Entropy **2001**, *3*, 227–246

entropy
ISSN 1099-4300
www.mdpi.org/entropy/

Major Transitions in Language Evolution

Joshua B. Plotkin* and **Martin A. Nowak**

Institute For Advanced Study, Olden Lane, Princeton, NJ 08540, U.S.A.
E-mail: plotkin@ias.edu, nowak@ias.edu

Submitted: 26 June 2001 / Accepted: 1 October 2001 / Published 10 October 2001

Abstract: Language is the most important evolutionary invention of the last few million years. How human language evolved from animal communication is a challenging question for evolutionary biology. In this paper we use mathematical models to analyze the major transitions in language evolution. We begin by discussing the evolution of coordinated associations between signals and objects in a population. We then analyze word-formation and its relationship to Shannon's noisy coding theorem. Finally, we model the population dynamics of words and the adaptive emergence of syntax.

Keywords: Language evolution; evolutionary game theory; Shannon's noisy coding theorem; phoneme.

---

*The authors gratefully acknowledge support from the Alfred P. Sloan Foundation, The Ambrose Monell Foundation, The Florence Gould Foundation, and the J. Seward Johnson Trust. J.B.P also acknowledges support from the National Science Foundation and the Burroughs Wellcome Fund

1 Introduction

Everyone who reads this paper knows on the order of 50,000 words of his primary language. These words are stored in the 'mental lexicon' together with one or several meanings, some information 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 waking minutes. A six-year-old knows about 13,000 words ([41], [38], [51]-[53]).

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 novel combination. The brain contains a program that can build an unlimited number of sentences out of a finite list of words. This program is called 'mental grammar'([27]). Children develop this grammar rapidly and without formal instruction. Experimental observations show that 3 year old children apply grammatical rules correctly 90% of the time.

The most complicated mechanical motion that the human body can perform is the simple 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 millimeters and synchronized within a few 100th of a second ([38]).

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 sped-up speech. These numbers are surprising given the physical limitations of our auditory system: if a clicking sounds is repeated at a rate of about 20 per second, 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 ([36], [30], [14]).

The preceding paragraphs demonstrate that human language is an enormously complex trait. Our language performance relies on precisely coordinated interactions of various parts of anatomy, and we are amazingly good at it. We can all speak without thinking. In contrast, we often cannot perform basic mathematical operations without concentration. Why is doing math 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.

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 ([34], [35]), being of comparable importance as the origin of genetic replication, the origin of eukaryotes, or the emergence of multi-cellular organisms.

Attempts to shed light on the evolution of language have come from many areas including studies of primate social behavior ([55], [6], [10]) or animal communication ([14], [22], [56]), the diversity of existing human languages ([21], [7]), the development of language in children ([43], [3], [24]), theoretical studies of cultural evolution ([8], [9], [60], [2]) and learning theory ([44], [45]). Our objective here and in several related papers ([47]-[50], [54])) is to bring discussions of language evolution within the precise mathematical framework of modern evolutionary biology. For mathematical models of language evolution see also [23]-[24], [57].

In Section 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 errors during communication and show how such errors limit the repertoire a simple communication system. In Section 3, we show that word formation can overcome this error limit – a phenomenon explained by Shannon's noisy coding theorem. In Section 4, we design a framework for modelling the population dynamics of words. We define the basic reproductive ratio of words and calculate the maximum size of a lexicon. We discuss how natural selection can guide the emergence of syntactic communication. Section 5 is a conclusion.

2 Evolving Arbitrary Signals

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 speaking matrix, a listening matrix, a payoff function, and finally the evolutionary dynamics.

Consider a population of individuals that can communicate via signals. Signals may include gestures, facial expressions, or spoken sounds. We are interested how an arbitrary association between signals and 'objects' can evolve.

In the most simple model, each individual is described by an active matrix, P, and a passive matrix, Q ([25]). The entry P_{ij} denotes that the probability that the individual, as a speaker, will refer to object i by using signal j. The entry Q_{ji} denotes the probability that the individual, as a listener, will interpret signal j as referring to object i. Both P and Q are stochastic matrices; their entries lie in $[0, 1]$, and their rows each sum to one. The 'language' of an individual, $L = (P, Q)$, is defined by these two matrices.

When one individual using $L = (P, Q)$ communicates with another individual using $L' =$

(P', Q'), we define the payoff as the number of objects communicable between the individuals, weighted by their probability of correct communication. Thus the payoff for L versus L' is given by

$$F(L, L') = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij}Q'_{ji} + P'_{ij}Q_{ji} = \frac{1}{2}\left((PQ') + (P'Q)\right). \tag{2.1}$$

There are n objects and m signals. Loosely speaking, this payoff function reflects the total amount of information that L can convey to L', and *vica versa*. In this basic model, any possible miscommunication results from a discrepancy between the signal-object associations of the speaker and the listener. The maximum possible payoff to two individuals who share a common language is the smaller of n or m.

2.1 Evolutionary dynamics

This framework can be used to study how signals can become associated with arbitrary meaning (an approach pioneered by [23]). Consider a population of size N. Each individual is characterized by a language $L = (P, Q)$ matrix. The fitness is evaluated according to eq (2.1). Every individual talks to every other individual with equal probability. For the next generation, individuals produce children proportional to their payoff. This is the standard assumption of evolutionary game theory; the payoff of the game is related to fitness ([33]). 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* – a strategy how to acquire language. In the idealized case children learn the exact language spoken by their parents. In more realistic cases, children might sample their parents' languages, building an internal association matrix which, when normalized, gives a P and Q matrix. If children sample their parents' languages infinitely many times, they will adopt acquire an the same language as their parents. Under finite sampling, the child's language will only resemble its parents'. Alternatively, children might sample the languages of the most fit individual or individuals in the population. Children might even alternatively sample the language of a random individual in the population. In all cases, over time the population will evolve towards a coherent language $L = (P, Q)$. If children sample their parents' languages or the languages of well-spoken individuals, however, the resulting population will have a higher mean equilibrium fitness ([47], [49]).

It is interesting to ask which languages are Nash-equilibria in this evolutionary setting. It turns out that a language can satisfy $F(L, L) > F(L', L)$ for all L' if and only if $n = m$, P is permutation, and Q is transpose of P ([58]).

2.2 Errors during transmission

In this section, we analyze the consequences of transmission 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.

Denote by U_{ij} the probability of mistaking signal i for signal j. The corresponding signal-error matrix, U, is a stochastic $m \times m$ matrix. Its rows sum to 1. The diagonal values, U_{ii}, define the probabilities of correct communication. Given this error matrix, the fitness a language becomes

$$F = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} P_{ij} U_{jk} Q_{ki}. \tag{2.2}$$

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

$$F = \sum_{i=1}^{m} U_{ii}. \tag{2.3}$$

The signal-error matrix, U, can be constructed to reflect the similarities of the signals. In particular, we denote the similarity between signal i and signal j by S_{ij}. We stipulate that S_{ii}=1 and $S_{ij} \leq 1$. The probability of mistaking signal i for signal j quantifies how similar signal i is to j compared with all other signals: $U_{ij} = S_{ij} / \sum_{k=1}^{m} S_{ik}$. In these terms, the fitness of a common language, $n = m$ can be expressed as

$$F(L, L) = \sum_{i=1}^{m} \frac{1}{\sum_{k=1}^{m} S_{ik}}. \tag{2.4}$$

We imagine that the signals of the language are embedded in some compact metric space, X, and that d_{ij} denotes the distance between signals i and j. The similarity between two signals, then, is a decreasing function of the distance $S_{ij} = f(d_{ij})$.

For example, if we embed signals i into the unit interval $v_i \in [0, 1]$, and if similarity is an exponentially decreasing function of distance $S_{ij} = \exp(-\alpha|v_i - v_j|)$, then the maximal fitness satisfies

$$F(L, L) = \sum_{i=1}^{m} 1 / \sum_{k=1}^{m} \exp(-\alpha|v_i - v_k|). \tag{2.5}$$

As the number of signals increases, $m \to \infty$, then $F(L, L)$ is bounded by $1 + \alpha/2$, regardless of the choice of embedding ([48]). In other words, the maximal fitness is bounded, regardless of the number of signals and objects used.

This error-limit is an example of a general phenomenon. It can be shown that the maximal fitness is always bounded by some constant depending only on the X and f, but not on n ([17]).

In other words, even as the signal repertoire of a language increases, the fitness cannot exceed a fixed value.

If communication about different objects leads to different payoff 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 ([47], [49]). Increasing the repertoire of the language can reduce fitness. Hence natural selection will prefer communication systems with limited repertoires. 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 analog 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?

3 Word Formation

The error limit can be overcome by combining sounds into words. We will provide a very simple and intuitive argument for this which is closely related to the noisy coding theorem of Shannon.

Communicating with words

Words are strings of sounds. Linguists call these sounds 'phonemes'. We now develop a more general framework for word-based language. A language will now be described by four components: a *lexicon*, an *active matrix* P, a *passive matrix* Q, and a *phoneme error-matrix* V.

Our model of word-based language uses words which are l-phonemes long. The lexicon of the language, however, does not necessarily include all possible m^l words. Instead, the lexicon contains a subset of all possible words. We denote the phonemes of the language by the set $\Phi = \{\phi_1, \ldots, \phi_m\}$. We denote the lexicon by some subset $C \subset \Phi^l$. We refer to the words in C as the *lexicon* or *proper vocabulary* of the language. Let us denote the size of the lexicon by $n = |C|$ (*i.e.* n is the cardinality of the set C). Notice that n also denotes the number of objects expressible in the language.

The active matrix P defines the (probabilistic) association between objects and words for the speaker. P is now an n-by-m^l stochastic matrix whose ijth entry denotes the probability that a speaker will attempt to use word j to denote object i. By definition, nonzero entries in P may occur only at columns corresponding to words in the lexicon C. The passive matrix Q maps all possible perceived words (probabilistically) back into the n objects. We specify the passive matrix via a stochastic m^l-by-n matrix Q. The entry Q_{ji} represents the probability that a listener who perceives the jth word will interpret it as the ith object.

Finally, we must provide a description of transmission errors. As before, we use an m^l-by-m^l word error-matrix U. The entry U_{ij} denotes the probability that, when a speaker attempts to vocalize the ith word, the listener perceives the jth word. Notice that only the rows of U corresponding to lexicon words matter; we have assumed that a speaker will never *attempt* to vocalize an improper vocabulary word (although a speaker may, in fact, utter a word outside of the lexicon via a transmission error).

In strict analogy with previous models, the U-matrix is built upon the similarity between the phonemes of which the words are comprised. In particular, we start with a stochastic m-by-m *phoneme error-matrix* V. The entry V_{ij} denote the probability that, when a speaker attempts to vocalize the ith phoneme, the listener hears the jth phoneme. The similarity between words α and β is defined by the product of the similarity of their phonemes. In other words, we have the following expression for the word error-matrix:

$$U_{\alpha\beta} = \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}}. \tag{3.6}$$

where $\alpha^{(k)}$ denotes the kth phoneme of word α.

Thus, a language L is described completely by the three matrices $L = (P, Q, V)$. The matrix U is derived from V, and C is determined by those columns of P containing nonzero entries. Finally, we stipulate that all individuals in a population share the same V-matrix. In other words, all individuals use the same phonemic alphabet, and they share the same imperfections in their vocal and auditory organs.

In this setting, the proper payoff function (in strict analogy with previous models) is given by the sum of the number of objects which speaker L can convey to speaker L', weighted by the probability of communicating the objects correctly. In other words, letting w_i denote the ith object, we define

$$\begin{aligned} F(L, L') &= \sum_{i=1}^{n} \sum_{\alpha\in\Phi^l} \sum_{\beta\in\Phi^l} P_{w_i\alpha} U_{\alpha\beta} Q_{\beta w_i} &\quad (3.7) \\ &= \sum_{i=1}^{n} \sum_{\alpha\in\Phi^l} P_{w_i\alpha} \sum_{\beta\in\Phi^l} Q_{\beta w_i} \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}}. &\quad (3.8) \end{aligned}$$

We now ask what is the maximum possible fitness a language can obtain. Of course, the maximum is obtained when the speaker and listener share a common language given by binary active and passive matrices. But we do not yet know, given P and V, what is the optimal listening matrix Q.

Moreover, there remains another issue to be addressed: is it possible, by increasing the word length l, to increase a language's payoff without bound? In light of the error limit discussed in Section 2, this inquiry addresses a fundamental question regarding the adaptive benefits of word formation.

3.1 Shannon's noisy coding theorem

The adaptive benefit of word formation is clarified by appealing to the noisy coding theorem of Shannon. In this section we briefly review Shannon's fundamental result.

Shannon considers a discreet memoryless source I which emits characters from an *alphabet* $\Phi = \{\phi_1, \ldots, \phi_m\}$ according to some discreet probability distribution. The discreet source I is linked to a noisy channel used to transmit information. The channel is summarized by a channel matrix V. The entry V_{ij} gives the conditional probability $(\phi_j received|\phi_i sent)$.

Given a channel V and an input-source, we obtain a natural output stream J. The capacity $C(V) \in [0,1]$ measures the maximum rate at which information about an input stream may be inferred by inspecting the output stream:

$$C(V) = \sup_I [H(I) + H(J) - H(I,J)] \tag{3.9}$$

where H denotes the entropy of a source.

In order to increase fidelity, Shannon defines a set of n *codewords*, C, each codeword being a string of l characters from Φ. The *encoder* takes input messages from the source I, encodes the information into codewords, and sends the codeword on to the noisy channel, letter by letter. The *decoder* is a (deterministic) map from all possible outputs of the noisy channel, Φ^l, back to C. Shannon defines the *error probability* of this communication system as

$$e(C) = \frac{1}{n}\sum_{i=1}^{n}(errorincommunication$$

— $codeword \mathrm{w}_i$ is transmitted). (3.10)

Clearly one would like to construct codes with error probability as small as possible. This is precisely the problem which Shannon's fundamental theorem addresses.

Theorem 3.1 (Shannon, 1948) *If a discrete memoryless channel V has capacity $C > 0$ and R is any positive quantity with $R < C$, then there exists a sequence of codes $(C_i|1 \leq i < \infty)$ such that*

$$\begin{array}{ll} (a) & C_i \text{ has } 2^{\lfloor R\cdot i \rfloor} \text{ codewords of length } l = i \\ (b) & \text{the error probability satisfies } e(C_i) \leq Ae^{-Bi}, \end{array} \tag{3.11}$$

where the constants A and B depend only on the channel V and on R.

Shannon's theorem provides a sequence of communication systems with linearly increasing codeword length, exponentially increasing number of codewords (and thus describable objects), and exponentially decreasing error probability. (In essence, Shannon constructs each successive code C_i by choosing random codewords and decoding via the maximum likelihood method.) Shannon's coding theorem provides us with exponentially good codes. There is, however, an important converse to this theorem. The converse tells us that we could hope for nothing better:

Theorem 3.2 (Wolfowitz, 1961) *For a discreet memoryless channel of capacity C and for any $R > C$, there cannot exist a sequence of codes C_i such that C_i has $2^{R\cdot i}$ codewords of length i and error probability tending to zero. In fact, such a sequence of codes must have error probability which approaches one as $i \to \infty$.*

3.2 Coding theory and word-formation

Shannon's communication system is clearly related to our model of word-based language. A Shannon-encoder may be expressed as a binary n-by-m^l matrix P whose rows sum to one. The entry P_{ij} indicates whether or not the encoder uses word j to denote object (or message) i. Similarly, the decoder may be expressed as a binary m^l-by-n matrix Q. The entry Q_{ji} denotes whether or not the jth word is included in the subset words decoded as the ith codeword (or ith message).

In this setting, Shannon's codeword communication through a noisy channel is easily seen to be equivalent to our model for language. Shannon's alphabet Φ plays the role of the phonemes, the encoder plays the role of the active matrix, and the decoder the passive matrix. Shannon's "codewords" are simply strings of phonemes. Similarly, the noisy channel V plays the role of the phoneme error matrix. Shannon's communication system is always deterministic, however; it requires that the matrices P and Q are binary. Notice that, when P is binary, there is an unambiguous one-to-one correspondence between lexicon words and objects. In this situation, the "objects" expressible in our original language model may be identified with Shannon's codewords.

In light of the equivalence of these two systems, it is important to relate the information-theoretic definition of error probability – whose behavior is described by Shannon's theorem and its converse – with our definition of language fitness. Such a relation will allow us to use Theorem 3.1 to derive the maximal fitness of our word-based model.

Towards this end, consider the expression $\tilde{F}(C) = |C| \cdot (1 - e(C)) = n \cdot (1 - e(C))$. By Shannon's theorem, given a channel V with nonzero capacity, we can find a sequence of codes C_i with linearly increasing codeword length and with exponentially increasing $\tilde{F}(C)$. Thus Shannon's theorem (together with its converse) reveals the maximal properties of $\tilde{F}(C)$. It is not difficult to see, however, that $\tilde{F}(C)$ is equivalent to the fitness of language in our evolutionary model: $\tilde{F}(C) = F(L, L)$. The proof of this statement is little more than an exercise in unravelling definitions ([54]).

Therefore, if all the individuals in a population use the same language, and if that language has *binary* P and Q-matrices, then the fitness $F(L, L)$ agrees with the information-theoretic quantity $\tilde{F}(C)$. As a consequence, Shannon's coding theorem implies the following result.

Theorem 3.3 (Word Formation) *Given a phoneme error-matrix V (with nonzero capacity), there exists a sequence of languages L_i with linearly increasing word-length and exponentially increasing fitness.*

Thus word formation overcomes the error limit which constrains strictly phonemic communication; increasing word-length can increase fitness without bound. This result highlights the importance of word formation, which is more or less unique to the human species.

4 The emergence of syntax

We now study a later stage in the evolution of language when the population has agreed upon a common association between objects and words. But individuals vary in the extent and composition of their lexica. We now develop a model to study the population dynamics of the words themselves – the frequency which they are found among the lexica of different individuals.

4.1 Population dynamics of 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: 1 means that the corresponding word is known, 0 means it is not. Let us enumerate them by $I = 0, .., \nu$ 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

$$\dot{x}_I = \delta_I - x_I + b \sum_{J=0}^{\nu} \sum_{K=0}^{\nu} (x_J x_K Q_{JKI} - x_I x_J Q_{IJK}) \qquad I = 0, .., \nu \tag{4.12}$$

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 1, thereby defining a time scale. The quantities Q_{IJK} denote the probabilities that individual I learning from J will become K. Eq (4.12) 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 life-time. In this case, we obtain for the population dynamics of x_i, which is the relative abundance of individuals who know word W_i:

$$\dot{x}_i = -x_i + R_i x_i (1 - x_i). \tag{4.13}$$

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 $\phi_i \leq 1/n$, 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 ([61], [**?**], [36]): the frequency of the i-th most frequent word is given by a constant divided by i. Therefore we have

$$\phi_i = C/i. \tag{4.14}$$

The constant is given by $C = 1/\sum_i (1/i)$. Nobody knows the significance of Zipf's law for language. Miller & Chomsky ([39]), 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

$$n_{\max}(\gamma + \ln n_{\max}) = bq. \tag{4.15}$$

We have used Euler's gamma: $\gamma = 0.5772...$. 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 $b \approx 10^6$.

4.2 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 Section 4.1 in order to study the transition from non-syntactic to syntactic communication.

Imagine a group of individuals that 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

$$x^*(W_{ij}) = 1 - 1/R(W_{ij}). \tag{4.16}$$

As in Section 4.1, 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

$$F_{ns} = \sum_{i,j} x^*(W_{ij})^2 r_{ij}. \tag{4.17}$$

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 $R(N_i) = (b/2)q_s\phi(N_i)$ and $R(V_j) = (b/2)q_s\phi(V_j)$. 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 $\phi(N_i) > 2/(bq_s)$. Similarly for verb V_j we find $\phi(V_j) > 2/(bq_s)$. 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 $(bq_s)^2/4$. 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 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 eq (4.12) we obtain the dynamics

$$\dot{x}(N_iV_j) = -x(N_iV_j) + R(N_i)x(N_i)[x(V_j) - x(N_iV_j)] \tag{4.18}$$

$$+ R(V_j)x(V_j)[x(N_i) - x(N_iV_j)]. \tag{4.19}$$

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

$$x^*(N_iV_j) = \frac{x^*(N_i)x^*(V_j)}{1 - 1/[R(N_i) + R(V_j)]}. \tag{4.20}$$

At equilibrium, the fitness of syntactic communication is given by

$$F_s = \sum_{i,j} x^*(N_iV_j)^2 r_{ij}. \tag{4.21}$$

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 $R(N_i) = bq_s/(2n)$ and $R(V_j) = bq_s/(2m)$. 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 one, we find that $F_s > F_{ns}$ leads to

$$\frac{m^2n + mn^2}{m^2 + mn + n^2} > \frac{2q}{pq_s}. \tag{4.22}$$

If this inequality holds then syntactic communication will be favored by natural selection. Otherwise non-syntactic communication will win. For $m = n$ condition (4.22) reduces to

$$n > 3q/(pq_s). \tag{4.23}$$

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 must exceed thee before syntax can 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 favor 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 eq (4.23). 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.

5 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 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 one: 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 assumed that language was about information transfer. Efficient and unambiguous communication as well as easy learnability of the language is rewarded in terms of payoff 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 toward increased and unnecessary complexity. Such a process can drive the numbers of words and rules beyond what would be best for efficient information exchange. We hope this will be the subject of papers to come.

References

[1] 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.

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

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

[4] Bickerton, D. 1990 *Language and Species*. Chicago: University of Chicago Press.

[5] Brandon, R. N. & Hornstein, N. 1986 From icons to symbols: Some speculations on the origin of language. *Biology and Philosophy* **1**, 169–189.

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

[7] 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.

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

[9] Cavalli-Sforza, L. L. & Feldman, M. W. 1981 *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton: Princeton University Press.

[10] Cheney, D. & Seyfarth, R. 1990 *How Monkeys See the World*. Chicago: University of Chicago Press.

[11] Chomsky, N. 1965 *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

[12] Chomsky, N. 1972 *Language and Mind*. New York: Harcourt Brace Jovanovich.

[13] Chomsky, N. 1988 *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: MIT Press.

[14] Cole, R. A. & Jakimik, J. 1980 A model of speech perception. In R. A. Cole (ed) *Perception and Production of Fluent Speech*. Hillsdale, NJ: Erlbaum.

[15] Corballis, M. 1991 *The Lopsided Ape*. New York: Oxford University Press.

[16] Deacon, T. 1997 *The Symbolic Species*. London: Penguin Books.

[17] Dress, A., Mueller, S., & Nowak, M. 2001. The information storage capacity of compact metric spaces. preprint.

[18] Estoup, J. B. 1916 *Gammes Stenographique*. Paris: Gauthier-Villars.

[19] Gopnik, M. & Crago, M. 1991 Familial aggregation of a developmental language disorder. *Cognition* **21**, 1–50.

[20] Grassly, N. C., Krakauer, D. C., & Von Haessler, A. 2001 Population structure and the origin of referential sign systems. *J. theor. Biol*, in press.

[21] Greenberg, J. H. 1971 *Language, Culture, and Communication*. Stanford, CA: Stanford University Press.

[22] Hauser, M. D. 1996 *The Evolution of Communication*. Cambridge, MA: Harvard University Press.

[23] Hurford, J. R. 1989 Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187–222.

[24] Hurford, J. R. 1991 The evolution of the critical period for language acquisition. *Cognition* **40**, 159–201.

[25] Hurford, J. R., Studdert-Kennedy, M., & Knight, C. (eds) 1998 *Approaches to the Evolution of Language*. Cambridge: Cambridge University Press.

[26] Hutsler, J. J. & Gazzaniga, M. S. 1997 The organization of human language cortex: Special adaptation or common cortical design? *Neuroscientist* **3**, 61–72.

[27] Jackendoff, R. S. 1997 *The Architecture of the Language Faculty*. Cambridge, MA: MIT Press.

[28] Komarova, N. L. & Nowak, M. A. 2001 Natural selection of the critical period for grammar acquisition, *Proc R Soc Lond B*, submitted

[29] Krakauer, D. C. 2001 Kin imitation for a private sign system. *Evolution*, submitted.

[30] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. 1967 Perception of the speech code. *Psychological Review* **74**, 431–461.

[31] Lieberman, P. 1984 *The Biology and Evolution of Language.* Cambridge, MA: Harvard University Press.

[32] Mandelbrot, B. 1958 Les lois statistique macroscopiques du comportment. *Psychol. Francaise* **3**, 237–249.

[33] Maynard Smith, J. 1982 *Evolution and the Theory of Games.* Cambridge: Cambridge University Press.

[34] Maynard Smith, J. & Szathmary, E. 1995 *The Major Transitions in Evolution.* Oxford; New York: W. H. Freeman Spektrum.

[35] Maynard Smith, J. & Szathmary, E. 1999 *The Origins of Life.* Oxford: Oxford University Press.

[36] Miller, G. A. 1967 *The Psychology of Communication.* London: Penguin Books.

[37] Miller, G. A. 1981 *Language and Speech.* San Francisco: W. H. Freeman and Company.

[38] Miller, G. A. 1991 *The Science of Words.* New York: Scientific American Library.

[39] Miller, G. A. & Chomsky, N. 1963 Finitary models of language users. In Luce, R. D., Bush, R., & Galanter, E. (eds) *Handbook of Mathematical Psychology, Vol. 2.* New York: Wiley, 419–491.

[40] Nagy, W. E., Diadikoy, L., & Anderson, R. 1993 The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of Reading Behavior* **25**, 155–170.

[41] Nagy, W. E. & Anderson, R. C. 1984 How many words are there in printed school English? *Reading Research Quarterly* **19**, 304–330.

[42] Newmeyer, F. 1991 Functional explanation in linguistics and the origin of language. *Language and Communication* **11**, 3–96.

[43] Newport, E. 1990 Maturational constraints on language learning. *Cognitive Sci.* **14**, 11–28.

[44] Niyogi, P. 1998 *The Informational Complexity of Learning.* Boston: Kluwer Academic Publishers.

[45] Niyogi, P. & Berwick, R. C. 1996 A language learning model for finite parameter spaces. *Cognition* **61**, 161–193.

[46] Nobre, A., Allison, T. & McCarthy, G. 1994 Word recognition in the human inferior temporal lobe. *Nature* **372**, 260–263.

[47] Nowak, M. A. & Krakauer, D. C. 1999 The evolution of language. *Proc. Nat. Acad. Sci. USA* **96**, 8028–8033.

[48] Nowak, M. A., Krakauer, D. C., & Dress, A. 1999 An error limit for the evolution of language. *Proc. R. Soc. Lond. B.* **266**, 2131–2136.

[49] Nowak, M. A., Plotkin, J. B., & Krakauer, D. C. 1999 The evolutionary language game. *J. theor. Biol.* **200**, 147–162.

[50] Nowak, M. A., Plotkin, J. B., & Jansen, V. A. A. 2000 The evolution of syntactic communication. *Nature* **404**, 495–498.

[51] Pinker, S. 1994 *The Language Instinct.* New York: William Morrow and Company.

[52] Pinker, S. 1999 *Words and Rules.* New York: Basic Books.

[53] Pinker S. & Bloom, P., & commentators 1990 Natural language and natural selection. *Behavioral and Brain Sciences* **13**, 707–784.

[54] Plotkin, J.B., & Nowak, M. A. 2000 Language evolution and information theory. *J. theor. Biol.* **205**: 147-159.

[55] 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.

[56] Smith, W. J. 1977 *The Behavior of Communicating.* Cambridge, MA: Harvard University Press.

[57] Steels, L. 1997 The synthetic modelling of language origins. *Evolution of Communication Journal* **1**.

[58] Trapa, P. E. & Nowak, M. A. 2000 Nash equilibria for an evolutionary language game. *Journal of Mathematical Biology*, **41**: 172-188.

[59] Von Frisch, K. 1967 *The Dance Language and Orientation of Bees*. Cambridge, MA: Harvard University Press.

[60] 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.

[61] Zipf, G. K. 1935 *The Psychobiology of Language*. Boston: Houghton-Mifflin.