J. theor. Biol. (2003) **221,** 445–457
doi:10.1006/jtbi.2003.3199, available online at http://www.idealibrary.com on IDEAL®

Language Dynamics in Finite Populations

NATALIA L. KOMAROVA†‡ AND MARTIN A. NOWAK*†

†*Institute for Advanced Study, Einstein Drive, Princeton, NJ* 08540, *U.S.A. and*
‡*Department of Applied Mathematics, University of Leeds, Leeds LS*2 9*JT, U.K.*

(*Received on* 17 *April* 2002, *Accepted in revised form on* 22 *October* 2002)

Any mechanism of language acquisition can only learn a restricted set of grammars. The human brain contains a mechanism for language acquisition which can learn a restricted set of grammars. The theory of this restricted set is universal grammar (UG). UG has to be sufficiently specific to induce linguistic coherence in a population. This phenomenon is known as "coherence threshold". Previously, we have calculated the coherence threshold for deterministic dynamics and infinitely large populations. Here, we extend the framework to stochastic processes and finite populations. If there is selection for communicative function (selective language dynamics), then the analytic results for infinite populations are excellent approximations for finite populations; as expected, finite populations need a slightly higher accuracy of language acquisition to maintain coherence. If there is no selection for communicative function (neutral language dynamics), then linguistic coherence is only possible for finite populations.

1. Introduction

Grammar is the computational system of human language. It is a machinery that allows us to make "infinite use of finite means". Each of the 6000 human languages currently spoken contains a finite number of phonemes, but can generate infinitely many sentences.

Children learn grammar by interacting with speakers of their native language. They receive sample sentences and construct a sub-conscious representation of the underlying grammatical rules. Such a learning process is based on generalization as opposed to memorization. The learner "generalizes" from a finite number of sentences to a rule system that generates a much larger (and in the case of human language, an infinite) number of sentences. Generalization is equivalent to "looking for rules". Generalization is only possible if the learner has a limited search space. This means the learner can only acquire a restricted set of languages. Universal grammar (UG) provides a theory of the restricted set of languages learnable by the human brain in the context of language acquisition (Chomsky, 1972, 1981, 1984). More generally, UG is what is innate about human language. UG is what the child brings to the process of language acquisition. UG includes a procedure for language acquisition, sometimes called "language acquisition device", and a description of the restricted set of languages learnable by this device.

Consider a population of speakers with the same UG. In this paper, we will assume that UG

*Corresponding author. Tel.: +1-609-734-8389; fax: +1-609-951-4438.
E-mail address: nowak@ias.edu (M. A. Nowak).

admits only a finite number of languages, $L_1, \ldots, L_n$. This is, for example, in agreement with the principles-and-parameters theory of language acquisition (Chomsky, 1981; Baker, 2001): children learn the grammars of natural languages by setting binary parameters. For k parameters, there are $n = 2^k$ possible languages. A finite search space is also implied by optimality theory (Prince & Smolensky, 1997): children learn grammar by ranking a finite number of constraints. For k constraints, there are $n = k!$ languages. In principle, however, UG could also admit an infinite number of candidate languages. In this case, UG must contain a prior ranking or probability distribution specifying which languages are more likely than others. In this paper, we only consider a finite search space; the extension to infinite search spaces is possible.

We formulate language dynamics similar to evolutionary game dynamics. There is a population of individuals. They talk to each other. Successful communication results in a payoff, which contributes to biological fitness (Nowak *et al.*, 2001). Children learn the language of their parents. Alternatively we can assume that successful communication also leads to "cultural fitness" and that children are more likely to learn the language of individuals with a high payoff.

We will also study the case where language does not affect any fitness. Let us coin the term *neutral language dynamics* for this situation. In contrast, *selective language dynamics* refers to the situation where language affects (biological or cultural) fitness.

We consider a language to be a mapping between form and meaning. Denote by a_{ij} the probability that a speaker of L_i generates a sentence that is compatible with L_j. The payoff for L_i conversing with L_j is $F_{ij} = (a_{ij} + a_{ji})/2$. The matrix, $A = [a_{ij}]$, measures the similarity between languages $L_1, \ldots, L_n$ and is a property of UG.

Language acquisition is not perfect but subject to mistakes. The matrix Q contains entries q_{ij} denoting the probability that a child learner will acquire language L_j when learning from a teacher who uses L_i. The Q matrix will depend on the A matrix and the actual learning mechanism used by children.

Throughout the paper, we consider the fully symmetric case $a_{ij} = a$ for $i \neq j$ and $a_{ii} = 1$. Hence all languages have the same expressive power and are equally distant from each other. For this case, analytic investigations are possible. The symmetry of the A matrix is also reflected in the Q matrix. We have $q_{ij} = u/(n-1)$ for $i \neq j$ and $q_{ii} = 1 - u$. Here, u is the error rate of language acquisition. For most learning mechanisms, u will depend on a and n.

In Section 2, we summarize the results for infinite population size. For selective dynamics, there is a coherence threshold. If the error rate, u, is less than a critical value, u_1, then there are equilibrium solutions where a single grammar is predominating. If u is even smaller (less than a critical value $u_2 < u_1$), then such equilibrium solutions are the *only* stable solutions of the system. For selective dynamics, linguistic coherence is a requirement of language adaptation. Only a UG that induces linguistic coherence admits language adaptation. In turn, only a UG that admits language adaptation can be selected for linguistic function. For neutral dynamics, there is no linguistic coherence for infinite populations.

In Section 3, we introduce a stochastic process for language dynamics in finite populations. For selective dynamics, we find coherence threshold phenomena similar to the deterministic case, but language acquisition has to be slightly more accurate to maintain coherence in finite populations. For neutral dynamics, we find linguistic coherence whenever $u < u_3$, where $u_3 \sim 1/M$. In this case, it is likely that the stochastic process is at a state where all individuals speak the same language. This is coherence without adaptation. The model for neutral language dynamics is equivalent to stochastic processes analysed in the context of the neutral theory of molecular evolution (Moran, 1962; Kimura, 1983).

In Section 4, we estimate the lifetime, T, of a language in a finite population in the limit $u \ll 1/M$. For neutral dynamics, we find $T = 1/u$ generations. For selective dynamics, we find that T is proportional to $M^{-1/2}e^{\gamma M}$, where $\gamma = (1-a)/2/(1+a)$.

In Section 5, we summarize our findings and compare selective and neutral language dynamics.

There is a fascinating and growing literature on the evolution of language (Pinker, 1979; Lieberman, 1984; Brandon & Hornstein, 1986; Bickerton, 1990; Pinker & Bloom, 1990; Newmeyer, 1991; Hawkins & Gell-Mann, 1992; Aitchinson, 1996; Deacon, 1997; Jackendoff, 1999; Lightfoot, 1999). The modeling framework that we present in this paper is a part of a larger effort to obtain a mathematical description of language evolution (Batali, 1994; Hashimoto & Ikegami, 1996; Kirby & Hurford, 1997; Huford *et al.*, 1998; Nowak & Krakauer, 1999; Nowak *et al.*, 2000; Ferrer i Cancho & Sole, 2001; Kirby, 2001; Kamarova & Nowaka, 2001; Lachmann *et al.*, 2001; Christiansen *et al.*, 2002). The traditional approach is to discuss ideas of language in the context of language acquisition. A linguistic theory has "explanatory adequacy" if it describes how the grammar develops in the brain of a child learner (if it describes UG). The new approach ["Beyond explanatory adequacy" (Chomsky, 2002)] is to discuss theories of language acquisition in the context of evolution, which requires the synthesis of mathematical models of language, learning and evolutionary dynamics (Nowak, *et al.*, 2001, 2002; Komarova *et al.*, 2001).

2. The Limit of Infinite Populations

2.1. MODEL WITH SELECTION

Let us describe the dynamics of learning and evolution of languages. We denote by x_i the frequency of individuals who speak language i. Then we can set the fitness of those who speak language i to be

$$f_i = \sum_{k=1}^{n} x_k F_{ki}. \tag{1}$$

F_{ki} denotes the payoff for somebody who uses language k communicating with somebody who uses language i. We have $F_{ki} = (a_{ki} + a_{ki})/2$, where a_{ki} is the probability for a speaker of language k to produce a sentence compatible with language i. We assume that successful communication contributes to fitness. The average fitness of the population, ϕ, is given by

$$\phi = \sum_{k=1}^{n} f_k x_k. \tag{2}$$

The function f_i is the probability for a person who speaks language i to understand and be understood by the rest of the population. The function ϕ measures the *linguistic coherence* of the population and in general, takes values between zero and one (perfect coherence).

Individuals reproduce proportional to their fitness and the children learn the language of their parents, possibly with mistakes. In the limit of the infinite population size, the selective language dynamics can be described by the following system of equations (Nowak *et al.*, 2001):

$$\dot{x}_i = \sum_{j=1}^{n} f_j x_j Q_{ji} - \phi x_i, \quad 1 \leqslant i \leqslant n. \tag{3}$$

Here, Q_{ij} are elements of a row-stochastic matrix; the quantity Q_{ij} is the probability for the child to end up speaking language j, given that the parent speaks language i.

In what follows, we assume for simplicity that the similarity matrix has a symmetrical form:

$$a_{ij} = \begin{cases} a, & i \neq j, \\ 1, & i = j, \end{cases} \tag{4}$$

where $0 < a < 1$, and that the learning accuracy is given by

$$Q_{ij} = \begin{cases} u/(n-1), & i \neq j, \\ 1-u, & i = j, \end{cases} \tag{5}$$

where u is the probability to make a learning mistake; $0 \leqslant u \leqslant 1$.

The deterministic model with assumptions (4) and (5) has been studied in detail by Komarova *et al.* (2001). In particular, we know that the system undergoes a bifurcation as the probability of mistakes, u, decreases to zero. Namely, for large values of u the only stable fixed point corresponds to the "uniform solution" where each of the n languages occurs with the same frequency, $1/n$. As u decreases below a threshold given by

$$u_1^{\infty} = \frac{(n-1)[a(n-2)+n-2\sqrt{(n-1)[1+a(n-2)]}]}{(1-a)(n-2)^2} = \frac{1-\sqrt{a}}{1+\sqrt{a}} + O\left(\frac{1}{n\sqrt{a}}\right), \quad (6)$$

a new type of solutions appears which we call one-grammar solutions. There, one of the languages is dominant in that its share in the population is the largest; all the rest of the languages (the secondary languages) are spoken with equal frequencies. There are exactly n one-grammar solutions, one for each of the n languages. The linguistic coherence of the population corresponding to one-grammar solutions is higher than the coherence of the uniform solution, and it tends to the maximum coherence ($\phi = 1$) as $u \to 0$.

As long as u is not too small, the one-grammar solutions coexist with the uniform solution; we refer to this regime as bistability. When u crosses a second threshold, defined as

$$u_2^{\infty} = \frac{(1-a)(n-1)}{n[2+a(n-2)]} = \frac{1-a}{2+an} + O\left(\frac{1}{a^2n^2}\right), \quad (7)$$

the uniform solution loses stability, and for $u < u_2^{\infty}$, the only stable fixed points of the system are the one-grammar solutions.

The bifurcation diagram for eqn (3) is presented in Fig. 1(a). The values u_1^{∞} and u_2^{∞} are the two *coherence thresholds* of the language dynamics system with selection; the superscript refers to the fact that these values are defined for infinite populations. The value u_1^{∞} defines the maximum learning error compatible with coherence (i.e. the existence of one-grammar solutions). The value u_2^{∞} is the smallest learning error for which incoherent (uniform) solutions are still possible.

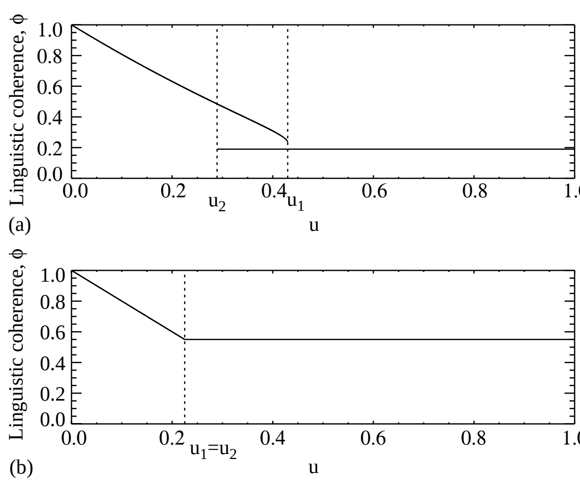

FIG. 1. Linguistic coherence, ϕ, of the deterministic system as a function of u, the probability of learning mistakes. Here $a = 0.1$, $n = 10$ in (a) and $a = 0.1$, $n = 2$ in (b).

For $n \geqslant 3$, the two coherence thresholds are distinct, and we have $u_1^{\infty} > u_2^{\infty}$. The case $n = 2$ is special because there,

$$u_1^{\infty} = u_2^{\infty} = \frac{1-a}{4}, \quad (8)$$

and there is no bistability regime, see diagram in Fig. 1(b).

2.2. THE NEUTRAL MODEL

Equation (3) together with eqn (1) describes a selection-mutation model; it implies that language-dependent fitness plays a role in the reproductive success of the individuals. The behavior of the system changes drastically if we do not include selection and consider neutral evolution of languages. The neutral limit of eqns (3) and (1) corresponds to setting

$$a = 1,$$

which is equivalent to taking

$$f_i = 1 \quad \forall i, \qquad \phi = 1;$$

hence we get

$$\dot{x}_i = \sum_{j=1}^{n} x_j Q_{ji} - x_i.$$

For our particular choice of the matrix Q, eqn (5), the only stable fixed point of the deterministic system (3) is the uniform solution, and linguistic coherence is never reached even for high learning accuracy.

3. Finite Populations

3.1. STOCHASTIC MODEL

Let us suppose that there are M people in a population, each speaking one of n languages. At each moment of time, one person produces a child, and one person dies, so that the population size stays constant (this is known as the Moran process in population genetics). The child learns the language of the parent, possibly with mistakes. As before, we set the probability for the child to learn language j given that the parent speaks language i to be Q_{ij}.

In what follows we will consider the following two birth–death processes.

- The *neutral model* assumes that all individuals reproduce with the same rate.
- In the *model with selection*, the chance for an individual to reproduce is weighted with its biological fitness.

In both models, we assume that the probability to die is the same for all.

Each model gives rise to a Markov process on an n-dimensional simplex, S_n. The stochastic variable is the vector $(I_1, I_2, \ldots, I_n)$, where $0 \leqslant I_k \leqslant M$ defines the number of people who speak language L_k; we have $\sum_{k=1}^{n} I_k = M$.

In the selection model, the fitness of individual k is given by its ability to communicate with the rest of the population,

$$f_k = \frac{1}{M-1} \sum_{i \neq k} F[L(k), L(i)], \tag{9}$$

where we sum over all other individuals in the population. Here $L(i)$ is the language of the i-th person and $F[L(k), L(i)]$ are entries of the payoff matrix A. The average fitness of the population, ϕ, is given by

$$\phi = \frac{1}{M} \sum_{k=1}^{M} f_k. \tag{10}$$

It is easy to see that in the limit $M \to \infty$, formulas (9) and (10) tend to eqns (1) and (2), respectively.

In the next subsections we will study the dynamics of the neutral model and the model with selection. We will see that the behavior of the model with selection is similar to that in the deterministic limit. Namely, if the accuracy of learning is high enough, the system finds itself above a coherence threshold. What is interesting is that linguistic coherence can also be found for the neutral model.

3.2. THE CASE OF TWO LANGUAGES

Let us assume that $n = 2$, i.e. that only two languages are possible. Let our stochastic variable, j, be the number of people who speak language 1. The variable j takes values between 0 and M. The following transition matrix governs the stochastic process:

$$P_{i \to i+1} = [f_1(i) i (1-u) + f_2(i)(M-i)u] \frac{M-i}{M^2 \phi(i)}, \tag{11}$$

$$P_{i \to i-1} = [f_1(i) i u + f_2(i)(M-i)(1-u)] \frac{i}{M^2 \phi(i)}, \tag{12}$$

$$P_{i \to i} = 1 - P_{i \to i-1} - P_{i \to i+1}, \quad 0 \leqslant i \leqslant M, \tag{13}$$

where the fitness functions, $f_{1,2}$, of speakers of language one and two, are given, respectively, by

$$f_1(i) = \frac{i - 1 + a(M-i)}{M-1},$$

$$f_2(i) = \frac{ai + M - i - 1}{M-1}$$

for the model with selection and they are taken to be $f_1 = f_2 = 1$ for the neutral process. The average fitness, ϕ, is defined as

$$\phi(i) = \frac{f_1(i) i + f_2(i)(M-i)}{M}.$$

3.2.1. *Selective Dynamics*

If $u = 0$, the states $j = 0$ and $j = M$ are absorbing, so that the stationary probability distribution is given by $(1/2, 0, \ldots, 0, 1/2)$. In the model with selection, stationary probability distribution is U-shaped for small positive values of u, see Fig. 2(a). As u increases, the two maxima of the distribution move towards the

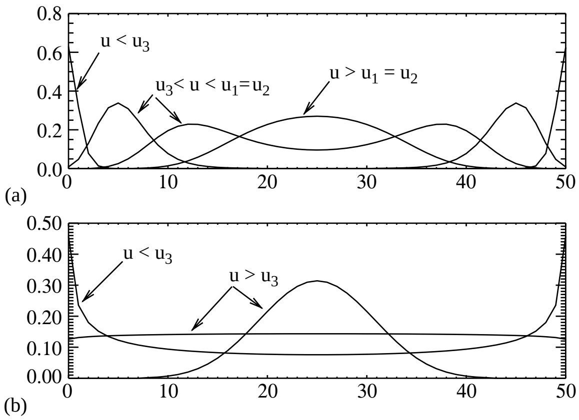

FIG. 2. Stationary probability distribution in the case $n = 2$, $M = 50$, for different values of u. (a) Model with selection, $u = 0.01$, 0.1, 0.2 and 0.3. We have $u_3 = 0.0196$ and $u_1 = u_2 = 0.259$. (b) Neutral model, $a = 0.1$, $u = 0.01$, 0.02 and 0.2. In this case, $u_3 = 0.0192$.

center and after a certain threshold value, the distribution becomes a one-humped function. The value of u for which a maximum appears in the center is the coherence threshold, u_2, for the stochastic selective model with $n = 2$. Incidentally, it coincides with the value of u for which the two distinct maxima corresponding to each of the two languages disappear, $u = u_1$; this is the degeneracy of the case $n = 2$, exactly as we had in the deterministic model. Now we present our first definition of the coherence threshold in the case $n = 2$.

Definition 1. We define the coherence threshold, $u_1 = u_2$, to be the value of u such that for $u < u_1$, the stationary probability distribution has a minimum at $i = M/2$, and for $u > u_1$ it has a maximum at $i = M/2$.

The coherence threshold can be calculated from the equation

$$P_{i \to i_1} + P_{i \to i+1} = P_{i-1 \to i} + P_{i+1 \to i},$$

where we set $i = M/2$, which gives

$$u_1 = u_2 = \frac{M^2 + 4M - 8 - a(M^2 - 4)}{4M(M + 2)}.$$

In the limit of large M, we obtain

$$u_1 = u_2 = \frac{1 - a}{n},$$

which coincides with the deterministic results for $n = 2$, eqn (8).

3.2.2. *Neutral Dynamics*

In the neutral model, stationary probability distribution is also U-shaped for small positive values of u, but behaves differently for larger u, see Fig. 2(b). As u increases, the distribution becomes more flat, and then "flips over" to become a one-humped function with a maximum at $j = M/2$. The value of u when this happens corresponds to the "neutral" coherence threshold, u_3, defined below.

Definition 2. We define the coherence threshold, u_3, to be such a value of u that for $u < u_3$, stationary probability distribution has two maxima at $j = 0$ and $j = M$, and for $u > u_3$ they become minima.

The coherence threshold, u_3, can be found from the equation

$$P_{0 \to 1} = P_{1 \to 0}, \tag{14}$$

which gives

$$u_3 = \frac{M - 1}{M^2 + M - 2}. \tag{15}$$

For large values of M, we have

$$u_3 = \frac{1}{M}. \tag{16}$$

In the model with selection, the coherence threshold, u_3, can also be found from eqn (14). It is given by

$$u_3 = \frac{M + a - 2}{M^2 + M(2a - 1) - 2}, \tag{17}$$

which gives $u_3 = 1/M$ for large population sizes. Expression (17) is a decreasing function of a, so that it is slightly easier to reach the coherence threshold u_3 for systems with strong selection (small a). For $a = 1$ (the neutral case) we recover eqn (15).

For any a, there is always a value M_0, such that for all $M > M_0$, $u_1 > u_3$. However, for a given M there can be found a value a_0 close to 1 such that for $a > a_0$, $u_3 > u_1$. More precisely, for

$$1 \geqslant a \geqslant 1 - 4/M^2,$$

we have $u_3 > u_1$. For the values of a in this interval, selection plays no role, and the system can be considered neutral.

The coherence threshold, u_3, can also be defined in the deterministic case. It is such value, u_3^∞, of u that for $u \leqslant u_3^\infty$, the solution where the frequency of one of the languages is equal to one, is stable. We have

$$u_3^\infty = 0.$$

3.3. THE CASE OF MANY LANGUAGES

The stationary probability distribution for $n > 2$ behaves in the following way. In the neutral system, it has n maxima at the corners of the simplex for small values of u, and a minimum in the middle, see the diagram in Fig. 3(a) corresponding to $n = 3$. As u grows through a certain value (the coherence threshold u_3), the stationary probability distribution changes concavity and acquires a maximum in the middle, whereas the maxima at the corners turn into minima.

For the model with selection we have a more complex picture, Fig. 3(b). As before, for small u there are n maxima in the corners and a minimum in the middle of the simplex . As u grows, the n maxima move towards the center of the simplex and become less sharp. At some point ($u = u_2$), an additional maximum appears in the center of the simplex. This corresponds to the bistability regime. As u continues to grow, the middle maximum becomes higher and the n symmetrical maxima get smaller, until they disappear (at $u = u_1$) and we are left with only one maximum at the center.

3.3.1. *Coherence Threshold u_3 for $n > 2$*

We have simulated the neutral process for different values of u and M. For each value of u, we started with the initial condition where all individuals spoke different languages, with $n = 1000$. The stochastic process was running for $10\,000M$ time steps (which is 10 000 generations), and then the linguistic coherence was calculated by averaging over the next 10 000 generations. Results for $M = 10$ and 100 are presented in Fig. 4. We observe that coherence is higher for smaller values of M. For small values of u, the average coherence is high and tends to its maximum ($\phi = 1$) as u decreases to zero.

The definition of coherence threshold u_3 given in the previous section can be generalized in the case of n languages:

Definition 3. Coherence threshold u_3 is the value of u, such that for $u < u_3$, the stationary probability distribution has maxima in the corners of the simplex, and for $u > u_3$ it has minima in the corners (and a maximum in the middle).

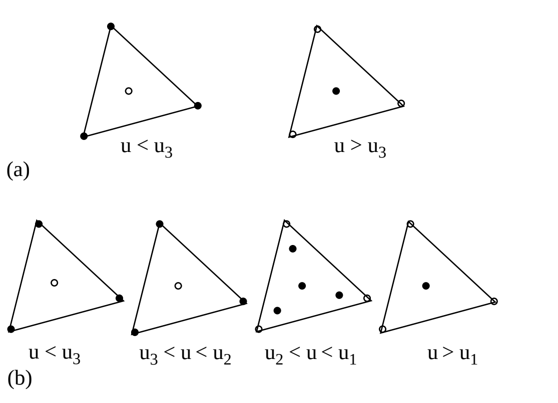

FIG. 3. Stationary probability distribution $n = 3$, for different values of u. (a) Neutral dynamics. (b) Selective dynamics. Solid circles indicate maxima and empty circles minima.

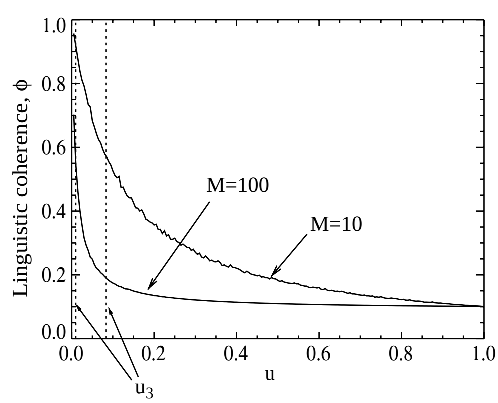

FIG. 4. Average linguistic coherence, ϕ, of the stochastic system as a function of u, the probability of learning mistakes. Here $a = 0.1$, $n = 10$ in (a) and $a = 0.1$, $n = 2$ in (b).

The value u_3 can be found in the following way. Let us consider the states $\mathbf{x}_0 = (M, 0, \ldots, 0)$ (a corner of the simplex) and for $2 \leqslant k \leqslant n$, $\mathbf{x}_1^k = (M-1, 0, \ldots, 0, 1, 0, \ldots, 0)$, where the unit entry corresponds to the k-th slot (the neighboring states of $\mathbf{x}_0$). Further we will denote as $\pi(\mathbf{x})$ the probability to attain the state $\mathbf{x}$. We have

$$\pi(\mathbf{x}_0) \sum_{k=2}^{n} P_{\mathbf{x}_0 \to \mathbf{x}_1^k} = \sum_{k=2}^{n} \pi(\mathbf{x}_1^k) P_{\mathbf{x}_1^k \to \mathbf{x}_0}.$$

Let us use the fact that $P_{\mathbf{x}_0 \to \mathbf{x}_1^k} = P_{\mathbf{x}_0 \to \mathbf{x}_1^m}$ and $P_{\mathbf{x}_1^k \to \mathbf{x}_0} = P_{\mathbf{x}_1^m \to \mathbf{x}_0}$ for all k and m, which is a consequence of the symmetries of the system. Setting $\pi(\mathbf{x}_0) = \pi(\mathbf{x}_1^k)$ for all k, we have

$$P_{\mathbf{x}_0 \to \mathbf{x}_1^k} = P_{\mathbf{x}_1^k \to \mathbf{x}_0} \quad \forall k.$$

This gives

$$u_3 = \frac{(n-1)(M+a-2)}{(n-1)[M^2 + M(2a-1) + a - 2] - a}. \quad (18)$$

For $n \gg 1$ we have

$$u_3 = \frac{M+a-2}{M^2 + M(2a-1) + a - 2}.$$

Again, for large values of M we have

$$u_3 = \frac{1}{M}.$$

In Fig. 4, estimate (18) where we take $a = 1$ is shown with dashed vertical lines.

3.3.2. *Coherence Thresholds u_1 and u_2 for $n > 2$, $M \gg n$*

The prediction of the deterministic neutral model cannot be used to describe the behavior of the stochastic neutral model. On the other hand, the stochastic model with selection behaves in the way predicted by the deterministic system, see Fig. 5. As M grows, the average coherence measured in the stochastic simulation approaches the one calculated in the limit $M \to \infty$.

Let us assume that $M \gg n$ and define the coherence thresholds for the system with n languages.

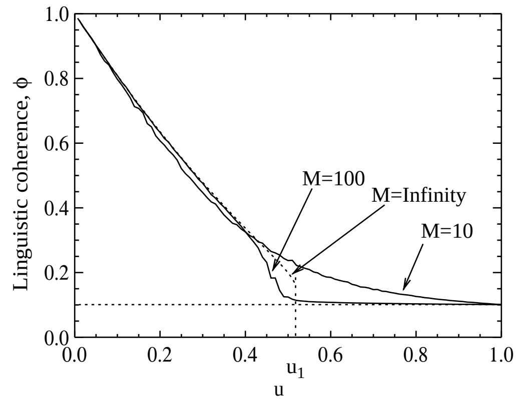

FIG. 5. Average linguistic coherence as a function of u for the stochastic system with $a = 0.1$ and $n = 1000$. Numerical results are presented for $M = 10$ and $M = 100$. The dotted lines represent the analytic solution for infinite M, both for the one-grammar and the uniform solutions (the latter one corresponds to the horizontal line). The vertical dashed line denotes the coherence threshold u_1^{∞} for the analytic solution in the case of infinite populations.

Definition 4. Coherence threshold, u_1, is the value of u such that for $u > u_1$, the stationary probability distribution has a unique maximum at the state $\bar{\mathbf{x}} = (M/n, \ldots, M/n)$, and for $u < u_1$ it has $n+1$ maxima. One of the maxima is at $\bar{\mathbf{x}}$ and n are at points equidistant from $\bar{\mathbf{x}}$ and situated on the straight lines that connect $\bar{\mathbf{x}}$ to the corners of the simplex.

In the next section we will use a simplified model to show that as M increases, the stochastic coherence threshold approaches the value u_1 calculated for the deterministic system and given by eqn (6). Here we will look at the other coherence threshold, u_2.

Definition 5. Coherence threshold, u_2, is the value of u such that for $u > u_2$, the stationary probability distribution has a maximum at the state $\bar{\mathbf{x}} = (M/n, \ldots, M/n)$, and for $u < u_2$ it has a minimum there.

As in the case $n = 2$, this can be found by equating the probability to leave the state $\bar{\mathbf{x}}$ with the probability to enter this state from all neighboring states. Each of the neighboring states has $n-2$ entries M/n, one entry $M/n + 1$ and one entry $M/n - 1$. We skip the details of the calculation, which is straightforward, and give the result:

$$u_2 = \frac{(n-1)[(1-a)M(M+n+2)+(a(M+1)-1)n^2]}{n(M+n)[(M+1)(2+a(n-2))-n]}.$$

Expanding in $1/M$, we have

$$u_2 = u_2^\infty + \frac{(n-1)[2+n+2a(n^2-2)+a^2(n-2)(n^2+1)]}{n(2+a(n-2))^2}\frac{1}{M} + O\left(\frac{1}{M^2}\right),$$

where u_2^∞ is the deterministic value of this threshold defined by formula (7). In the case $1 \ll n \ll M$ we get a simpler expression:

$$u_2 = u_2^\infty + \frac{n}{M} + O\left(\frac{1}{M^2}\right).$$

3.4. A SIMPLIFIED PROCESS

In order to estimate the threshold u_1, we will consider a simplified stochastic process. Let us assume that $n \gg 1$. The following simplified one-dimensional process can mimic the population dynamics of language learning. Let us consider such states where j people speak a dominant language, and the rest $M-j$ people all have different languages. Let j be our stochastic variable. It takes values from 1 to M. The fitness of people who speak the dominant language is given by

$$f_d(j) = [j-1+(M-j)a]/(M-1),$$

and the fitness of each of the rest of the individuals is

$$f_s = a.$$

Here the subscripts in f_d and f_s refer to "dominant" and "secondary" grammars, respectively. The average fitness is given by

$$\phi(j) = [f_d(j)j + f_s(M-j)]/M.$$

We have

$$P_{j\to j+1} = \frac{f_d(j)j(1-u)(M-j)}{M^2\phi(j)}, \tag{19}$$

$$P_{j\to j-1} = \frac{[f_d(j)ju+(M-j)f_s]j}{M^2\phi(j)}, \tag{20}$$

$$P_{j\to j} = 1 - P_{j\to j+1} - P_{j\to j-1}, \tag{21}$$

for $2 \leqslant j \leqslant M-1$, and

$$P_{1\to 2} = (1-u)(1-1/M), \quad P_{1\to 1} = 1 - P_{1\to 2}, \tag{22}$$

$$P_{M\to M-1} = u, \quad P_{M\to M} = 1-u. \tag{23}$$

Equation (22) means that once we are at a "uniform" state (everybody speaks a different language), there are M different ways to go to the state where two people speak the dominant grammar, because any of the M grammars spoken can become dominant. In eqns (19–23) we neglected back-mutations. This is a consequence of the fact that n is very large.

The model described by eqns (19–23) approximates the full process best when selection is strong, which corresponds to small values of a. Then as M gets large, the dynamics of the simplified process approach the dynamics of the actual, multi-dimensional stochastic process, see Fig. 6. This is because in systems with strong selection, there is a strong tendency to cluster. In Fig. 6 the parameter a is taken to be $a = 0.1$; we have checked that a very good agreement is achieved for values of a as high as $a = 0.5$. For weak selection and neutral systems (a near 1) the simplified process always predicts lower coherence (and higher coherence thresholds) than the full stochastic process.

The main difference between the full and the simplified processes is clear from eqn (20). There are two ways in which the number of people speaking the dominant grammar, j, can decrease by one. One way is to have $j-1$ people speaking the dominant grammar and $M-j+1$ people

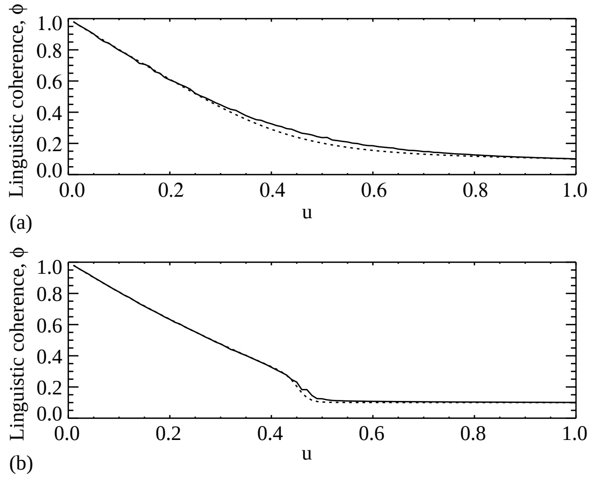

FIG. 6. Average linguistic coherence, ϕ, as a function of u, for selection models. ϕ for the full stochastic process (solid lines) is calculated numerically by averaging over a large number of generations. ϕ for the simplified process (dashed lines) is calculated from the stationary probability distribution as $\sum_{j=1}^{M} p_j\phi(j)$. (a) $M = 10$, (b) $M = 100$. Other parameters are $a = 0.1$ and $n = 1000$ in both plots.

speaking all different languages, and the other is to have *two* people speaking one of the secondary languages. The latter state is not a part of the simplified description, and in eqn (20) we simply include transitions of this kind into $P_{j\to j-1}$ [the second term on the right-hand side of eqn (20)]. The corresponding term is multiplied by a, so for small values of a the error introduced by it is small.

The stationary probability distribution, $(p_1, \dots, p_M)$, of process (19–23) is the left eigenvector of the matrix P corresponding to the top eigenvalue, and is found from equation

$$\pi_i(P_{i\to i+1} + P_{i\to i-1}) = \pi_{i-1}P_{i-1\to i} + \pi_{i+1}P_{i+1\to i},$$
$$2 \leqslant i \leqslant M-1 \qquad (24)$$

with the boundary conditions

$$\pi_1 P_{1\to 2} = \pi_2 P_{2\to 1}, \qquad (25)$$

$$\pi_M P_{M\to M-1} = \pi_{M-1}P_{M-1\to M}. \qquad (26)$$

For $u = 0$, we have $p_M = 1$, i.e. everyone has the same language. For small non-trivial values of u, the stationary probability distribution peaks at $j = M$, which means that the system spends most of the time at a state where all people have the same language. As u decreases, the maximum of p_j moves from $j = M$ to smaller values of j, which means that most of the time only a certain fraction of people speak the dominant language. As u grows, a second maximum near $j = 1$ becomes more visible. For a window of values of u, the two maxima coexist, and for yet larger values of u the only remaining peak corresponds to the uniform solution.

Let us estimate the coherence threshold u_1. The number of people typically speaking the dominant grammar can be found from eqns (24–26) for π_j. Let us denote $\kappa_i = \pi_i/\pi_{i+1}$. Then we have the following solution for $2 \leqslant i \leqslant M-1$:

$$\kappa_i = \frac{[(i+1)u + (M-i-1)a(M-1+(i+1)u)/i](i+1)[i(i-1)(1-a) + M(M-1)a]}{[(M-i)a + i - 1](1-u)(M-i)[(i+1)i(1-a) + M(M-1)a]}.$$

In particular, for $a = 0$ this is reduced to $\kappa_i = (i+1)u/(M-i)/(1-u)$. The equation

$$\kappa_i = 1 \qquad (27)$$

gives the value $i = i_0$ which corresponds to the maximum of the stationary probability distribution. For $a = 0$ we have $i_0 = M(1-u)$. For general values of a, we can solve eqn (27) in the limit of large values of M. If we denote $x = i/M$, expand in $1/M$ and only consider the leading order term, we will get

$$i_0 = \frac{1}{2}\left(1 - u + \sqrt{(1-u)^2 - \frac{4au}{1-a}}\right).$$

This maximum exists if $u < u_1$ with

$$u_1 = \frac{1-\sqrt{a}}{1+\sqrt{a}},$$

which is exactly the result obtained for infinitely large populations, see eqn (6). If the next term is taken into account in the expansion of u_3 in terms of $1/M$, we obtain

$$u_1 = u_1^{\infty}(1 - 3/M).$$

For completeness, we note that the coherence threshold u_3 can also be found for the simplified process; we have

$$u_3 = \frac{(M-1)(M+a-1)}{M^3 - M^2 + 1 + a(2M^2 - M - 1)}.$$

For large values of M, we again obtain

$$u_3 = 1/M.$$

4. Characteristic Time of Language Change

Let us estimate the time-scale on which a language can disappear from a population. Let us assume that u is small (more precisely, that $u \ll 1/M$). This means that new languages are produced very rarely, and once a new language is invented (a "mutant" speaking a different language is produced), then this new language typically has time to invade or go extinct until a new innovation occurs (another "mutant" is produced). Under this assumption we can calculate the expected time, T, it takes to go from the state where everybody in the population speaks a certain language until this language is extinct. The trick is to only consider two languages: the one spoken by everybody as the initial condition (language 1), and the second language "invented" (language 2).

In the case where $u \ll 1/M$, the rate of language change, $1/T$, equals the rate at which speakers of language 2 are produced, times the probability (p_1) for language 2 to reach fixation in the population. The latter quantity can be found under the assumption that $u = 0$.

Therefore, for our purposes, it is enough to consider the case $n = 2$ and calculate the expected time to go from the state $j = 0$ to the state $j = M$. For $u = 0$, the states $j = 0$ and $j = M$ are absorbing. Let us denote as p_j the probability that starting from the state j, the system gets absorbed in the state $j = M$. We have

$$p_j = P_{j\to M} + \sum_{k=1}^{M-1} P_{j\to k} p_k, \quad 1 \leqslant j \leqslant M-1,$$

where we set $u = 0$ in the expression for the matrix P. Then the time to change from one grammar to the other for small values of u is given by

$$T^{-1} = up_1.$$

4.1. NEUTRAL LANGUAGE DYNAMICS

In the neutral case ($a = 1$) we have the following system for p_j:

$$2p_j = p_{j+1} + p_{j-1}, \quad 2 \leqslant j \leqslant M-2, \qquad (28)$$

$$2p_1 = p_2, \quad 2p_{M-1} = 1 + p_{M-2}. \qquad (29)$$

The solution is $p_j = j/M$ which gives the time of language change in the neutral model (under the assumption that $u \ll 1/M$),

$$T = M/u.$$

It follows that it takes on average

$$T_{gen} = 1/u$$

generations before a language is lost from the population. Note that this quantity is M-independent in this limit.

4.2. SELECTIVE LANGUAGE DYNAMICS

In the case with fitness, let us denote $\xi_j = p_i - p_{i-1}$. We have

$$[j - 1 + a(M-j)]\xi_{j+1} = [aj + M - j - 1]\xi_j. \qquad (30)$$

Let us ignore 1 in comparison with M and change coordinates such that $k = j - M/2$. We can write

$$\frac{\xi_k}{\xi_{k+1}} = 1 - \frac{\xi_{k+1} - \xi_k}{\xi_{k+1}},$$

and from eqn (30) we have

$$\frac{\xi_{k+1} - \xi_k}{\xi_{k+1}} = \frac{-2(1-a)2k/M}{1 + a - 2k/M(1-a)}.$$

We can see that for $k \ll M$, the rate of change of the function ξ_k is slow, and we can use a continuous description to obtain

$$\frac{d\xi_k}{dk}\frac{1}{\xi_k} = -\frac{4(1-a)k}{(1+a)M}.$$

This gives the estimate

$$\xi_j = \exp\left[-\frac{2(1-a)(j-M/2)^2}{(1+a)M}\right].$$

Integrating in i, normalizing and setting $i = 1$, we obtain

$$p_1 = \frac{1}{2}\left(1 - \frac{Erf\left[\sqrt{\frac{2(1-a)}{1+a}}\left(\frac{\sqrt{M}}{2} - \frac{1}{\sqrt{M}}\right)\right]}{Erf\sqrt{\frac{M(1-a)}{2(1+a)}}}\right),$$

where $Erf(z) = 2/\sqrt{\pi}\int_0^z e^{-t^2}\,dt$. For large values of M, and assuming that $M(1-a) \to \infty$, this gives

$$p_1 = \sqrt{\frac{2(1-a)}{M\pi(1+a)}}\exp\left[-\frac{M(1-a)}{2(1+a)}\right].$$

Therefore, we obtain the time of language change in the model with fitness in the case where $u \ll 1/M$,

$$T = \frac{1}{u}\sqrt{\frac{M\pi(1+a)}{2(1-a)}}\exp\left[\frac{M(1-a)}{2(1+a)}\right]. \quad (31)$$

If we denote $\gamma = (1-a)/2/(1+a)$, then we can say that it takes on average

$$T_{gen} \propto M^{-1/2}e^{\gamma M}$$

generations before a language is lost from a population if there is a selection for linguistic ability. Note that in order to include the limit of (almost) no fitness (a close to 1), we need to use the expression

$$T = \frac{1}{u}\sqrt{\frac{M\pi(1+a)}{2(1-a)}} \exp\left[\frac{M(1-a)}{2(1+a)}\right] Erf\sqrt{\frac{M(1-a)}{2(1+a)}}.$$

This gives the right limit ($T = M/u$) as $a \to 1$.

We have compared estimate (31) with numerical experiments, see Fig. 7. For $a = 0.5$, $u = 10^{-4}$ and the values $n = 2$, $n = 100$ and $n = 1000$, and M between $M = 5$ and 30, we estimated the average time of language extinction. For each set of parameters, we performed 100 runs. Starting from the initial condition where everyone in the population spoke the same language, we measured the time it took for this language to disappear from the population. The numerical results are in an excellent agreement with formula (31).

5. Summary

In this paper, we have studied stochastic processes of language acquisition and evolution in finite populations. We have shown how coherence threshold phenomena that were originally calculated for deterministic dynamics of infinite populations carry over to stochastic dynamics of finite populations. We derive analytic expressions for coherence thresholds in finite population and find that they all approach the values derived for deterministic dynamics in the limit of large population size, M.

Of special interest is that, for neutral language dynamics, we find linguistic coherence for $u < 1/M$, where u is the error rate of language acquisition. This coherence threshold corresponds to an equilibrium distribution of the stochastic process which is peaked at homogeneous states (all individuals use the same language). This threshold is largely independent of the number of candidate languages, n, see eqn (18). Notice however that for most learning

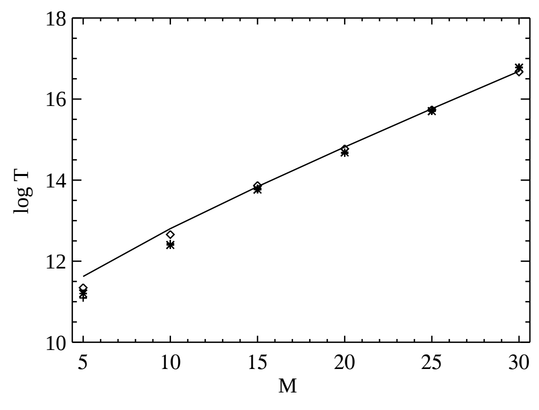

FIG. 7. Average time it takes to lose a language as a function of M, in the selection model. The solid line corresponds to the analytical result (31), diamonds represent the numerical results for $n = 2$, stars—numerical results for $n = 100$ and crosses—numerical results for $n = 1000$. Other parameters are $u = 10^{-4}$ and $a = 0.5$.

algorithm (such as memoryless learners or batch learners), u will depend on n.

Neutral coherence is an important finding because it explains how linguistic features that do not contribute to communicative fitness (efficacy) can be fairly homogeneous in a population. Neutral language dynamics provide an appropriate description for many language changes studied in historical linguistics (Kroch, 1989; Hopper & Traugott, 1993; Niyogi & Berwick, 1997; Wang, 1998; de Graff, 1999) where fitness effects can probably be neglected.

Selective language dynamics facilitates coherence at higher error rates, u. Selection needs to be taken into account whenever we want to derive a global understanding of language evolution.

REFERENCES

AITCHINSON, J. (1996). *The Seeds of Speech*. Cambridge, MA: Cambridge University Press.

BAKER, M. C. (2001). *Atoms of Language*. New York: Basic Books.

BATALI, J. (1994). Innate biases and critical periods. In: *Artificial Life IV* (Brooks, R. & Maes, P., eds), pp. 160–171. Cambridge, MA: MIT Press.

BICKERTON, D. (1990). *Language and Species*. Chicago: University of Chicago Press.

BRANDON, R. & HORNSTEIN, N. (1986). From icon to symbol: some speculations on the evolution of natural language. *Philosophy and Biology* **1,** 169–189.

CHOMSKY, N. (1972). *Language and Mind*. New York: Harcourt Brace Jovanovich.

CHOMSKY, N. (1981). Principles and parameters in syntactic theory. In: *Explanation in Linguistics* (Hornstein, N. & Lightfoot, D., eds), pp, 123–146. London: Longman.

CHOMSKY, N. A. (1984). *Lectures on Government and Binding: The Pisa Lectures*. Holland: Foris Publications, Dordrecht.

CHOMSKY, N. (2002). Preprint.

CHRISTIANSEN, M. H., DALE, R. A. C., ELLEFSON, M. R. & CONWAY, C. M. (2002). The role of sequential learning in language evolution: computational and experimental studies. In: *Simulating the Evolution of Language* (Cangelosi, A. & Parisi, D., eds), pp. 165–187. London: Springer.

DEACON, T. (1997). *The Symbolic Species*. London: Penguin Books.

FERRER I CANCHO, R. & SOLE, R. V. (2001). The small world of human language. *Proc. R. Soc. London B* **268,** 2261–2265.

de GRAFF, M. (1999). *Language Creation and Language Change: Creolization, Diachrony and Development*. Cambridge, MA: MIT Press.

HASHIMOTO, T. & IKEGAMI, T. (1996). Emergence of net-grammar in communicating agents. *BioSystems* **38,** 1–14.

HAWKINS, J. A. & GELL-MANN, M. (1992). *The Evolution of Human Languages*. Reading, MA: Addison-Wesley.

HOPPER, P. & TRAUGOTT, E. (1993). *Grammaticalization*. Cambridge: Cambridge University Press.

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

JACKENDOFF, R. (1999). Possible stages in the evolution of the language capacity. *Trends Cognit. Sci.* **3,** 272–279.

KIMURA, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

KIRBY, S. & HURFORD, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In: *ECAL97* (Husbands P. & Harvey, I., eds), pp. 493–502. Cambridge, MA: MIT Press.

KIRBY, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Trans. Evol. Comput.* **5,** 102–110.

KOMAROVA, N. L., NIYOGI, P. & NOWAK, M. A. (2001). Evolutionary dynamics of grammar acquisition. *J. theor. Biol.* **209,** 43–59.

KOMAROVA, N. L. & NOWAK, M. A. (2001). Evolutionary dynamics of the lexical matrix. *Bull. Math. Biol.* **63,** 451–485.

KROCH, A. (1989). Reflexes of grammar in patterns of language change. *Lang. Variation Change* **1,** 199–244.

LACHMANN, M., SZAMADO, S. & BERGSTROM, C. T. (2001). Cost and conflict in animal signals and human language. *Proc. Natl. Acad. Soc. U.S.A.* **98,** 13189–13194.

LIEBERMAN, P. (1984). *The Biology and Evolution of Language*. Cambridge, MA: Harvard University Press.

LIGHTFOOT, D. (1999). *The Development of Language: Acquisition, Changes and Evolution*. Maryland Lecture in Language and Cognition. Oxfordz: Blackwell.

MORAN, P. A. P. (1962). *The Statistical Processes of Evolutionary Theory*. Oxford: Clarendon Press.

NEWMEYER, F. (1991). Functional explanation in linguistics and the origins of Language. *Lang. Commun.* **11,** 3–28, 97–108.

NIYOGI, P. & BERWICK, R. C. (1997). Evolutionary consequences of language learning. *J. Linguistics Philos.* **20,** 697–719.

NOWAK, M. A. & KRAKAUER, D. C. (1999). The evolution of language. *Proc. Natl. Acad. Sci. U.S.A.* **96,** 8028–8033.

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

NOWAK, M. A., KOMAROVA, N. L. & NIYOGI, P. (2001). Evolution of universal grammar. *Science* **291,** 114–118.

NOWAK, M. A., KOMAROVA, N. L. & NIYOGI, P. (2002). computational and evolutionary aspects of language. *Nature*, **417,** 611–617.

PINKER, S. (1979). Formal models of language learning. *Cognition* **7,** 217–283.

PINKER, S. & BLOOM, A. (1990). Natural language and natural selection. *Behav. Brain Sci.* **13,** 707–784.

PRINCE, A. & SMOLENSKY, P. (1997). Optimality: from neural networks to universal grammar. *Science* **275,** 1604–1610.

WANG, W. S. Y. (1998). Language and the evolution of modern humans. In: *The Origins and Past of Modern Humans* (Omoto, K. & Tobias, P. V., eds), pp. 247–262. Singapore: World Scientific.