer 1999; Nowak 2000; Nowak *et al.* 1999*a*,*b*, 2000; Grassly *et al.* 2000; Krakauer 2000) is to bring discussions of language evolution within the precise mathematical framework of modern evolutionary biology. For mathematical models of language evolution see also Hurford (1989), Hurford *et al.* (1998), and Steels (1997).

In § 2, we describe how evolution can design a very basic communication system where arbitrary signals refer to specific objects (or concepts) of the world. We study the effect or errors during language acquisition. In § 3, we study errors during communication and show how such errors limit the repertoire of a simple communication system. In § 4, we show that word formation can overcome this error limit. In § 5, we design a framework for the population dynamics of words, define the basic reproductive ratio of words and calculate the maximum size of a lexicon. § 6 outlines how natural selection can guide the emergence of syntactic communication and § 7 is a conclusion.

2. THE EVOLUTION OF THE SIMPLEST COMMUNICATION SYSTEM

Let us first study the basic requirements for the evolution of the simplest possible communication system. We imagine a group of individuals (humans or other animals) using a number of arbitrary signals to communicate information about a number of objects (or concepts) of their perceived world. We will define an association matrix, a pay-off function, and finally study evolutionary dynamics.

(a) *The association matrix*

Suppose that the communicative behaviour of each individual is characterized by an association matrix, A. If there are n objects and m signals, then A is an $n \times m$ matrix. The entries, a_{ij}, can be non-negative real numbers (or integers) and denote the strength of the association between object i and signal j.

A speaker is described by a P matrix. The element p_{ij} denotes the probability of using signal j for object i. A hearer is described by a Q matrix. The element q_{ij} denotes the probability of interpreting signal j as referring to object i. The P and Q matrices are derived from the A matrix by normalizing rows and columns respectively. We have

$$p_{ij} = a_{ij} \Big/ \sum_{j=1}^{m} a_{ij}, \qquad q_{ij} = a_{ij} \Big/ \sum_{i=1}^{n} a_{ij}. \tag{1}$$

The denominators have to be greater than zero; otherwise simply take p_{ij} or q_{ij} as zero. The assumption here is that both speaking and hearing are derived from the same association matrix.

(b) *A pay-off function*

Let us now consider two individuals I and J with association matrices A_I and A_J. We can define the pay-off for I communicating with J as

$$F(A_I, A_J) = (1/2) \sum_{i=1}^{n} \sum_{j=1}^{m} (p_{ij}^{(I)} q_{ji}^{(J)} + p_{ij}^{(J)} q_{ij}^{(I)}). \tag{2}$$

The term $\sum_{j=1}^{m} p_{ij}^{(I)} q_{ij}^{(\mathcal{J})}$ denotes the probability that individual I will successfully communicate 'object i' to individual $\mathcal{J}$. This probability is then summed over all objects and averaged over the situation where individual I signals to individual $\mathcal{J}$ and vice versa. Note that equation (2) also assumes that communication about each object occurs with the same frequency.

If both individuals use the same matrix A then

$$F(A, A) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} q_{ij}. \tag{3}$$

The maximum pay-off is obtained if P is a permutation matrix (that is, it has exactly one entry of unity per row and column, all other entries being zero) and $P = Q$. The maximum pay-off is the smaller of n or m (Trapa & Nowak 2000).

Intuitively, our pay-off function characterizes the total amount of information that can be exchanged between two individuals. For each correct information transfer, both speaker and listener obtain a pay-off of one point. Other assumptions for the pay-off function are possible. In particular, it need not be the case that communication is of advantage to both speaker and listener. An alarm call, for example, may be of benefit to the receiver, but may even constitute a cost for the sender. The opposite extreme is if a sender uses a signal to deceive or manipulate another individual. Then the positive pay-off will be with the sender, while the receiver gets a negative pay-off. Thus equation (2) can be generalized to include the possibility that communication about different objects leads to different pay-off values for sender and receiver:

$$F(A_I, A_{\mathcal{J}}) = (1/2) \sum_{i=1}^{n} \sum_{j=1}^{m} (\alpha_i p_{ij}^{(I)} q_{ij}^{(\mathcal{J})} + \beta_i p_{ij}^{(\mathcal{J})} q_{ij}^{(I)}). \tag{4}$$

Here α_i and β_i are positive or negative numbers and denote, respectively, the pay-off to sender and receiver for correct communication about object i.

(c) *Evolutionary dynamics with errors during language acquisition*

The model can be used to study how signals can become associated with arbitrary meaning (Hurford 1989). Consider a population of size N. Each individual is characterized by an A matrix. The fitness is evaluated according to equation (2). Every individual talks to every other individual with equal probability. For the next generation, individuals produce children proportional to their pay-off. This is the standard assumption of evolutionary game theory; the pay-off of the game is related to fitness (Maynard Smith 1982). In the context of language evolution, it means that successful communication increases the survival probability or performance during life history and hence enhances the expected number of offspring. Thus, language is of adaptive value and contributes to biological fitness.

Children inherit from their parents a language acquisition device: this is a strategy for acquiring language. Children are born with an A matrix containing only zero entries. They sample their parents' language and thereby form their own A matrix. The language acquisition device will determine how accurately the child will learn the association matrix of its parent. The accuracy of this process will in turn determine whether or not a population can evolve a coherent language.

Let us for simplicity consider binary A matrices. If $a_{ij} = 1$ there is an association between object i and signal j; if $a_{ij} = 0$ there is none. Let w_0 denote the probability for the child's A matrix to have an entry of unity at a place where the parent's A matrix has a zero entry. Let w_1 denote the probability for the child's A matrix to have a zero entry at a place where the parent's A matrix has an entry of unity. Hence w_0 and w_1 represent the error rates of the language acquisition device. The parameter w_1 denotes the probability of losing an association, while w_0 denotes the probability of forming a new, possibly erroneous, association.

Figure 1 shows how the ability to evolve a coherent communication system depends on the error rates w_0 and w_1. The population size is $N = 100$. There are $n = 10$ objects and $m = 10$ signals. Hence each individual is characterized by a 10×10 binary A matrix. The figure shows the average pay-off of the population as a function of the error rate w_1, for two choices of w_0. The average pay-off of the population is given by $\bar{f} = (1/N) \sum_I f_I$, where $f_I = [1/(N-1)] \sum_{\mathcal{J} \neq I} F(A_I, A_{\mathcal{J}})$ is the average pay-off of individual I. The function $F(A_I, A_{\mathcal{J}})$ is given by equation (2). If all individuals speak the same 'perfect' language with n signals referring uniquely to n objects, then the average fitness is n. For $w_0 = 0.0001$, the average pay-off declines with increasing error-rate, w_1. To maintain a perfect language of a given size n, the error-rate w_1 has to be below a critical value. Similarly, w_0 has to be less than a critical value to get any evolutionary adaptation. For $w_0 = 0.005$, the average pay-off never increases above four and even declines for very low values of w_1; this is a consequence of individuals accumulating too many entries of unity in the association matrix.

For large population sizes, N, the evolutionary dynamics can be modelled by a system of ordinary differential equations. Denote by x_I the frequency of individuals with matrix A_I. There are $\nu = 2^{nm}$ binary A matrices of size $n \times m$. The evolutionary dynamics can be formulated as

$$\dot{x}_I = \sum_{\mathcal{J}=1}^{\nu} f_{\mathcal{J}} x_{\mathcal{J}} W_{I\mathcal{J}} - \phi x_I, \qquad I = 1, \ldots, \nu. \tag{5}$$

The fitness of individuals $\mathcal{J}$ is given by

$$f_{\mathcal{J}} = \sum_{I=1}^{\nu} F(L_{\mathcal{J}}, L_I) x_I. \tag{6}$$

This assumes that individual $\mathcal{J}$ talks to I with probability x_I. The quantity $f_{\mathcal{J}}$ denotes the expected pay-off of all interactions of individual $\mathcal{J}$. The average fitness of the population is given by

$$\phi = \sum_{I=1}^{\nu} f_I x_I. \tag{7}$$

For equation (5), the total population size is constant. We set $\sum_I x_I = 1$. The parameter $W_{I\mathcal{J}}$ denotes the probability that someone learning from an individual with $A_{\mathcal{J}}$ will end up with A_I. Thus, W_{II} denotes the probability of

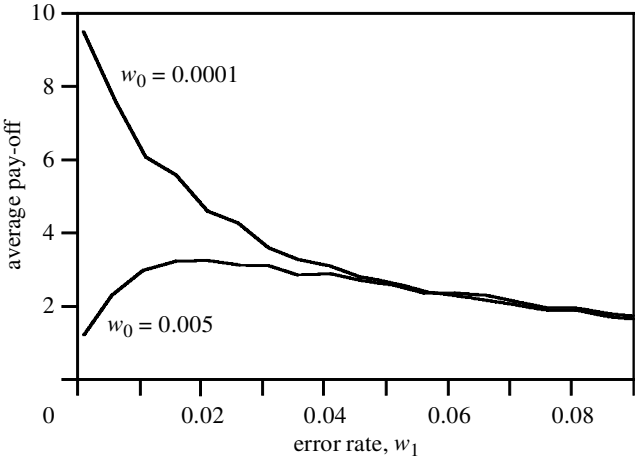

Figure 1. The maximum size of a communication system that can be stably maintained by a population depends on the error rate during language acquisition. Each individual is characterized by a binary association matrix, A, linking arbitrary signals to objects (or concepts). There are $n = 10$ objects and $m = 10$ signals. A is a binary 10×10 matrix. The population size is $N = 100$. In every generation, each individual talks to every other individual and the pay-offs are summed up. For the next generation, individuals produce offspring who learn their language. The error rate, w_0, is the probability that a child will have an entry of unity in the A matrix where the parent had a zero entry. Similarly, w_1 is the probability for the child to lose an entry of unity. The figure shows the average pay-off of the population averaged over 5000 generations. The average pay-off is indicative of the number of signals that can be maintained by the population. If all individuals speak a 'perfect' language with n signals referring uniquely to n objects, the average pay-off of the population is n. For $w_0 = 0.001$, the average pay-off is a declining function of w_1. For $w_0 = 0.005$, the average pay-off is a one-humped function; for very low w_1, individuals accumulate too many associations in their A matrix. For a population to maintain a signalling system of size $n \times n$, both error rates have to be below some critical threshold values.

correct learning, while $W_{I\bar{J}}$ with $I \neq \bar{J}$ denotes learning with mistakes.

Equation (5) is an interesting extension of the quasispecies equation (Eigen & Schuster 1979). Standard quasispecies theory has constant fitness values, whereas our equation has frequency-dependent fitness values. Thus equation (5) can be considered to be at the interface between quasispecies theory and evolutionary game dynamics.

Equation (5) can be used to calculate the maximum error rates that are compatible with language evolution (Komarova & Nowak 2001). Such relationships allow us to understand the basic requirements of a language acquisition device for the evolution and stability of simple communication systems.

3. ERRORS DURING COMMUNICATION

In this section, we analyse the consequences of errors during communication. We will show that such errors limit the maximum fitness of a language irrespective of the total number of objects that are being described by the language. If communication about different objects leads to different pay-off contributions, then the maximum fitness of a language can be achieved by concentrating only on a small number of the most valuable objects, all other objects being ignored (Nowak & Krakauer 1999; Nowak *et al.* 1999*a*).

Denote by u_{ij} the probability of mistaking signal i for signal j. The corresponding error matrix, U, is a stochastic $m \times m$ matrix. Its rows sum to unity. The diagonal values, u_{ii}, define the probabilities of correct communication. Given this error matrix, the fitness of a language becomes

$$F = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} p_{ij} u_{jk} q_{ik}. \tag{8}$$

In the best possible case, the language is given by a permutation matrix (assuming $m = n$) and the fitness is simply given by the sum over the diagonal entries of the error matrix,

$$F = \sum_{i=1}^{n} u_{ii}. \tag{9}$$

The error matrix can be defined in terms of similarity between signals. Denote by s_{ij} the similarity between signals i and j. 'Similarity' should be a number between zero and unity, with unity denoting 'identity'. Thus we have $s_{ii} = 1$. The probability of mistaking signal i for j is now given by $u_{ij} = s_{ij} / \sum_{k=1}^{n} s_{ik}$; hence, the probability of mistaking signal i for j is defined by how similar signal i is to signal j compared to how similar signal i is to all other signals in the language. The probability of correct communication is given by $u_{ii} = 1 / \sum_{k=1}^{n} s_{ik}$. Thus, the fitness function in terms of similarity becomes

$$F = \sum_{i=1}^{n} \left(1 \Big/ \sum_{j=1}^{n} s_{ij} \right). \tag{10}$$

Let us now imagine that signals (or, more specifically, 'sounds' if we consider a spoken language) can be embedded in some metric space, and that d_{ij} denotes the distance between sounds i and j. The similarity is a monotonically decreasing function of their distance, $s_{ij} = f(d_{ij})$. If this function is strictly positive on some interval $[0, \varepsilon)$, then it is possible to show that the maximum fitness of a language is bounded by a constant which is independent of n (Dress *et al.* 2001). In other words, adding the possibility of describing more and more objects (or concepts) to the repertoire of a language cannot increase the maximum amount of information transfer beyond a certain limit.

If, in addition, we assume that objects have different values, then we find that the maximum fitness is usually achieved by limiting the repertoire of the language to a small number of objects. Increasing the repertoire of the language can reduce fitness. Hence natural selection will prefer communication systems with limited repertoires.

In § 3, the repertoire size of the language was limited by errors during language learning. Here an error limit arises as a consequence of errors during communication: if signals can be mistaken for each other, it can be better to have fewer signals that can be clearly identified.

In our current understanding, all animal communication systems seem to be based on fairly limited repertoires. Bees use a three-dimensional analogue system.

Birds have alarm calls for a small number of predators. Vervet monkeys have a handful of signals, their best studied signals being 'leopard', 'eagle' and 'snake'. In contrast, human language has a nearly unlimited repertoire. How did we overcome the error limit?

4. WORD FORMATION

The error limit can be overcome by combining sounds into words. We will provide a very simple and intuitive argument: more detailed descriptions are possible.

Words are strings of sounds. Linguists call these sounds 'phonemes'. Suppose there are n phonemes. Let us at first only consider words of length two phonemes. There are n^2 such words. We assume that the similarity between two words is the product of the similarities between the phonemes in corresponding positions. Thus if word W_{ij} consists of phonemes i and j, then the similarity between the words W_{ij} and W_{kl} is

$$S(W_{ij}, W_{kl}) = s_{ik}s_{jl}. \tag{11}$$

The fitness of a language that contains n^2 words to describe the same number of objects is

$$F = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(1\Big/\sum_{k=1}^{n}\sum_{l=1}^{n}s_{ik}s_{jl}\right). \tag{12}$$

This can be rewritten as

$$F = \left[\sum_{i=1}^{n}\left(1\Big/\sum_{j=1}^{n}s_{ij}\right)\right]^2. \tag{13}$$

Similarly for word-length K, we obtain

$$F = \left[\sum_{i=1}^{n}\left(1\Big/\sum_{j=1}^{n}s_{ij}\right)\right]^K. \tag{14}$$

Hence, if $F_{\max}(K)$ denotes the maximum fitness that can be achieved for a given word-length K, we have

$$F_{\max}(K) = F_{\max}(1)^K. \tag{15}$$

This equation describes the maximum fitness for a language that contains words of constant length K. Hence the maximum fitness increases exponentially with word-length. Note that $F_{\max}(1) \geqslant 1$.

There is an interesting connection to a central result of information theory. Shannon's noisy coding theorem states that for a discrete, memoryless channel with a certain capacity, there exists a sequence of codes such that with increasing length of the code word the maximum error probability declines exponentially to zero. Shannon's framework is similar to ours: our P matrix describes 'encoding', our Q matrix describes decoding, the error matrix, U, is equivalent to Shannon's channel. It is possible to show that our fitness function is proportional to unity minus Shannon's maximum error probability. Hence equation (15) is directly related to Shannon's noisy coding theorem (Plotkin & Nowak 2000).

If objects have different values, then word formation leads to a much larger number of objects being described at maximum fitness value.

5. POPULATION DYNAMICS OF WORDS

Let us now study the population dynamics of words. Suppose a language contains n words. Each individual is born not knowing any of the words, but can acquire words by learning from other individuals. Individuals are characterized by the subset of words they know. There are 2^n possibilities for the internal lexicon of an individual. Internal lexica are defined by bit strings: unity means that the corresponding word is known, zero means it is not. Let us enumerate them by $I = 0, \ldots, \nu$ where $\nu = 2^n - 1$. The number I is the integer representation of the corresponding bit string. Denote by x_I the abundance of individuals with internal lexicon I. The population dynamics can be formulated as

$$\dot{x}_I = \delta_I - x_I + b\sum_{J=0}^{\nu}\sum_{K=0}^{\nu}(x_J x_K Q_{JKI} - x_I x_J Q_{IJK}), \qquad I = 0, \ldots, \nu. \tag{16}$$

We have $\delta_0 = 1$ and $\delta_I = 0$ for $I > 0$; thus all individuals are born not knowing any of the words. Individuals die at a constant rate, which we set to unity, thereby defining a time-scale. The quantities Q_{IJK} denote the probabilities that individual I learning from J will become K. Equations (16) can be studied analytically if we assume that in any one interaction between two individuals only a single new word can be acquired and if words are memorized independently of each other. Thus the acquisition of the internal lexicon of each individual proceeds in single steps. The parameter b is the total number of word learning events per individual per lifetime. In this case, we obtain for the population dynamics of x_i, which is the relative abundance of individuals who know word W_i,

$$\dot{x}_i = -x_i + R_i x_i(1 - x_i). \tag{17}$$

Here $R_i = bq\phi_i$ is the basic reproductive ratio of word W_i. It is the average number of individuals who acquire word W_i from one individual who knows it. The parameter q is the probability to memorize a single word, and ϕ_i is the frequency of occurrence of word W_i in the (spoken) language. If $R_i > 1$, then x_i will converge to the equilibrium $x_i^* = 1 - 1/R_i$. We can now derive an estimate for the maximum size of a lexicon. From $R_i > 1$ we obtain $\phi_i > 1/(bq)$. Suppose W_i is the least frequent word. We certainly have $\phi_i \leqslant 1/n$, and hence the maximum number of words is $n_{\max} = bq$. Note that this number is always less than the total number of words, b, that are presented to a learning individual. Hence, the combined lexicon of the population cannot exceed the total number of word-learning events for each individual.

A curious observation of English and other languages is that the word frequency distributions follow Zipf's law (Zipf 1935; Estoup 1916; Mandelbrot 1958): the frequency of the ith most frequent word is given by a constant divided by i. Therefore we have

$$\phi_i = C/i. \tag{18}$$

The constant is given by $C = 1/\sum_i (1/i)$. Nobody knows the significance of Zipf's law for language. Miller & Chomsky (1963), however, point out that a random source emitting symbols and spaces will also generate

word frequency distributions that follow Zipf's law. This seems to suggest that Zipf's law is a kind of null hypotheses of word distributions.

We can use Zipf's law to derive an improved equation for the maximum lexicon size. Assuming that word frequency distributions follow Zipf's law, we find that the maximum number of words is approximately given by the equation

$$n_{\max}(\gamma + \ln n_{\max}) = bq. \tag{19}$$

We have used Euler's gamma: $\gamma = 0.5772\ldots$. Suppose we want to maintain a language with $n = 100$ words. If the probability of memorizing a word after one encounter is given by $q = 0.1$, we need $b \approx 5000$ word-learning events. For $n = 10^4$ and $q = 0.1$ we need $b \approx 10^6$.

6. THE EVOLUTION OF SYNTAX

Animal communication is believed to be non-syntactic: signals refer to whole events. Human language is syntactic: signals consist of components that have their own meaning. Syntax allows us to formulate a nearly unlimited number of sentences. Let us now use the mathematical framework of § 5 to study the transition from non-syntactic to syntactic communication.

Imagine a group of individuals who communicate about events in the world. Events are combinations of objects, places, times and actions. (We use 'object' and 'action' in a very general way as everything that can be referred to by nouns and verbs of current human languages.) For notational simplicity, suppose that each event consists of one object and one action. Thus event E_{ij} consists of object i and action j. Denote by r_{ij} the rate of occurrence of event E_{ij}. Denote by ϕ_{ij} the frequency of occurrence of event E_{ij}. We have $\phi_{ij} = r_{ij}/\sum_{ij} r_{ij}$. Non-syntactic communication uses words for events, while syntactic communication uses words for objects and actions.

Let us first consider the population dynamics of non-syntactic communication. The word, W_{ij}, refers to event E_{ij}. The basic reproductive ratio of W_{ij} is given by $R(W_{ij}) = bq\phi_{ij}$. If $R(W_{ij}) > 1$ then the word W_{ij} will persist in the population, and at equilibrium the relative abundance of individuals who know this word is given by

$$x^*(W_{ij}) = 1 - 1/R(W_{ij}). \tag{20}$$

As in § 5, the maximum number of words that can be maintained in the population is limited by bq.

For natural selection to operate on language design, language must confer fitness. Assume that correct communication about events confers some fitness advantage to the interacting individuals. In terms of our model, the fitness contribution of a language can be formulated as the probability that two individuals know the correct word for a given event summed over all events and weighted with the rate of occurrence of these events. Hence, at equilibrium, the fitness of individuals using non-syntactic communication is given by

$$F_{\mathrm{ns}} = \sum_{i,j} x^*(W_{ij})^2 r_{ij}. \tag{21}$$

Let us now turn to syntactic communication. Noun N_i refers to object i, and verb V_j refers to action j; hence the event E_{ij} is described by the sentence N_iV_j. For the basic reproductive ratios we obtain $R(N_i) = (b/2)q_{\mathrm{s}}\phi(N_i)$ and $R(V_j) = (b/2)q_{\mathrm{s}}\phi(V_j)$. The frequency of occurrence of noun N_i is $\phi(N_i) = \sum_j \phi_{ij}$, and of verb V_j it is $\phi(V_j) = \sum_i \phi_{ij}$. The factor $1/2$ appears because either the noun or the verb is learned in any one of the b learning events. The probability to memorize a noun or a verb is given by q_{s}. We expect q_{s} to be (slightly) smaller than q, which simply means that it is a more difficult task to learn a syntactic signal than a non-syntactic signal. For both signals, the (arbitrary) meaning has to be memorized; for a syntactic signal one also has to memorize how it relates to other signals (whether it is a noun or a verb, for example).

For noun N_i to be maintained in the lexicon of the population, we require $R(N_i) > 1$, which implies $\phi(N_i) > 2/(bq_{\mathrm{s}})$. Similarly for verb V_j we find $\phi(V_j) > 2/(bq_{\mathrm{s}})$. This means that the total number of nouns plus verbs is limited by bq_{s}, which is always less than b. The maximum number of grammatical sentences, however, which consist of one noun and one verb, is given by $(bq_{\mathrm{s}})^2/4$. Hence syntax makes it possible to maintain more sentences than the total number of sentences, b, that are said to a learning individual by all of his or her teachers together. Therefore all words have to be learned, but syntactic signals enable the formulation of new sentences that have not been learned beforehand.

For calculating the fitness of syntactic communication, note that two randomly chosen individuals can communicate about event E_{ij} if they both know noun N_i and verb V_j. Denote by $x(N_iV_j)$ the relative abundance of individuals who know N_i and V_j. From equations (1) we obtain the dynamics

$$\begin{aligned}\dot{x}(N_iV_j) = &-x(N_iV_j) + R(N_i)x(N_i)[x(V_j) - x(N_iV_j)] \\ &+ R(V_j)x(V_j)[x(N_i) - x(N_iV_j)].\end{aligned} \tag{22}$$

If $R(N_i) > 1$ and $R(V_j) > 1$, the abundances converge to the equilibrium

$$x^*(N_iV_j) = \frac{x^*(N_i)x^*(V_j)}{1 - 1/[R(N_i) + R(V_j)]}. \tag{23}$$

At equilibrium, the fitness of syntactic communication is given by

$$F_{\mathrm{s}} = \sum_{i,j} x^*(N_iV_j)^2 r_{ij}. \tag{24}$$

When does syntactic communication lead to a higher fitness than non-syntactic communication? Suppose there are n objects and m actions. Suppose a fraction, p, of these mn events occur, while the other events do not occur. In this case $R(W_{ij}) = bq/(pmn)$ for those events that occur, and $R(N_i) = bq_{\mathrm{s}}/(2n)$ and $R(V_j) = bq_{\mathrm{s}}/(2m)$. We make the (somewhat rough) assumption that all nouns and all verbs, respectively, occur on average at the same frequency. If all involved basic reproductive ratios are well above unity, we find that $F_{\mathrm{s}} > F_{\mathrm{ns}}$ leads to

$$\frac{m^2n + mn^2}{m^2 + mn + n^2} > \frac{2q}{pq_{\mathrm{s}}}. \tag{25}$$

If this inequality holds then syntactic communication will be favoured by natural selection. Otherwise non-syntactic communication will win. For $m = n$, condition (25) reduces to

$$n > 3q/(p q_s). \quad (26)$$

Therefore the size, n, of the communication system has to exceed a threshold value before natural selection can see the advantage of syntactic communication. This threshold value depends crucially on the parameter p, which describes the syntactic structure of the relevant events. If p is small then most events are unique object–action pairings and syntax will not evolve. The number np is the average number of relevant events that contain a particular noun or verb. This number has to exceed three before syntax could evolve.

'Relevant event' means there is a fitness contribution for communicating about this event. As the number of such 'relevant communication topics' increased, natural selection could begin to favour syntactic communication and thereby lead to a language design where messages could be formulated that were not learned beforehand. Syntactic messages can encode new ideas or refer to extremely rare but important events. Our theory, however, does not suggest that syntactic communication is always at an advantage. It is likely that many animal species have a syntactic understanding of the world, but natural selection did not produce a syntactic communication system for these species, because the number of relevant signals was below the threshold illustrated by equation (25). Presumably the increase in the number of relevant communication topics was caused by changes in the social structure and interaction of those human ancestors who evolved syntactic communication.

7. CONCLUSIONS

We have outlined some basic mathematical models that enable us to study a number of the most fundamental steps that are necessary for the evolution of human language by natural selection. We have studied the basic requirements for a language acquisition device that are necessary for the evolution of a coherent communication system described by an association matrix that links objects of the world (or concepts) to arbitrary signals. Errors during language learning lead to evolutionary change and adaptation of improved information transfer. Misunderstandings during communication lead to an error limit: the maximum fitness is achieved by a system with a small number of signals referring to a small number of relevant objects. This error limit can be overcome by word formation, which represents a transition from an analogue to a digital communication system.

Words are maintained in the lexicon of a language if their basic reproductive ratio exceeds unity: a person who knows a word must transmit knowledge of this word to more than one new person on average. Since there is a limit on how much people can say to each other and how much they can memorize, this implies a maximum size for the lexicon of a language (in the absence of written records).

Words alone are not enough. The nearly unlimited expressibility of human language comes from the fact that we use syntax to combine words into sentences. In the most basic form, syntax refers to a communication system where messages consist of components that have their own meaning. Non-syntactic communication, in contrast, has signals that refer to whole situations. Natural selection can only see the advantages of syntactic communication if the size of the system is above a critical value. Below this value non-syntactic communication is more efficient.

Throughout the paper we have assumed that language is about information transfer. Efficient and unambiguous communication together with easy learnability of the language is rewarded in terms of pay-off and fitness. While we think that these are absolutely fundamental and necessary assumptions for much of language evolution, we also note the seemingly unnecessary complexity of current languages. Certainly, systems designed by evolution are often not optimized from an engineering perspective. Moreover, it seems likely that at times evolutionary forces were at work to make things more ambiguous and harder to learn, such that only a few selected listeners could understand the message. If a good language performance enhances the reputation within the group, we can also imagine an arms race towards increased and unnecessary complexity. Such a process can drive the numbers of words and rules beyond what would be best for efficient information exchange. This should be the subject of papers to come.

This paper is dedicated to the 80th birthday of John Maynard Smith. The first theoretical biology paper I ever read was a *Scientific American* article by John Maynard Smith on evolutionary game theory. The first time I came across questions of language evolution was during a lecture delivered by John Maynard Smith at the University of Oxford.

Support from the Leon Levy and Shelby White Initiatives Fund, the Florence Gould Foundation, the J. Seward Johnson, Sr Charitable Trusts, the Ambrose Monell Foundation and the Alfred P. Sloan Foundation is gratefully acknowledged.

REFERENCES

Aboitiz, F. & Garcia, R. 1997 The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. *Brain Res. Rev.* **25**, 381–396.

Aoki, K. & Feldman, M. W. 1989 Pleiotropy and preadaptation in the evolution of human language capacity. *Theor. Popul. Biol.* **35**, 181–194.

Bates, E. 1992 Language development. *Curr. Opin. Neurobiol.* **2**, 180–185.

Bickerton, D. 1990 *Language and species*. University of Chicago Press.

Brandon, R. N. & Hornstein, N. 1986 From icons to symbols: some speculations on the origin of language. *Biol. Phil.* **1**, 169–189.

Burling, R. 1993 Primate calls, human language, and nonverbal communication. *Curr. Anthropol.* **34**, 25–53.

Cavalli-Sforza, L. L. 1997 Genes, peoples, and languages. *Proc. Natl Acad. Sci. USA* **94**, 7719–7724.

Cavalli-Sforza, L. L. & Cavalli-Sforza, F. 1995 *The great human diasporas: the history of diversity and evolution*. Translated by Sarah Thomas. Reading, MA: Addison-Wesley.

Cavalli-Sforza, L. L. & Feldman, M. W. 1981 *Cultural transmission and evolution: a quantitative approach*. Princeton University Press.

Cheney, D. & Seyfarth, R. 1990 *How monkeys see the world*. University of Chicago Press.
Chomsky, N. 1965 *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. 1972 *Language and mind*. New York: Harcourt Brace Jovanovich.
Chomsky, N. 1988 *Language and problems of knowledge: the Managua lectures*. Cambridge, MA: MIT Press.
Cole, R. A. & Jakimik, J. 1980 A model of speech perception. In *Perception and production of fluent speech* (ed. R. A. Cole), pp. 103–124. Hillsdale, NJ: Erlbaum.
Corballis, M. 1991 *The lopsided ape*. New York: Oxford University Press.
Darwin, C. R. 1871 *The descent of man, and selection in relation to sex*. London: John Murray.
Deacon, T. 1997 *The symbolic species*. London: Penguin Books.
Dress, A., Mueller, S. & Nowak, M. A. 2001 The information storage capacity of compact metric. (In preparation.)
Eigen, M. & Schuster, P. 1979 *The hypercycle*. Berlin: Springer.
Estoup, J. B. 1916 *Gammes stenographique*. Paris: Gauthier-Villars.
Gopnik, M. & Crago, M. 1991 Familial aggregation of a developmental language disorder. *Cognition* **21**, 1–50.
Grassly, N. C., Von Haeseler, A. & Krakauer, D. C. 2000 Error, population structure and the origin of diverse sign systems. *J. Theor. Biol.* **206**, 369–378.
Greenberg, J. H. 1971 *Language, culture, and communication*. Stanford University Press.
Hauser, M. D. 1996 *The evolution of communication*. Cambridge, MA: Harvard University Press.
Hurford, J. R. 1989 Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187–222.
Hurford, J. R. 1991 The evolution of the critical period for language acquisition. *Cognition* **40**, 159–201.
Hurford, J. R., Studdert-Kennedy, M. & Knight, C. (eds) 1998 *Approaches to the evolution of language*. Cambridge University Press.
Hutsler, J. J. & Gazzaniga, M. S. 1997 The organization of human language cortex: special adaptation or common cortical design? *Neuroscientist* **3**, 61–72.
Jackendoff, R. S. 1997 *The architecture of the language faculty*. Cambridge, MA: MIT Press.
Komarova, N. L. & Nowak, M. A. 2001 Mathematical models for language evolution. (Submitted.)
Krakauer, D. C. 2000 Kin imitation for a private sign system. (Submitted.)
Liberman, A. M., Cooper, F. S., Shankweiler, D. P. & Studdert-Kennedy, M. 1967 Perception of the speech code. *Psychol. Rev.* **74**, 431–461.
Lieberman, P. 1984 *The biology and evolution of language*. Cambridge, MA: Harvard University Press.
Mandelbrot, B. 1958 Les lois statistique macroscopiques du comportment. *Psychol. Francaise* **3**, 237–249.
Maynard Smith, J. 1982 *Evolution and the theory of games*. Cambridge University Press.
Maynard Smith, J. & Szathmáry, E. 1995 *The major transitions in evolution*. Oxford and New York: W. H. Freeman Spektrum.
Maynard Smith, J. & Szathmáry, E. 1999 *The origins of life*. Oxford University Press.
Miller, G. A. 1967 *The psychology of communication*. London: Penguin Books.
Miller, G. A. 1981 *Language and speech*. San Francisco, CA: W. H. Freeman & Co.
Miller, G. A. 1991 *The science of words*. New York: Scientific American Library.
Miller, G. A. & Chomsky, N. 1963 Finitary models of language users. In *Handbook of mathematical psychology*, vol. 2 (ed. R. D. Luce, R. Bush & E. Galanter), pp. 419–491. New York: Wiley.
Nagy, W. E. & Anderson, R. C. 1984 How many words are there in printed school English? *Reading Res. Q.* **19**, 304–330.
Nagy, W. E., Diadikoy, L. & Anderson, R. 1993 The acquisition of morphology: learning the contribution of suffixes to the meanings of derivatives. *J. Reading Behav.* **25**, 155–170.
Newmeyer, F. 1991 Functional explanation in linguistics and the origin of language. *Lang. Commun.* **11**, 3–96.
Newport, E. 1990 Maturational constraints on language learning. *Cogn. Sci.* **14**, 11–28.
Niyogi, P. 1998 *The informational complexity of learning*. Boston, MA: Kluwer Academic.
Niyogi, P. & Berwick, R. C. 1996 A language learning model for finite parameter spaces. *Cognition* **61**, 161–193.
Nobre, A., Allison, T. & McCarthy, G. 1994 Word recognition in the human inferior temporal lobe. *Nature* **372**, 260–263.
Nowak, M. A. 2000 The basic reproductive ratio of a word, the maximum size of a lexicon. *J. Theor. Biol.* **204**, 179–189.
Nowak, M. A. & Krakauer, D. C. 1999 The evolution of language. *Proc. Natl Acad. Sci. USA* **96**, 8028–8033.
Nowak, M. A., Krakauer, D. C. & Dress, A. 1999*a* An error limit for the evolution of language. *Proc. R. Soc. Lond.* B **266**, 2131–2136.
Nowak, M. A., Plotkin, J. B. & Krakauer, D. C. 1999*b* The evolutionary language game. *J. Theor. Biol.* **200**, 147–162.
Nowak, M. A., Plotkin, J. B. & Jansen, V. A. A. 2000 The evolution of syntactic communication. *Nature* **404**, 495–498.
Pinker, S. 1994 *The language instinct*. New York: William Morrow.
Pinker, S. 1999 *Words and rules*. New York: Basic Books.
Pinker S., Bloom, P. & commentators 1990 Natural language and natural selection. *Behav. Brain Sci.* **13**, 707–784.
Plotkin, J. B. & Nowak, M. A. 2000 Language evolution and information theory. *J. Theor. Biol.* **205**, 147–159.
Seyfarth, R., Cheney, D. & Marler, P. 1980 Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science* **210**, 801–803.
Smith, W. J. 1977 *The behavior of communicating*. Cambridge, MA: Harvard University Press.
Steels, L. 1997 The synthetic modelling of language origins. *Evol. Commun. J.* **1**, 1–17.
Trapa, P. E. & Nowak, M. A. 2000 Nash equilibria for an evolutionary language game. *J. Math. Biol.* **41**, 172–188.
Von Frisch, K. 1967 *The dance language and orientation of bees*. Cambridge, MA: Harvard University Press.
Yasuda, N., Cavalli-Sforza, L. L., Skolnick, M. & Moroni, A. 1974 The evolution of surnames: an analysis of their distribution and extinction. *Theor. Popul. Biol.* **5**, 123–142.
Zipf, G. K. 1935 *The psychobiology of language*. Boston, MA: Houghton-Mifflin.