J. theor. Biol. (2000) **205,** 147–159
doi:10.1006/jtbi.2000.2053, available online at http://www.idealibrary.com on IDEAL®

Language Evolution and Information Theory

JOSHUA B. PLOTKIN* AND MARTIN A. NOWAK

Institut for Advanced Study, Princeton, NJ 08540, *U.S.A.*

(*Received on* 21 *October* 1999, *Accepted in revised form* 24 *March* 2000)

This paper places models of language evolution within the framework of information theory. We study how signals become associated with meaning. If there is a probability of mistaking signals for each other, then evolution leads to an error limit: increasing the number of signals does not increase the fitness of a language beyond a certain limit. This error limit can be overcome by word formation: a linear increase of the word length leads to an exponential increase of the maximum fitness. We develop a general model of word formation and demonstrate the connection between the error limit and Shannon's noisy coding theorem.

1. Introduction

If we want to understand the evolution of primitive communication, then we should first consider how signals acquire specific meanings. In other words, we should explore how evolution can lead to an association between signals and objects of the world. Here "object" is used in a very broad sense to include everything which can be referred to.

In previous papers, we have used evolutionary game theory to study this question (Nowak & Krakauer, 1999; Nowak *et al*., 1999a). We have assumed that communication is of benefit to both speaker and listener. Correct communication leads to a payoff. Each individual is characterized by two matrices. The active matrix contains the probabilities that a speaker will use a certain signal when attempting to communicate about a certain object. The passive matrix contains the probabilities that a listener will associate a specific signal with a specific object. If there is a possibility of mistaking signals for each other then we obtain an error limit. Using more and more signals cannot increase the fitness of a language beyond a certain limit. In other words, natural selection will design a communication system that has only a small number of signals referring to a few important concepts. It seems that this is the case for animal communication, while human language is (almost) unlimited.

The mechanism that can overcome the error limit is word formation. We find that linearly increasing word length can lead to exponentially increasing maximum fitness (Nowak *et al*., 1999b). The evolution of word formation is a transition from an analog to a digital communication system. All human languages use a limited number of phonemes to generate a large number of words. Moreover, word formation is probably unique to human communication (Pinker, 1995; Miller, 1991). Bird song is certainly combinatorial as well, but the interpretation of bird song is most likely not combinatorial (Marler, 1970; Hauser, 1996).

*Author to whom correspondence should be addressed. E-mail: plotkin@ias.edu

0022–5193/00/130147 + 13 $35.00/0

An important feature which has been missing from previous investigations is the connection between language evolution and information theory as conceived by Shannon & Weaver (1949). We build this bridge in the present paper. We will develop a more realistic model of word formation and discuss how the payoff in our evolutionary language game is related to Shannon's error probability and to the capacity of a channel. Specifically, we show the relationship between Shannon's noisy coding theorem and our finding that word formation can overcome the error limit.

Information theory is a mathematical discipline devoted to a precise definition and understanding of "information" in the vernacular sense. Following the seminal works of Shannon, information theorists have used notions from probability theory to define uncertainty and information. Information theory addresses such questions as the maximum rate at which information can be transfered over a noisy (or imperfect) channel. In particular, a coding system (e.g. repeating each message five times) may be used to increase the fidelity of a noisy channel. Shannon's noisy coding theorem quantifies the benefits which coding systems may confer on an otherwise noisy communication channel.

Although information theory has been widely applied in today's "information age", it has seldom been used—nor was it conceived—in the setting of language evolution. Despite this fact, we shall see that Shannon's formalism provides an excellent framework for considering language evolution, and especially word formation. In fact, when properly placed in this framework, the theory of language evolution should benefit from the insights of information theorists over the past 50 years.

This paper contains seven sections. In Section 2, we will outline the basic model of language evolution. In Section 3, we present a new approach for describing word formation. In Section 4, some basic concepts of information theory are discussed. In Section 5, we compare these concepts of information theory, in particular Shannon's noisy coding theorem, with our model of word formation. In Section 6, we present a specific example of a very simple communication system and analyse it from the perspective of language evolution and information theory. Section 7 is a short conclusion.

2. Evolving Arbitrary Signals

Consider a population of individuals who 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 (Hurford *et al.*, 1998). 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 of L vs. L' is given by

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

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 vice 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. TRANSMISSION ERROR

Miscommunication can arise from errors that occur during the transmission of a signal. Such "transmission errors" can be described by a signal-error matrix U. The entry U_{ij} denotes the probability that, when a speaker sends signal i, the listener receives signal j. In this setting, assuming that two individuals share a common language $L = (P, Q)$, the payoff is defined by

$$F(L, L) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{m} P_{ij} U_{jk} Q_{ki}. \qquad (1)$$

Again, this payoff function reflects the sum of the information content which a speaker can convey accurately to a listener and vice versa.

2.2. THE ERROR LIMIT

Let us now calculate the maximum payoff that a language, L, can achieve. Suppose $n = m$. The maximum payoff will be obtained if P and Q are identical permutation matrices. (A permutation matrix has a single 1 entry per row and column, all other entries being 0.) Without loss of generality, we set $P_{ii} = Q_{ii} = 1$ for all i, and $P_{ij} = Q_{ji} = 0$ for all $i \neq j$. In this case, we obtain the payoff function

$$F(L, L) = \sum_{k=1}^{n} U_{kk}.$$

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 $s_{ij} \leqslant 1$. The probability of mistaking signal i for signal j quantifies how similar signal i is to j compared with all other signals: $u_{ij} = s_{ij}/\sum_{k=1}^{n} s_{ik}$. In these terms, the fitness of a common language can be expressed as

$$F(L, L) = \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{n} s_{ik}}.$$

We imagine that the signals of the language are embedded in some pre-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 $s_{ij} = f(d_{ij})$. It can be shown that in this situation the fitness is always bounded by some constant depending only on X and f, but not on n (Dress & Nowak, 2000). In other words, even as the signal repertoire of a language increases, the fitness cannot exceed a fixed value.

3. Word Formation

In Nowak *et al.* (1999b), we demonstrated that word formation can overcome the constraint of the error limit. We considered languages whose basic signals consist of m phonemes. The words of the language were all assumed to be l-phonemes long. For simplicity, we also assumed that a language includes all possible m^l words in its lexicon. The similarity between words α and β was defined by the product of the similarity of their phonemes. In other words,

$$S(\alpha, \beta) = \prod_{k=1}^{l} s_{\alpha^{(k)}\beta^{(k)}},$$

where $\alpha^{(k)}$ denotes the k-th phoneme of word α. Using this definition, we showed that the maximum fitness of the language increases exponentially with word length l. In this sense, word formation allows the language to overcome the error limit.

We now develop a more general framework for word-based language. A language will be described by four components: a *lexicon*, an *active matrix P*, a *passive matrix Q*, and a *phoneme error-matrix V*.

As before, our model is based upon 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. Specifically, let us denote the phonemes of the language by the set $\Phi = \{\phi_1, \ldots, \phi_m\}$. We denote the lexicon by some subset $\mathfrak{C} \subset \Phi^l$. We refer to the words in $\mathfrak{C}$ as the *lexicon* or *proper vocabulary* of the language. Let us denote the size of the lexicon by $n = |\mathfrak{C}|$ (i.e. n is the cardinality of the set $\mathfrak{C}$). 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 $n \times m^l$ stochastic matrix whose ij-th entry denotes the probability that a speaker will attempt to use word j to denote object i. By definition, non-zero entries in P may occur only at columns corresponding to words in the lexicon $\mathfrak{C}$.

The passive matrix Q maps all possible perceived words (probabilistically) back into the n objects. We specify the passive matrix via a stochastic $m^l \times n$ matrix Q. The entry Q_{ji} represents the probability that a listener who perceives the j-th word will interpret it as the i-th object.

Finally, we must provide a description of transmission errors. As before, we use an $m^l \times m^l$ word error-matrix U. The entry U_{ij} denotes the probability that, when a speaker attempts to vocalize the i-th word, the listener perceives the j-th 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 $m \times m$ *phoneme error-matrix* V. The entry V_{ij} denote the probability that, when a speaker attempts to vocalize the i-th phoneme, the listener hears the j-th phoneme. Therefore, as before, for words α and β, we have the following expression for the word error-matrix (notice that, since V is stochastic, U is as well):

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

Thus, a language L is described completely by the three matrices $L = (P, Q, V)$. The matrix U is derived from V, and $\mathfrak{C}$ is determined by those columns of P-containing non-zero 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 i-th object, we define

$$\begin{aligned} F(L, L') &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} \sum_{\beta \in \Phi^l} P_{w_i\alpha} U_{\alpha\beta} Q_{\beta w_i} \\ &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i\alpha} \sum_{\beta \in \Phi^l} Q_{\beta w_i} \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}}. \qquad (2) \end{aligned}$$

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, this inquiry addresses a fundamental question regarding the adaptive benefits of word formation.

In order to answer these questions, we will first take a detour into the information theory of Shannon. We will use Shannon's classical theorem on noisy communication to show that word formation does, indeed, remove the error limit which constrains strictly phonemic communication. In fact, as we shall see, word formation provides an exponential increase in fitness with word length l. This result places our evolutionary theory of language within the larger framework of information theory.

4. Shannon's Information Theory

We will outline some of the primary components of Shannon's theory as needed for our purposes. For a detailed accounts of information theory, we refer to Welsh (1988) and van der Lubbe (1988).

4.1. THE NOISY CHANNEL

Shannon considers a discrete memoryless source $\mathfrak{I}$ which emits characters from an *alphabet* $\Phi = \{\phi_1, \ldots, \phi_m\}$ according to some discrete probability distribution. Most often, Shannon considers the special case of a binary alphabet, but his noisy coding theorem applies to arbitrary alphabets as well. The discrete source $\mathfrak{I}$ 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 $\Pr(\phi_j \text{ received} \mid \phi_i \text{ sent})$.

Given a channel V and an input source, we obtain a natural output stream. In this situation, Shannon introduces the notion of the *capacity* of V. 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. (A more precise definition will be given in Section 6.) Given a channel V with capacity $C(V)$, Shannon asks how one can improve the reliability of the communication system by constructing codes while simultaneously keeping the required number of transmissions small.

4.2. ENCODING AND DECODING

In order to increase fidelity, Shannon defines a set of n *codewords*, $\mathfrak{C}$, each codeword being a string of l characters from Φ. The *encoder* takes input messages from the source $\mathfrak{I}$, encodes the information into codewords, and sends the codeword on to the noisy channel, letter by letter. Shannon also requires the specification of a deterministic *decoder*. The *decoder* is a map from all possible outputs (from the noisy channel) back to $\mathfrak{C}$. In other words, the decoder is a partition of Φ^l into n disjoint subsets. (Of course, any good decoder will certainly include each codeword w within the subset of Φ^l which is decoded as w.)

Shannon defines the *error probability* of this communication system (Fig. 1) as

$$e(\mathfrak{C}) = \frac{1}{n} \sum_{i=1}^{n} \Pr(\text{error in communication} \mid \text{codeword } w_i \text{ is transmitted}).$$

Thus, the error probability measures the average number of mis-interpreted codewords, assuming codewords are transmitted with equal probabilities. 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.

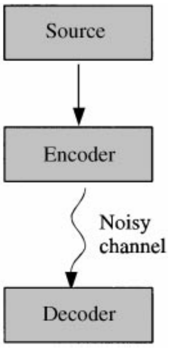

FIG. 1. Schematic diagram of Shannon's communication system. A source emits messages which are encoded into codewords, transmitted over a noisy channel, and then decoded.

4.3. THE NOISY CODING THEOREM

In this situation, Shannon's noisy coding theorem states the following:

Theorem 4.1 (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 $(\mathfrak{C}_n \mid 1 \leqslant n < \infty)$ such that*
(a) *$\mathfrak{C}_n$ has $2^{\lfloor Rn \rfloor}$ codewords of length $l = n$,*
(b) *the error probability satisfies $e(\mathfrak{C}_n) \leqslant A\mathrm{e}^{-Bn}$, where the constants A and B depend only on the channel V and on R.*

In other words, 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. [For a proof of Shannon's theorem, we refer the reader to Gallager (1968). In essence, Shannon constructs each successive code $\mathfrak{C}_n$ 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. Specifically, we have the following result.

Theorem 4.2 (Wolfowitz, 1961). *For a discrete memoryless channel of capacity C and for any $R > C$, there cannot exist a sequence of codes $\mathfrak{C}_n$ such that $\mathfrak{C}_n$ has 2^{Rn} codewords of length n and error probability tending to zero. In fact, such a sequence of codes must have error probability which approaches 1 as $n \to \infty$.*

5. Information Theory and Word-based Language

In this section, we reinterpret Shannon's communication system and the noisy coding theorem in terms of our language model.

In order to relate Shannon's theory to our language model, we make two trivial remarks.

Notice first that a Shannon-encoder may be expressed as a binary $n \times m^l$ 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 $m^l \times n$ matrix Q. The entry Q_{ji} denotes whether or not the j-th word is included in the subset words decoded as the i-th codeword (or i-th 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 4.1 to derive the maximal fitness of our word-based model.

Towards this end, consider the information-theoretic expression $\tilde{F}(\mathfrak{C}) = |\mathfrak{C}|(1-e(\mathfrak{C})) = n(1-e(\mathfrak{C}))$. By Shannon's theorem, given a channel V with non-zero capacity, we can find a sequence of codes $\mathfrak{C}_n$ with linearly increasing codeword length and with exponentially increasing $\tilde{F}(\mathfrak{C})$. Thus, Shannon's theorem (together with its converse) reveals the maximal properties of $\tilde{F}(\mathfrak{C})$.

We will show that $\tilde{F}(\mathfrak{C})$ is equivalent to the fitness of language in our evolutionary model. Thus, we claim

$$\tilde{F}(\mathfrak{C}) = F(L, L).$$

The proof of this statement is little more than an exercise in unraveling definitions. We start with the definition of $\tilde{F}$:

$$\begin{aligned}\tilde{F}(\mathfrak{C}) &= n[1 - e(\mathfrak{C})]\\ &= n\left[1 - \frac{1}{n}\sum_{i=1}^{n} \Pr(\text{communication error}|\right.\\ &\qquad\qquad \left. w_i \text{ transmitted})\right]\\ &= n - \sum_{i=1}^{n} (1 - \Pr(\text{no communication error}|\\ &\qquad\qquad w_i \text{ transmitted}))\\ &= \sum_{i=1}^{n} \Pr(\text{no communication error}|\\ &\qquad\qquad w_i \text{ transmitted}).\end{aligned}$$

Recall that in Shannon's codeword system, P and Q are binary matrices whose rows sum to 1. Therefore, given $w \in \mathfrak{C}$, $P_{w\alpha} = 1$ only when $\alpha = w$. Thus, we may rewrite the last equality above as follows:

$$\tilde{F}(\mathfrak{C}) = \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i\alpha} \Pr(\text{no communication error}|\alpha \text{ transmitted}).$$

In order to calculate the probability of correct communication when α is transmitted, we consider all possible outputs β from the noisy channel, weighted by their respective probabilities. For each output β, correct communication occurs only if β is decoded as codeword w_i. Thus, we derive the equations

$$\begin{aligned}\tilde{F}(\mathfrak{C}) &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i\alpha} \sum_{\beta \in \Phi^l} \Pr(\beta \text{ received}|\alpha \text{ transmitted})\Pr(\beta \text{ is decoded as codeword } w_i)\\ &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i\alpha} \sum_{\beta \in \Phi^l} \Pr(\beta \text{ received}|\alpha \text{ transmitted}) Q_{\beta w_i}.\end{aligned}$$

Finally, we must calculate the probability that, upon input α into the noisy channel, we receive output β. The noisy channel produces its output phoneme-by-phoneme (or "letter by letter"). Therefore, the probability we want equals the product of the probabilities that each phoneme of α will be transmitted as the respective phoneme of β. Such quantities are given by the channel matrix V. Thus, we see that

$$\tilde{F}(\mathfrak{C}) = \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w_i\alpha} \sum_{\beta \in \Phi^l} Q_{\beta w_i} \prod_{k=1}^{l} V_{\alpha^{(k)}\beta^{(k)}}.$$

But this last formula coincides with our expression for the language fitness $F(L, L)$, defined in eqn (2). Hence, we have shown $\tilde{F}(\mathfrak{C}) = F(L, L)$.

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 $\tilde{F}(\mathfrak{C})$. As a consequence, Shannon's coding theorem implies the following result.

Theorem 5.1 (word formation)**.** *Given a phoneme error-matrix V (with non-zero capacity), there exists a sequence of languages L_n with linearly increasing word length and exponentially increasing fitness.*

Thus, word formation overcomes the error limit which constraints 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.

Note that this result is not limited by the restriction to binary active and passive matrices; as is usual in linear programming, we know *a priori* that a maximally fit language must use binary matrices.

As we have seen, information theory helps to answer fundamental questions about our model of word-based language. In one sense, however, our model is more general than Shannon's communication system; P and Q are not necessarily deterministic, and they are not necessarily shared by the speaker and listener. This generality allows us to view language in an evolutionary context, as we shall explore in the following extended example.

6. A Specific Example

In order to illustrate our somewhat abstract model of language—as well as its relationship to information theory—we now present a specific example. Let us assume that the individuals in a population all speak the same language. In this section, we will step through a complete specification of the language L and eventually calculate $F(L, L)$. Then, we will reconsider the example in light of information theory and Theorem 5.1.

6.1. THE PHONEMES

We first stipulate that the organisms in the population are all limited, by their similar vocal and auditory organs, to the use of three phonemes. We denote these phonemes by a, m and p. Therefore, $\Phi = \{\mathsf{a}, \mathsf{m}, \mathsf{p}\}$ and $m = 3$. We also assume the population uses a language with word length $l = 2$. Therefore, the possible words which an individual can utter are $\Phi^l = \{\mathsf{aa}, \mathsf{am}, \mathsf{ap}, \mathsf{ma}, \mathsf{mm}, \mathsf{mp}, \mathsf{pa}, \mathsf{pm}, \mathsf{pp}\}$. Nevertheless, the population chooses only to describe three objects: father, mother and food. Hence, the population uses a lexicon of $n = 3$ proper vocabulary words. Specifically, we stipulate that the population uses the lexicon $\mathfrak{C} = \{\mathsf{pa}, \mathsf{ma}, \mathsf{mm}\} \subset \Phi^l$.

6.2. THE ACTIVE MATRIX

Next we describe the active matrix, P, of the common language. We assume that, generally speaking, a speaker attempts to use pa to convey father, ma to convey mother and mm to convey food. Nevertheless, the speaker does not have a deterministic (binary) active matrix. Instead, there is always a slight probability that the speaker will try to describe an object with the "wrong" vocabulary word. This leads to the following expression for the P-matrix:

$$P = \begin{array}{c|ccccccccc} & \mathsf{aa} & \mathsf{am} & \mathsf{ap} & \mathsf{ma} & \mathsf{mm} & \mathsf{mp} & \mathsf{pa} & \mathsf{pm} & \mathsf{pp} \\ \hline \mathsf{Mother} & 0 & 0 & 0 & 1-2\varepsilon & \varepsilon & 0 & \varepsilon & 0 & 0 \\ \mathsf{Food} & 0 & 0 & 0 & \varepsilon & 1-2\varepsilon & 0 & \varepsilon & 0 & 0 \\ \mathsf{Father} & 0 & 0 & 0 & \varepsilon & \varepsilon & 0 & 1-2\varepsilon & 0 & 0 \end{array}.$$

Notice that the columns of P with non-zero entries correspond to proper vocabulary words. In other words, when a speaker wants to communicate food, there is no chance that he will attempt to use a word outside of his language's lexicon. The active matrix represents the speaker's association between objects and *proper* vocabulary words.

6.3. THE TRANSMISSION ERROR MATRICES

Despite the form of the active matrix, there is always a chance that when trying to communicate an object, the *perceived* output will not fall in the lexicon. This phenomenon arises from errors in vocalizing the intended word or hearing the output correctly. Such a phenomenon is a *transmission* error, as opposed to a *interpretation* error, and it is quantified by the word error-matrix U.

In order to specify U, we must first define the phoneme error-matrix V. For illustrative purposes we derive V (and thus eventually U) by using the notion of phoneme similarity. We embed our phonemes into the compact metric space $X = [0, 1]$, and we define their pairwise similarity as a declining function f of their pairwise distances.

In particular let us embed $\Phi \rightarrow [0, 1]$ by placing p at 0, m at 1/4, and a at 3/4. We have chosen this particular embedding for purely illustrative purposes; there has been no attempt to reflect the actual phonetics of m, p, and a in the English language (Fig. 2).

We define phoneme similarity via a simple, exponentially decreasing function of distance: $s_{ij} = f(d_{ij}) = e^{-5d_{ij}}$. For example, $s_{\mathsf{a,m}} = e^{-5d_{\mathsf{a,m}}} = e^{-5|3/4-1/4|} = e^{-5/2} \approx 0.082$. Similarly, $s_{\mathsf{a,a}} = e^{-5d_{\mathsf{a,a}}} = e^0 = 1$. This leads us to the following values for phoneme similarity. (For ease of presentation, we henceforth round all reported values to two decimal places. Nevertheless, all of our calculations employ the precise parameter values.)

$$s_{\mathsf{a,a}} = s_{\mathsf{m,m}} = s_{\mathsf{p,p}} = 1,$$

$$s_{\mathsf{a,m}} = s_{\mathsf{m,a}} \approx 0.08,$$

$$s_{\mathsf{a,p}} = s_{\mathsf{p,a}} \approx 0.02,$$

$$s_{\mathsf{m,p}} = s_{\mathsf{p,m}} \approx 0.29.$$

As in Section 3, we use the similarity between phonemes to define the phoneme error matrix $V_{ij} = s_{ij}/\sum_{k=1}^{m} s_{ik}$. For example, $V_{\mathsf{a,a}} \approx 1/(1+0.08+0.02) = 0.90$. This leads to the following stochastic phoneme error matrix V:

$$V = \begin{array}{cccc} & \mathsf{a} & \mathsf{m} & \mathsf{p} \\ \mathsf{a} & 0.90 & 0.07 & 0.02 \\ \mathsf{m} & 0.06 & 0.73 & 0.21 \\ \mathsf{p} & 0.02 & 0.22 & 0.76 \end{array}.$$

(Notice that, if a row of a stochastic matrix appears not to sum to one, this is only an artifact of our two-decimal presentation.)

Although the matrix above is stochastic, it is not symmetric. For example, given V above, it is more likely to mis-speak or mis-hear the phoneme p as m than it is to slip from m to p. Upon a moment's reflection, this situation is realistic; certain phonemes lend themselves more easily to transmission error into others than vice versa.

Using the V matrix, we can derive the corresponding word-error matrix U. For each pair of words we compare constitutive phonemes in turn. For example, $U_{\mathsf{ma,aa}} = V_{\mathsf{m,a}} V_{\mathsf{a,a}} \approx 0.06 \cdot 0.90 = 0.05$. Of course, the only rows of U which matter correspond to words with positive probability in P—i.e. to the lexicon words. Thus, we only report such rows below. This leads to the following U matrix:

$$U = \begin{array}{cccccccccc} & \mathsf{aa} & \mathsf{am} & \mathsf{ap} & \mathsf{ma} & \mathsf{mm} & \mathsf{mp} & \mathsf{pa} & \mathsf{pm} & \mathsf{pp} \\ \mathsf{ma} & 0.05 & 0.00 & 0.00 & 0.66 & 0.05 & 0.02 & 0.19 & 0.02 & 0.00 \\ \mathsf{mm} & 0.00 & 0.04 & 0.01 & 0.04 & 0.53 & 0.15 & 0.01 & 0.15 & 0.04 \\ \mathsf{pa} & 0.02 & 0.00 & 0.00 & 0.20 & 0.02 & 0.00 & 0.69 & 0.06 & 0.02 \end{array}.$$

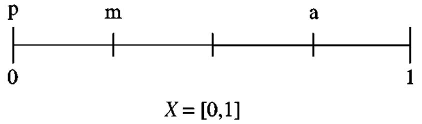

FIG. 2. Schematic diagram of the embedding of phonemes $\Phi = \{\mathsf{a}, \mathsf{m}, \mathsf{p}\}$ into the metric space $X = [0, 1]$. Once embedded, phoneme similarity is defined by a decreasing function of distance.

6.4. THE PASSIVE MATRIX

Finally, we must specify the passive matrix Q. This matrix provides, for each possible perceived output word, the probability of interpreting that word as each respective object. For example, $Q_{\mathsf{aa},mother}$ denotes the probability of interpreting perceived output word aa as the object mother. Given P and U, what is a reasonable choice of listening matrix Q?

We will soon derive the optimal, deterministic choice of Q. For now, however, we follow a reasonable rule of interpretation: a listener should interpret perceived output word α as object i with a probability which equals the probability that, when trying to communicate object i, the perceived output would be α. In other words, we set $Q_{\alpha,i} = \Pr(\text{output } \alpha | i \text{ transmitted})$. For example, we set $Q_{\mathsf{aa},mother}$ equal to the chance of producing output aa when trying to communicate mother. In particular, this probability is given by $(1-2\varepsilon)U_{\mathsf{ma,aa}} + \varepsilon U_{\mathsf{mm,aa}} + \varepsilon U_{\mathsf{pa,aa}} \approx 0.05(1-2\varepsilon) + 0.00\varepsilon + 0.02\varepsilon \approx 0.05 - 0.09\varepsilon$. Following this rule alone, however, the rows of the resulting Q matrix do not necessarily sum to one. Thus, we normalize each row to derive the following, non-deterministic Q-matrix:

$$
\begin{array}{rlll}
 & \text{Mother} & \text{Food} & \text{Father} \\
\mathsf{aa} & 0.73-1.2\varepsilon & 0.05+0.85\varepsilon & 0.22+0.34\varepsilon \\
\mathsf{am} & 0.09+0.73\varepsilon & 0.89-1.65\varepsilon & 0.03+0.92\varepsilon \\
\mathsf{ap} & 0.09+0.73\varepsilon & 0.88-1.65\varepsilon & 0.03+0.92\varepsilon \\
\mathsf{ma} & 0.73-1.2\varepsilon & 0.05+0.85\varepsilon & 0.22+0.34\varepsilon \\
Q = \mathsf{mm} & 0.09+0.73\varepsilon & 0.88-1.65\varepsilon & 0.03+0.92\varepsilon \\
\mathsf{mp} & 0.09+0.73\varepsilon & 0.88-1.65\varepsilon & 0.03+092\varepsilon \\
\mathsf{pa} & 0.21+0.36\varepsilon & 0.01+0.96\varepsilon & 0.77-1.3\varepsilon \\
\mathsf{pm} & 0.07+0.79\varepsilon & 0.68-1.04\varepsilon & 0.25+0.25\varepsilon \\
\mathsf{pp} & 0.07+0.79\varepsilon & 0.68-1.04\varepsilon & 0.25+0.25\varepsilon.
\end{array}
\tag{3}
$$

6.5. THE PAYOFF

Since the language $L = (P, Q, V)$ has now been completely specified, we may calculate its payoff via eqn (2):

$$
\begin{aligned}
F(L, L) &= \sum_{i=1}^{n} \sum_{\alpha \in \Phi^l} P_{w,\alpha} \sum_{\beta \in \Phi^l} Q_{\beta w_i} U_{\alpha\beta} \\
&= \sum_{\alpha \in \Phi^l} P_{mother,\alpha} \sum_{\beta \in \Phi^l} Q_{\beta,mother} U_{\alpha\beta} \\
&\quad + \sum_{\alpha \in \Phi^l} P_{food,\alpha} \sum_{\beta \in \Phi^l} Q_{\beta,food} U_{\alpha\beta} \\
&\quad + \sum_{\alpha \in \Phi^l} P_{father,\alpha} \sum_{\beta \in \Phi^l} Q_{\beta,father} U_{\alpha\beta} \\
&\approx 1.96 - 5.79\varepsilon + 8.68\varepsilon^2.
\end{aligned}
$$

Having derived an expression for the fitness of the language L, we may now inspect its properties. The only free parameter is ε, which is a measure of interpretive error. (By completely specifying V, on the other hand, we have fixed the amount of transmission error.) We may only allow ε to vary in $[0, 0.5]$ so that P remain a stochastic matrix. Figure 3 shows the graph of $F(L, L)$ for this range of ε values.

As expected, the maximum fitness occurs when ε equals zero—i.e. when there are no errors in interpretation. In general, as we usually find in linear programming problems, the maximal fitness always occurs when P and Q are binary,

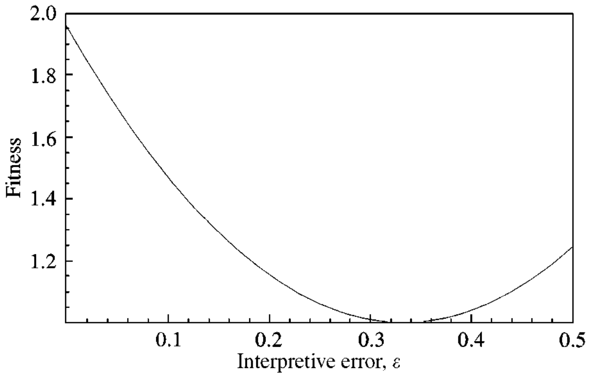

FIG. 3. Graph of the language fitness $F(L, L)$ obtained as a function of ε. The parameter ε measures the amount of interpretive error in the language. The fitness of the language L is maximized when there is no chance for misinterpretation ($\varepsilon = 0$).

deterministic matrices. Also notice that the minimum fitness occurs when $\varepsilon = 1/3$; in this situation each speaker has an indiscriminate association between objects and lexicon words.

6.6. IN THE CONTEXT OF INFORMATION THEORY

We now reconsider the example in the context of information theory. The phoneme error-matrix V now plays the role of the channel matrix. Similarly, we use the original lexicon as the code-words. The encoder and decoder, however, must now be binary. Therefore, in a strict, information theoretic context, we must take $\varepsilon = 0$. In this context, we should also use deterministic, *maximum likelihood* decoding, as in the proof of Shannon's theorem. Specifically, Q must decode an output word α as that particular codeword i which maximizes the conditional probability Pr(output $\alpha | i$ transmitted). Recall that in eqn (3) we defined $Q_{\alpha,i} = \Pr(\text{output } \alpha | i$ transmitted), and normalized each row. Therefore, the maximum likelihood decoder simply replaces each row of the non-deterministic Q with zeroes, except for the largest entry in the row. More explicitly, for each $\alpha \in \Phi^l$, the deterministic, maximum-likelihood decoder Q^{ML} satisfies

$$Q^{ML}_{\alpha,i} = \begin{cases} 1 & \text{when } \alpha \text{ maximizes} \\ & \quad \Pr(\text{output } \alpha | i \text{ transmitted}) \\ 0 & \text{otherwise.} \end{cases}$$

$$= \begin{cases} 1 & \text{when } \alpha \text{ maximizes } \quad \sum_{\beta \in \Phi^l} P_{i\beta} U_{\beta\alpha} \\ 0 & \text{otherwise.} \end{cases}$$

In our particular example, when $\varepsilon = 0$ we have the following maximum-likelihood decoder Q^{ML}:

$$Q^{ML} = \begin{array}{cccc} & \text{Mother} & \text{Food} & \text{Father} \\ \text{aa} & 1 & 0 & 0 \\ \text{am} & 0 & 1 & 0 \\ \text{ap} & 0 & 1 & 0 \\ \text{ma} & 1 & 0 & 0 \\ \text{mm} & 0 & 1 & 0 \\ \text{mp} & 0 & 1 & 0 \\ \text{pa} & 0 & 0 & 1 \\ \text{pm} & 0 & 1 & 0 \\ \text{pp} & 0 & 1 & 0 \end{array} . \tag{4}$$

(Notice, for example, that mp is decoded a food (mm) as opposed to mother (ma). This makes sense given that p is phonetically closer to m than to a.)

When $\varepsilon = 0$ and when we use maximum-likelihood decoding, we calculate that $F(L, L) \approx 2.35$. This maximum fitness is significantly larger than 1.96, which was the fitness obtained via the non-deterministic encoder of eqn (3). In fact, as ε varies, we can evaluate the fitnesses obtained via the maximum-likelihood vs. the non-deterministic encoders. A graph of these fitnesses is shown in Fig. 4.

Notice that it is always better (except when $\varepsilon = 1/3$) to use the Shannon-type decoder Q^{ML} than the non-deterministic matrix defined in eqn (3). This is a graphic illustration of the optimality ensured by Shannon's theorem.

6.7. THE CAPACITY CALCULATION

In this section, we compute the capacity of the 3×3 phoneme error-matrix V. Recall that Shannon's theorem only applies to channels with $C(V) > 0$.

First, we must state the precise definition of capacity. Given a channel V and an input-source $\mathfrak{I}$, we obtain a natural output stream $\mathfrak{J}$. The capacity of the channel V is defined by

$$C(V) = \sup_{\mathfrak{I}} [H(\mathfrak{I}) + H(\mathfrak{J}) - H(\mathfrak{I}, \mathfrak{J})],$$

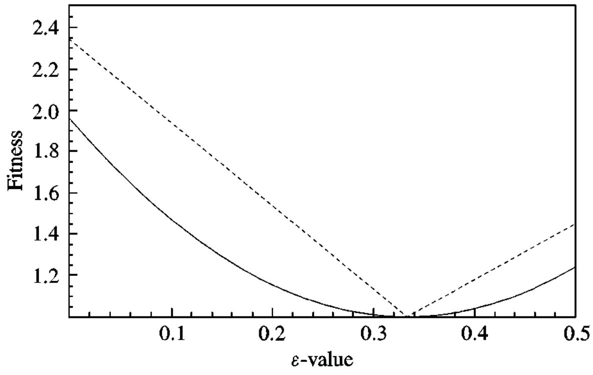

FIG. 4. Graph of the language fitness $F(L, L)$ obtained via the non-deterministic decoder Q as opposed to the deterministic, maximum-likelihood decoder Q^{ML}. A language is always better served by the maximum-likelihood decoder. Thus, we expect that languages should evolve towards maximum-likelihood decoding. (- - - -) Shannon decoder (Q^{ML}); (——) non-deterministic decoder (Q).

where H denotes the entropy of a source. The entropy of a discrete source $\Im$ with probability distribution $(p_1, \ldots, p_m)$ is defined as $H(\Im) = -\sum_{i=1}^{m} p_i \log(p_i)$. See Welsh (1988) for a description of entropy.

In our particular example, a source $\Im$ is determined by a discrete probability distribution (x, y, z), $x + y + z = 1$, where x denotes the probability that the source will emit letter a, y denotes the probability for m, and z for p. The output stream $\mathfrak{J}$ is calculated as follows; the probability of an a is $xV_{11} + yV_{21} + zV_{31}$, and similarly for the other two letters. The stream $(\Im, \mathfrak{J})$ has nine possible "letters": a followed by a, a followed by m, etc. These probabilities are given by xV_{11}, xV_{12}, etc.

Once the probability distributions for $\Im$, $\mathfrak{J}$, and $(\Im, \mathfrak{J})$ have been expressed in terms of (x, y, z), we must maximize the function $H(\Im) + H(\mathfrak{J}) - H(\Im, \mathfrak{J})$ over all input distributions for $\Im$. This maximization amounts to nothing more than a straightforward calculus problem via Lagrange multipliers. The resulting answer is given by

$$C(V) \approx 0.7988,$$

which is obtained when $(x, y, z) \approx (0.43,\ 0.19,\ 0.38)$.

Thus, since $C(V) > 0$, Shannon's theorem indeed applies to our explicit example. In particular, Shannon's theorem guarantees a sequence of languages L_n, $n = 1, 2, 3, \ldots$, each with a lexicon of $2^{\lfloor 0.79n \rfloor}$ words of length $l = n$, with exponentially increasing fitnesses.

6.8. EVOLUTION TOWARDS DETERMINISM

We can use evolutionary dynamics to test if the language will evolve toward deterministic passive matrices, as information theory predicts.

We ran a simple computer simulation to test whether, when V and P are fixed, Q will evolve towards Q^{ML}, the deterministic, maximum-likelihood decoder. In particular, referring to our explicit example, we fix $\varepsilon = 0$ and specify P and V as above. We choose a population of 50 asexual, semelperous individuals. In the first generation, each individual starts with the non-deterministic Q-matrix from eqn (3). In each successive generation, we calculate the payoff obtained by each individual communicating with every other one. Each individual then produces progeny (each progeny with the same Q-matrix as the parent) in proportion to its total payoff relative to the other individual's payoffs. We normalized so that there are 50 offspring, and thus 50 individuals, in every generation. At each generation and for each offspring, we stipulate a 0.4% chance that the offspring will be a "mutant". A mutant offspring possesses the same Q-matrix as its parent, but with each entry perturbed by a random value in $[-0.212, 0.212]$. (By normalizing, we ensure that a mutant's resulting Q-matrix has nonnegative values with rows summing to one.)

The result of this evolutionary simulation is summarized in Fig. 5. We have graphed the average payoff in the population at each of 5000 generations. The average payoff starts at 1.96, as derived in Section 6.5. Notice that the average payoff increases over time, approaching the Shannon-limit of 2.35, derived in Section 6.6.

We also report a typical Q-matrix found in the population after 5000 generations. This Q-matrix compares favorably with the matrix Q^{ML} of eqn (4):

$$Q_{5000} = \begin{array}{cccc} & \text{Mother} & \text{Food} & \text{Father} \\ \text{aa} & 1 & 0 & 0 \\ \text{am} & 0.19 & 0.81 & 0 \\ \text{ap} & 0.08 & 0.92 & 0 \\ \text{ma} & 0.99 & 0 & 0.01 \\ \text{mm} & 0 & 1 & 0 \\ \text{mp} & 0.09 & 0.91 & 0 \\ \text{pa} & 0 & 0 & 1 \\ \text{pm} & 0.13 & 0.77 & 0 \\ \text{pp} & 0 & 0.99 & 0.01 \end{array}.$$

As this simulation reveals, although the evolutionary dynamics of word-based language are

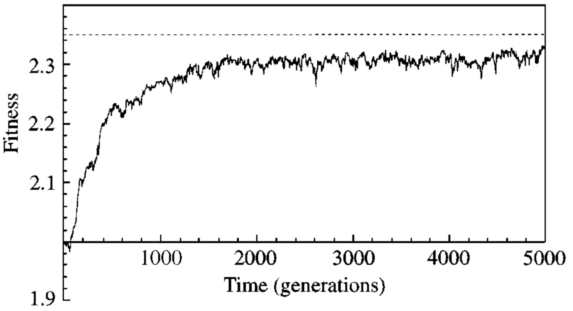

FIG. 5. Graph of the average payoff in a population of individuals during the course of simulated language evolution. The population begins with non-deterministic decoding and evolves towards maximum-likelihood decoding. This evolution towards determinism is reflected by the average payoff of the population, which ascends towards the Shannon limit of 2.35. Thus, the long-term, complex dynamics of language evolution are predicted by information theory. (- - - -) Shannon limit; (——) simulation.

complicated and noisy, we can still understand their long-term behavior via the bounds imposed by information theory.

7. Conclusions

We have compared models of language evolution with concepts from information theory. In particular, there is a connection between Shannon's noisy coding theorem and our results on word formation. Shannon's theorem states that, for a given noisy channel, there exists a sequence of codes with linearly increasing codeword length such that the probability of transmission error decreases exponentially. Our result on word formation states that, for a given phonemic error matrix, the maximum fitness of a language increases exponentially with word length. We demonstrated that Shannon's error probability is inversely proportional to our fitness function. Hence, the equivalence becomes obvious.

Although the maximum fitness of a language increases with word length, evolution will not lead to a run-away sequence of languages with longer and longer words. Clearly, there are natural restraints on this tendency: as word length increases, memorization difficulties increase and the rate of communication decreases.

Shannon's theory requires a deterministic encoder and decoder (equivalent to binary P and Q matrices in our notation). Any errors thus result from noisy transmission, which is equivalent in our terms to acoustic errors during communication: the sender emits signal A, but the receiver hears signal B. We believe that word formation was the crucial evolutionary invention to overcome this kind of transmission error. Instead of increasing the number of phonemes in a language, our ancestors invented a combinatorial signaling system. In this sense, word formation compensates for errors in signal transmission.

There are, however, other kinds of errors which are not captured by Shannon's basic model. These are coordination errors between the implementation of a signal by the sender and the interpretation of a signal by the receiver. We believe that these errors cannot be overcome by word formation *per se*, but that they instead require a sophisticated organization of the mental lexicon and, importantly, the invention of syntax. As the size of the mental lexicon reaches some memory capacity it becomes feasible to represent every message by an individual word. Instead, sentences comprised of individual words are required. Hence, there should be an error limit for words which is overcome by the use of sentences. This topic—the evolution of syntax—requires further investigation (Nowak *et al.*, 2000).

REFERENCES

DRESS, A. & NOWAK, M. A. (2000) (in preparation).

GALLAGER, R. G. (1968). *Information Theory and Reliable Communication*. New York: Wiley.

HAUSER, M. (1996). *The Evolution of Communication*. Cambridge, MA: MIT Press.

HOFBAUER, J. & SIGMUND, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

HURFORD, J., STUDDERT-KENNEDY, M. & KNIGHT, C. (1998). *Approaches to the Evolution of Language*. Cambridge: Cambridge University Press.

MARLER, P. (1970). Birdsong and speech development: could there be parallels? *Am. Sci.* **58,** 669–673.

MILLER, G. (1991). *The Science of Words*. New York: Scientific American Library.

NOWAK, M. A. & KRAKAUER, D. C. (1999a). The Evolution of Language. *Proc. Nat. Acad. Sci. U.S.A.* **96,** 8028.

NOWAK, M. A., PLOTKIN, J. B. & KRAKAUER, D. C. (1999a). The Evolutionary Language Game. *J. theor. Biol.* **200,** 147–162.

NOWAK, M. A., KRAKAUER, D. C. & DRESS, A. (1999b). An error-limit for the evolution of language. *Proc. Roy. Soc. London B* **266,** 2131–2136.

NOWAK, M. A., PLOTKIN, J. B. & JANSEN, V. A. A. (2000). The evolution of syntactic communication. *Nature* **404,** 495–498.

PINKER, S. (1995). *The Language Instinct*. New York: Penguin.

SHANNON, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* **27,** 379–423, 623–656.

SHANNON, C. E. & WEAVER, W. (1949). *The Mathematical Theory of Communication*. Illinois: University of Illinois Press.

VAN DER LUBBE, J. (1988). *Information Theory*. Cambridge: Cambridge University Press.

WELSH, D. (1988). *Codes and Cryptography*. Oxford University Press.

WOLFOWITZ, J. (1961). *Coding theorems of information theory*. Berlin: Springer-Verlag.