J. theor. Biol. (2001) **209,** 43–59
doi:10.1006/jtbi.2000.2240, available online at http://www.idealibrary.com on IDEAL®

The Evolutionary Dynamics of Grammar Acquisition

NATALIA L. KOMAROVA*†, PARTHA NIYOGI‡ AND MARTIN A. NOWAK*

* *Institute for Advanced Study, Einstein Drive, Princeton, NJ* 08540, *U.S.A.*,
† *Department of Applied Mathematics, University of Leeds, Leeds LS*2 9*JT, U.K. and*
‡ *Department of Computer Science, University of Chicago, Chicago, IL* 60637, *U.S.A.*

(*Received on* 15 *June* 2000, *Accepted in revised form on* 24 *July* 2000)

Grammar is the computational system of language. It is a set of rules that specifies how to construct sentences out of words. Grammar is the basis of the unlimited expressibility of human language. Children acquire the grammar of their native language without formal education simply by hearing a number of sample sentences. Children could not solve this learning task if they did not have some pre-formed expectations. In other words, children have to evaluate the sample sentences and choose one grammar out of a limited set of candidate grammars. The restricted search space and the mechanism which allows to evaluate the sample sentences is called universal grammar. Universal grammar cannot be learned; it must be in place when the learning process starts. In this paper, we design a mathematical theory that places the problem of language acquisition into an evolutionary context. We formulate equations for the population dynamics of communication and grammar learning. We ask how accurate children have to learn the grammar of their parents' language for a population of individuals to evolve and maintain a coherent grammatical system. It turns out that there is a maximum error tolerance for which a predominant grammar is stable. We calculate the maximum size of the search space that is compatible with coherent communication in a population. Thus, we specify the conditions for the evolution of universal grammar.

1. Introduction

Many sentences a person utters during life are new combinations of words appearing for the first time in the history of the universe. Hence, language is not simply a repertoire of memorized responses, but instead the brain has a powerful recipe book to construct sentences out of words (Chomsky, 1959; Pinker & Prince, 1988). This recipe book is called *mental grammar*. Children acquire the mental grammar of their native language rapidly and without formal education. When it comes to applying grammatical rules then 3-year old children are more than 90% on target.

Noam Chomsky (1965, 1980, 1993) points out that the mental grammar of a person (or "the computational system" of the language)' is a rich and complex structure which is hopelessly underdetermined by the fragmentary evidence available to the child. In other words, the sample sentences available to the child are not nearly enough to recreate all of the underlying grammatical rules. This is what linguists call the "poverty of input" and the "paradox of language acquisition" (Jackendoff, 1997; Hornstein & Lightfoot, 1981). Nevertheless, children growing up in the same speech community correctly deduce the underlying grammatical rules and consistently develop the same language. For many linguists this observation provides conclusive evidence that children must employ highly restrictive principles that guide the development of their grammar. This restriction is called *universal grammar*.

The acquisition of mental grammar is therefore not seen as a process where children simply try to memorize all syntactic structures they encounter. Instead it is conjectured that children have a set of candidate grammars available to them. Subsequently, they evaluate the environmental input (i.e. the sentences they hear from their parents and others) in order to determine which of the candidate grammars is being used. Universal grammar provides the restricted search space of all candidate grammars and perhaps the learning mechanism which allows to evaluate the sample sentences (Sorace *et al*., 1999; Prince & Smolensky, 1993; Wexler & Culicover, 1980; Lightfoot, 1991; Niyogi & Berwick, 1996, 1997). Universal grammar is therefore not learned but must be available to the child when the learning process starts. In other words, universal grammar is innate.

It is believed that universal grammar is the product of some special neuronal circuitry within the human brain, which is called "language organ" by Chomsky and "language instinct" by Pinker (1994, 1999). All humans, but no animals have a language instinct (Bickerton, 1990; Deacon, 1997; Hauser, 1996; Brandon & Hornstein, 1986; Pinker & Bloom, 1990).

The purpose of this paper is to develop a mathematical theory for the evolutionary and population dynamics of grammar acquisition (Nowak *et al*., 2001). In accordance with mainstream linguistic theory, we assume that children have a search space that consists of n candidate grammars, $G_1, \ldots, G_n$. Then they hear sample sentences and decide which of the possible grammars to use. Note that the number of candidate grammars can also be infinite, provided that children have a prior probability distribution specifying that some grammars are more likely than others. In this paper, however, we will restrict our analyses to the case of a finite search space, where all candidate grammars are equally likely at the beginning of the learning process. The extension to infinite, but biased, search spaces will be treated elsewhere.

One way to visualize this learning scenario is by imagining that the mental grammar is determined by *principles and parameters*. The principles are hardwired and innate. The principles restrict the infinitely large set of all conceivable grammars to a finite set of relevant grammars, $\{G_1, \ldots, G_n\}$. The parameters, on the other hand, need to be learned. A particular choice of parameters corresponds to one specific grammar, G_i. If, for example, the parameters are k independent Boolean variables (binary switches), then $n = 2^k$. Learning k independent parameters means identifying the right grammar among 2^k different alternatives. In principle, the child needs to hear sufficiently many sampling sentences to (uniquely) determine the setting of these k parameters (Gibson & Wexler, 1994).

A different approach is optimality theory. There are k constraints. A grammar is given by a particular ordering of these constraints. Hence, in total there are $n = k!$ candidate grammars, but many of them might be equivalent. The difference between constraints and parameters is as follows: parameters have to hold for all constructions of a given grammar, while constraints have to hold unless they are overruled by higher ranked constraints. Thus, for natural languages it is possible to specify simple constraints, while parameters are often complicated (Prince & Smolensky, 1997; Tesar & Smolensky, 2000).

Furthermore, we assume that communication among individuals has an effect on fitness. Someone who uses a grammar that is understood by others has a better performance during life history in terms of survival probability or reproductive success. Individuals who communicate successfully leave more offspring, who in turn learn their language, which puts the problem of grammar acquisition in an evolutionary context (Hashimoto & Ikegami, 1995, 1996; Nowak & Krakauer, 1999; Nowak *et al*., 1999, 2000).

Learning theory (Vapnik, 1995; Valiant, 1984; Niyogi, 1998; Haussler *et al*., 1997; Osherson *et al*., 1986) often asks the question how many sample sentences are needed for an individual learner to acquire the correct rule from a single teacher with a certain probability. In contrast, we study the following question: what are the conditions for the learning process which allow a *population* to evolve or maintain a unique grammar? To this end, we ask what is the maximum size of the search space for a specific learning algorithm, given the number of sample sentences. More generally, we calculate the minimum learning accuracy that is compatible with the whole population converging to a predominant grammar.

We will explore two different learning algorithms that represent opposite extremes of how much memory capacity is required during language acquisition. The *memoryless learner* holds at any one time a specific hypothesis (one of the n grammars). It switches at random to a new hypothesis should a sentence occur that is not consistent with the current hypothesis. The *batch learner* memorizes all sample sentences and decides at the end of the learning period which of the n grammars fits best. We assume that the learning mechanism employed by humans lies somewhere between these two extremes.

We will show that for memoryless learners, each child needs a number of sample sentences, b, which is greater than a constant times n, the number of possible grammars. For the batch learner we find that b must exceed a constant times $\log n$. If these conditions are fulfilled, then there will be a predominating grammar that is used by most individuals in the population. Clearly, this is a requirement for the evolution of a coherent language. Hence, the conditions $b > C_1 n$ or $b > C_2 \log n$, where $C_{1,2}$ denote some constants, specify how restrictive the search space has to be, that is how small the number of candidate grammars, n has to be compared to the number of sample sentences, b, for universal grammar to work (to evolve). Conversely, the capacity of the learning mechanism (which can also be seen as being a part of universal grammar) is specified by b and whichever algorithm is being used for evaluating the sample sentences.

Section 2 outlines the general model for the population dynamics of grammar acquisition. Section 3 provides a detailed bifurcation analysis of a special case where the relationship among the n candidate grammars is completely symmetric. Section 4 describes the memoryless learning algorithm together with a number of specific examples. Section 5 describes the batch learner. The conclusions are presented in the final section. Some examples of search spaces are worked out in Appendices A and B.

2. General Model

Mathematically speaking, a grammar mediates a mapping between form and meaning. The countably infinite number of possible linguistic expressions can be represented as strings over some finite alphabet, Σ_1. The set of all possible strings of Σ_1 is denoted by Σ_1^*. In our context, we can think of the alphabet as all words. Then a string would be a sequence of words, i.e. a sentence. A grammar specifies which sentences are valid and which are not. Sometimes it is convenient to consider a "compressed" alphabet consisting, for example, of nouns and verbs. In this case, a string is a sequence of nouns and verbs and can be seen as a *sentence type*. The exact interpretation of what is meant by the basic alphabet is not important for our analysis, and we will use the terms "sentence" and "sentence type" interchangingly.

On the other hand, grammar does not only specify whether a sentence is valid or not, but also conveys meaning. Hence, more generally, we have to see a grammar as a mapping between *syntactic forms* and *semantic forms*. Let us enumerate all possible meanings as strings over a primitive semantic alphabet Σ_2. Therefore, Σ_1^* is the set of all possible linguistic expressions and Σ_2^* is the set of all possible meanings. A grammar G_i generates a subset of $\Sigma_1^* \times \Sigma_2^*$, that is a (potentially infinite) set of sentence-meaning pairs. Each grammar, G_i, represents a measure μ_i on $\Sigma_1^* \times \Sigma_2^*$. Such measure specifies how often each grammar may use a certain syntactic construct to express a given meaning.

Let us assume that there are n possible grammars, $G_1, \ldots, G_n$. Matrix A relates grammars to each other. Let us define $a_{ij} = \mu_i(G_i \cap G_j)$ to be simply the proportion of sentence-meaning pairs that G_i and G_j have in common. Hence, a_{ij} is the probability that a user of G_i speaks an utterance that a user of G_j can understand. We have $a_{ii} = 1$ and $0 \leqslant a_{ij} \leqslant 1$. In general, $a_{ij} \neq a_{ji}$. The matrix A plays an important role in the dynamics of grammars, because it defines the fitness of individuals and the probability of making mistakes in learning.

Let us consider a population of constant size. Each person uses only one grammar. The fraction of people who speak grammar G_j is denoted as x_j. We have $\sum_{j=1}^{n} x_j = 1$. Individuals reproduce according to their fitness, and children learn the language of their parents. For simplicity, we assume that each person has only one parent, i.e. each child learns from one teacher. We define the fitness of an individual with grammar G_i as

$$f_i = f_0 + \frac{1}{2}\sum_{j=1}^{n}(a_{ij} + a_{ji})x_j. \qquad (1)$$

Here, f_0 is the background fitness which does not depend on the person's language. The language-related fitness depends on the individual's ability to communicate, i.e. the number (or fraction) of sentences he has in common with other people. Note that in this model, every grammar is as good as any other, and the ability to communicate depends only on the fraction of sentences that can be exchanged with other individuals.

We allow for mistakes during language acquisition. It is possible to learn from a person with grammar G_i and end up speaking grammar G_j. The probability of such transition is denoted as Q_{ij}. The matrix Q depends on the matrix A because the latter one defines how close different grammars are to each other (and therefore, how easy it is to confuse them with each other). The dependence of Q on A can be modeled if we make assumptions on how exactly the learning process takes place (see Sections 4 and 5).

The dynamics of a population $(x_1, \ldots, x_n)$ can be captured by the following general system of ordinary differential equations:

$$\dot{x}_j = \sum_i f_i x_i Q_{ij} - \phi x_j, \quad 1 \leqslant j \leqslant n, \qquad (2)$$

where $\phi = \sum_{m=1}^{n} f_m x_m$ is introduced to ensure the conservation of the population size. ϕ has the meaning of the average fitness of the population, and its language-dependent part is the *grammatical coherence*: it defines the probability that a sentence said by one person is understood by another person. Equation (2) is similar to a quasi-species equation (Eigen & Schuster, 1979), but has frequency-dependent fitness values (Nowak, 2000).

3. Dynamics of a Fully Symmetric System

In order to investigate system (2), we need to specify the matrices A and Q. Let us consider the simplest case where all $a_{ij} = a$, a constant, for all $i \neq j$, and $a_{ii} = 1$. We will refer to such a matrix as a *fully symmetric A* matrix. It corresponds to the situation where all grammars have the same distance from each other. The fitness in this case is simply

$$f_i = (1 - a)x_i + a + f_0. \qquad (3)$$

Next, we introduce the notion of a *learning accuracy*, *q*, which is the probability to learn grammar G_i given that the teacher speaks G_i. Our assumption of an equidistant configuration of grammars suggests that all G_i are equally easy (or hard) to confuse with each other. Namely, if a mistake is made, then it is equally likely that the person will speak G_j, $j \neq 1$, for any j. The probability to be taught G_i and learn G_j is $u = (1 - q)/(n - 1)$, for each $j \neq i$. The quantity u is called the *error rate* of grammar learning. Thus, the Q matrix is defined by

$$Q_{ii} = q, \quad Q_{ij} = u = (1 - q)/(n - 1), \quad i \neq j. \qquad (4)$$

The learning accuracy satisfies $1/n \leqslant q \leqslant 1$, where $q = 1$ means that no mistakes are made and $q = 1/n$ means that it does not matter what the teacher's grammar is, the choice of the resulting language is completely random. With these assumptions, system (2) becomes

$$\dot{x}_j = (1 - a)\left[-x_j^3 + x_j^2 q + \sum_{i \neq j} x_i^2\left(\frac{1-q}{n-1} - x_j\right)\right] - \frac{(a + f_0)(1 - q)(nx_j - 1)}{n - 1} \qquad (5)$$

for all $1 \leqslant j \leqslant n$.

3.1. FIXED POINTS

To begin, we will look for fixed points of system (5). Let us set $x_l = X$, $x_m = (1 - X)/(n - 1)$, $m \neq l$. This corresponds to the case where all grammars except one are used with the same frequency. Without loss of generality, we can take $l = 1$. From system (5) with a zero left-hand side, we obtain n equations for the unknown X. They are compatible, because the equations for $x_2, x_3, \ldots, x_n$ are identical, and their sum is just the equation for x_1 (due to the conservation of the number of people). In other words, each of the equations from the second to the last one is nothing but the first equation divided by $n - 1$. Therefore, we only need to solve the first equation,

$$X^3 - X^2 q + \frac{(1-X)^2}{n-1}\left(X - \frac{1-q}{n-1}\right) + \frac{(1-q)(a+f_0)(nX-1)}{(1-a)(n-1)} = 0. \qquad (6)$$

It has three solutions. One of them is

$$X_0 = 1/n \qquad (7)$$

and corresponds to the uniform distribution (i.e. all grammars occur in the population equally often). The other two solutions are

$$X_\pm = \frac{-(1-a)(1+(n-2)q) \mp \sqrt{D}}{2(a-1)(n-1)}, \qquad (8)$$

where

$$D = 4[-1 - a(n-2) - f_0(n-1)](1-q)(n-1) \times (1-a) + (1-a)^2[1+(n-2)q]^2. \qquad (9)$$

These two solutions describe a less symmetrical situation, when one grammar is the most (least) preferred one and is used with frequency $X_\pm$, and the rest of the grammars are used equally often. These solutions only exist if $D \geqslant 0$. Therefore, the existence condition is $q \geqslant q_1$, where

$$q_1 = \frac{4 + 2W(n-1)^{3/2} - 2f_0(n-1)^2 - 3n - a(2n^2 - 7n + 6)}{(1-a)(n-2)^2} \qquad (10)$$

and $W = \sqrt{(1+f_0)[1 + a(n-2) + f_0(n-1)]}$. In the special case of $n = 2$, q_1 is given by $q_1 = (3 + a + 4f_0)/(4(1+f_0))$. For $n \gg 1/(a + f_0)$, we have

$$q_1 = \gamma + O\left(\frac{1}{n}\right), \qquad (11)$$

where

$$\gamma = \frac{2}{1-a}\left(\sqrt{(a+f_0)(1+f_0)} - (a+f_0)\right). \qquad (12)$$

We observe (see Fig. 1) that γ is a monotonically increasing function of a and it is equal to 1 when $a = 1$. If a is close to 1, so that $a = 1 - \varepsilon$ and $\varepsilon \to 0$, we have $\gamma = 1 - \varepsilon/(4(f_0 + 1)) + O(\varepsilon^2)$. The coefficient γ also grows with f_0 reaching 1 as $f_0 \to \infty$. More precisely, we have $\gamma = 1 - (1-a)/(4f_0) + O(1/f_0^2)$.

In the special case of $a = f_0 = 0$, the existence condition looks like

$$q_1 = \frac{4 + 2(n-1)^{3/2} - 3n}{(n-2)^2}. \qquad (13)$$

For $n \gg 1$ we obtain $q_1 = 2/\sqrt{n} + O(1/n)$, i.e. the asymptotic behavior is quite different.

Solution (8) is shown in Fig. 2. For all values of a and f_0, at $q = 1$ we have $X_+ = 1$ and $X_- = 0$. At the point where the solution first appears ($q = q_1$), the value is $X_\pm = \sqrt{a}/(1 + \sqrt{a})$.

We note that because of the choice of the A and Q matrices, system (5) is highly symmetrical and its solutions are degenerate. Namely, by relabeling variables, we can pick any of the n grammars to be the "chosen" one, and then we will have n equivalent solutions of the form

$$x_l = X, \quad x_j = \frac{1-X}{n-1}, \quad \text{where } X = X_0, \; X_+ \text{ or } X_- \quad \forall j \neq l \qquad (14)$$

for any l such that $1 \leqslant l \leqslant$ n. Perturbations of the A or Q matrix will in general lift the degeneracy, which may result in the following changes: (i) in general, all values of x_j, $j \neq l$, will be different from each other, and (ii) solutions of form (14) will have different shapes for different values of

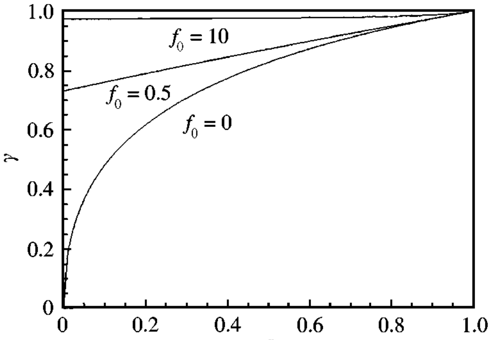

FIG. 1. The threshold value, γ, of learning accuracy, in the limit of large values of n. For $q > q_1 \approx \gamma$, asymmetric solutions become possible. The coefficient γ is plotted as a function of a for different values of the background fitness, f_0.

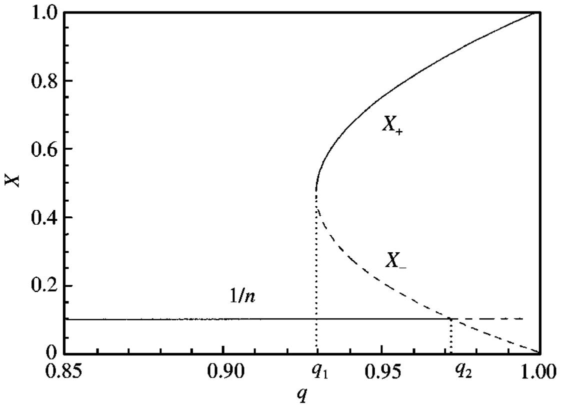

FIG. 2. The solutions $X = X_0$, X_+ and X_-. Here, $a = 0.5, f_0 = 1$ and $n = 10$. Stable solutions are represented by solid lines, and unstable ones by dashed lines (see Section 3.2).

l (in other words, X_0, X_+ and X_- will depend on l).

In the next section, we will prove that solution (14) with $x_l = X_-$ is always unstable and the one with $x_l = X_0$ (the uniform solution) loses stability as q grows. Only solutions with $x_l = X_+$ remain stable for high values of learning accuracy. When the A matrix is not fully symmetric, the X_+-type solutions have a more complicated form, but one important feature persists. Namely, these solutions can be characterized by one grammar whose share grows as q approaches unity, whereas the frequency of other grammars decreases. We will refer to such solutions as *one-grammar* solutions. G_l will be called the *preferred*, or "chosen", grammar and the grammars G_j with $j \neq l$—*secondary grammars.*

We would like to emphasize that the fixed points found in this section are not the only possible fixed points of system (2). In Appendix A.1 we demonstrate the existence of other classes of steady-state solutions. It turns out that for fully symmetric systems such solutions are always unstable. In the rest of this section, we only concentrate on the three fixed points found above.

3.2. STABILITY OF THE FIXED POINTS

Let us check the stability of solution (14); we will take $l = 1$. Following the well-developed techniques of a linear stability analysis, we perturb the solution by taking $x_1 = X + \tilde{y}_1$, $x_j = (1 - X)/(n - 1) + \tilde{y}_j$, $j > 1$ (here X can be X_0, X_+ or X_-). We substitute this into system (5) and linearize with respect to $\tilde{y}_j$. Next, we assume the exponential behavior of the perturbation, i.e. $(\tilde{y}_1, \dots, \tilde{y}_n)^{\mathrm{T}} = e^{\Gamma t}(y_1, \dots, y_n)^{\mathrm{T}}$. The system of linear equations for $y_1, \dots, y_n$ has the form

$$ay_1 + b \sum_{m>1} y_m = 0, \quad cy_j + d \sum_{\substack{m>1\\ m\neq j}} y_m + ey_1 = 0$$

$$2 \leqslant j \leqslant n, \tag{15}$$

where a, b, c, d and e are constants. Because of the conservation of the number of people, we have $\sum_{j=1}^{n} y_j = 0$. Replacing y_1 by $-\sum_{m=2}^{n} y_m$, we obtain

$$(a - b) \sum_{m=2}^{n} y_m = 0, \tag{16}$$

$$(c - d)y_j + (d - e) \sum_{m=2}^{n} y_m = 0, \quad 2 \leqslant j \leqslant n. \tag{17}$$

Here, the first equation is the sum of the other $(n - 1)$ equations [by construction of eqn (2)] and is therefore satisfied as long as the other $(n - 1)$ equations are satisfied. To ensure the existence of non-trivial solutions of linear system (17), we require that the determinant of the corresponding $(n - 1) \times (n - 1)$ matrix is zero. The matrix $[M_{ij}]$ has the form $M_{ii} = c - d$, $M_{ij} = d - e$ for $i \neq j$, and its determinant is given by

$$(c-d)^{n-2}(c-e+(n-2)(d-e)). \qquad (18)$$

The expressions for c, d and e are

$$c = \frac{\Gamma}{a-1} + (n+1)\left(\frac{1-X}{n-1}\right)^2 - 2q\frac{1-X}{n-1} + X^2 + \frac{n(a+f_0)(1-q)}{(n-1)(1-a)},$$

$$d = 2\frac{(1-X)(q-X)}{(n-1)^2}, \quad e = 2\frac{X(q-X)}{n-1}.$$

Determinant (18) is zero if $c = d$ (the corresponding Γ is denoted as Γ_1) or if $c - e + (n-2)(d-e) = 0$ (the corresponding Γ is denoted as Γ_2). Note that in the special case of $n = 2$ we only have the latter condition. By examining the sign of $\Gamma_{1,2}$, we can study the stability of solutions X_0, X_+ and X_-. If at least one of the growth rates is positive, the corresponding solution is unstable.

3.2.1. *The Uniform Solution*

For $X = X_0 = 1/n$, we have

$$\Gamma_1 = \Gamma_2 = \frac{1}{n(n-1)}[(n(2q-1)-1)(1-a) - n^2(1-q)(f_0+a))]. \qquad (19)$$

This gives a threshold condition for learning accuracy. Namely, for $q > q_2$, $\Gamma_{1,2}$ become positive and the uniform solution loses stability. The value q_2 is given by

$$q_2 = \frac{n^2(f_0+a)+(n+1)(1-a)}{n[n(f_0+a)+2(1-a)]}. \qquad (20)$$

The value q_2 corresponds to the point where $X_- = X_0$. Thus, the uniform solution loses stability at the point q where it meets solution X_-. For large n ($n \gg 1/(a+f_0)$), we have

$$q_2 = 1 - \frac{1}{n}\left(\frac{1-a}{a+f_0}\right) + O\left(\frac{1}{n^2}\right). \qquad (21)$$

Note that in the case $a = f_0 = 0$, we have $q_2 = 1/2 + 1/(2n)$.

3.2.2. *The Asymmetric Solutions*

First, we examine the case $n > 2$. The growth rate for the two asymmetric solutions is presented in Fig. 3. It turns out that for the solution X_+, both Γ_1 and Γ_2 are non-positive for all $q \geqslant q_1$ (the solid lines in Fig. 3). This means that the asymmetric solution X_+ is stable everywhere in the domain of its existence. Thus, for higher values of learning accuracy, the system prefers a state where one of the grammars is used very often, whereas the rest of them have an equal (and small) share.

For X_-, the situation is different. In the domain $q_1 \leqslant q \leqslant 1$, one of the growth rates is positive whereas the other is negative (at the point $q = q_2$ they are both zero, the dotted lines in Fig. 3). This means that the solution X_- is unstable (it is neutrally stable for $q = q_2$). It is instructive to compare the eigenvectors corresponding to the eigenvalues Γ_1 and Γ_2. The former one has $y_1 = 0$, and the latter one has $y_1 \neq 0$. For $q > q_2$, $\Gamma_1 > 0$, which means that the solution X_- loses stability in such a way that x_1 stays the same, but the rest of the grammars fail to keep a uniform distribution.

For completeness we consider the value $n = 2$. In this special case, q_2 coincides with q_1. Therefore, for $q < q_1 = q_2$, the uniform solution $x_{1,2} = 1/2$ is stable, and for higher values of

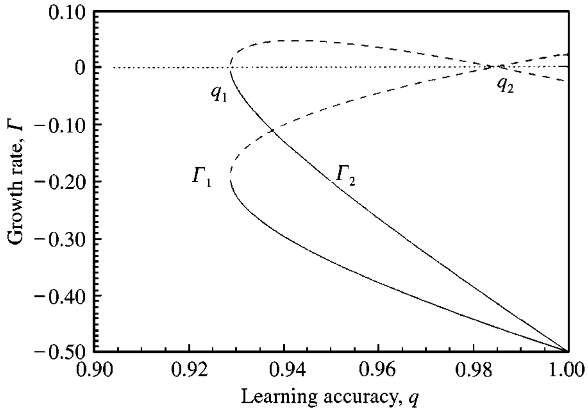

FIG. 3. The growth rates for the one-grammar solutions X_+ (——) and X_-(– – – –), as functions of q. Here, $a = 0.5$, $f_0 = 1$ and $n = 20$.

learning accuracy, it loses stability. We have a pitchfork bifurcation with two equivalent stable solutions, $(x_1, x_2) = (X_+, X_-)$ and $(x_1, x_2) = (X_-, X_+)$.

3.3. THE BIFURCATION SCENARIO

To sum up the bifurcation picture (Fig. 2), we note that for $0 \leqslant q < q_1$ the only stable solution is the uniform solution $1/n$, then between q_1 and q_2 both the uniform solution and solutions (14) with X_+ (the one-grammar solutions) are stable, and finally, for $q > q_2$ the uniform solution loses its stability and the one-grammar solutions remain stable.

At the point $q = q_1$, where the non-uniform solutions first appear, the corresponding average fitness (assuming that n is large) is

$$\phi_{asym} = \frac{(\sqrt{(a+f_0)(1+f_0)} - f_0 - a)^2}{1-a} + a + f_0, \tag{22}$$

whereas the average fitness of the uniform solution (for large n) is

$$\phi_{unif} = a + f_0. \tag{23}$$

One can see that as the system goes to a one-grammar solution, the average fitness (and the grammatical coherence) experience a jump, $\Delta f = (1-a)(\gamma/2)^2 + O(1/n)$, see Fig. 4. Note that if $a = 1 - \varepsilon$, then $\Delta f \sim \varepsilon/4$. As q increases to 1, the total fitness of the one-grammar solution monotonically increases to $1 + f_0$, whereas the fitness of the uniform solution stays constant.

It is convenient to present the stability diagram in terms of the error rate, u (see Fig. 5). Clearly, as n grows, it becomes harder and harder to maintain one grammar. Also, one can see that there is always a bistability region where the uniform solution and X_+ coexist. Indeed, for the existence of a one-grammar solution we need

$$u \leqslant u_1 = c_1/n, \quad c_1 \equiv 1 - \gamma. \tag{24}$$

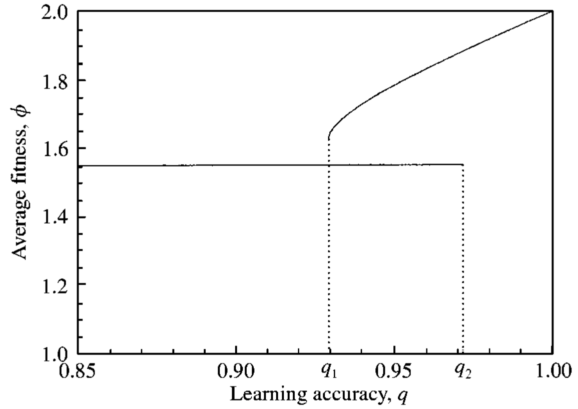

FIG. 4. Total fitness of the stable solutions of a system with a fully symmetric A matrix, as a function of learning accuracy, q. Parameters of the system are as in Fig. 2.

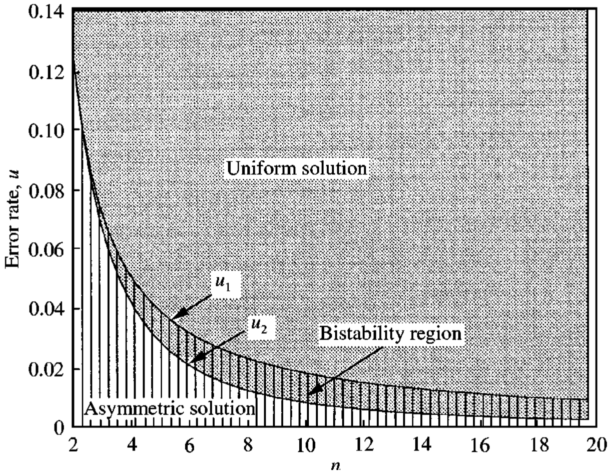

FIG. 5. The stability diagram in terms of the error rates, $u_{1,2}$. Here, $a = 0.5, f_0 = 1$.

For the uniform solution to lose stability we need

$$u \leqslant u_2 = c_2/n^2, \quad c_2 \equiv (1-a)/(a+f_0). \tag{25}$$

The above inequalities are derived in the case of large n and $a + f_0 > 0$.

4. Memoryless Learning

In this section, we will consider a particular learning algorithm, namely, the memoryless learner algorithm. It will allow us to relate the entries of the A matrix to the error rate. Then we will present some examples, where some or all symmetries of the A matrix are broken and the Q matrix is determined according to the memoryless learner formulation. The first example assumes that all but one grammars are in some sense equivalent, which means that certain symmetries remain in the system, even though the corresponding A matrix is no longer fully symmetric (Section 4.2). In Section 4.3, we describe a very general system where no symmetries remain. As a result, a very interesting bifurcation diagram is observed, where some of the grammars are suppressed and others are enhanced. Appendices A.2 and B present more examples.

4.1. THE ALGORITHM

As an example of a simple learning mechanism, we will use the *memoryless learner* algorithm (Niyogi, 1998). We suppose that the learner starts by (randomly) choosing one of the n grammars as an initial state. Then b sample sentences are received from the teacher. For each sampling, the learner compares the sentence uttered by the teacher with his own grammar. If the sentence is consistent with the learner's grammar, no action is taken; otherwise, the learner randomly picks a different grammar. The initial probability distribution of the learner is uniform: $\mathbf{p}^{(0)} = (1/n, \ldots, 1/n)^{\mathrm{T}}$, i.e. each of the grammars has the same chance to be picked at the initial moment. The discrete time evolution of the vector $\mathbf{p}^{(t)}$ is a Markov process with a transition matrix, $T(k)$, which depends on the teacher's grammar, k. This matrix is defined by $T(k)_{ij} = (1 - a_{ki})/(n-1)$ for $i \neq j$ and $T(k)_{ii} = a_{ki}$. After b samplings, the k-th row of matrix Q is the string-vector $(\mathbf{p}^{(b)})^{\mathrm{T}}$ obtained with the transition matrix $T(k)$. Therefore, we can write

$$Q_{ij} = [(\mathbf{p}^{(0)})^{\mathrm{T}} T(i)^b]_j. \tag{26}$$

This expression models the connection between matrices A and Q. For instance, if we assume that the off-diagonal entries of the A matrix are constant and equal to each other (the fully symmetric case), then, according to eqn (26), the off-diagonal entries of the Q matrix are also equal to each other, and eqn (4) holds. Expression (26) can be used to evaluate the learning accuracy and the error rate in terms of a:

$$q = 1 - \left(1 - \frac{1-a}{n-1}\right)^b \frac{n-1}{n}, \tag{27}$$

$$u = \frac{1}{n}\left(1 - \frac{1-a}{n-1}\right)^b. \tag{28}$$

Note that $\lim_{b\to\infty} q = 1$ for any fixed n and $0 \leqslant a < 1$. This means that the more samplings are available, the more precise the learning process becomes. Also, when $a = 1$ (no difference between grammars), $q = 1/n$ which is the lowest possible value of learning accuracy. We can use results of the previous section to find conditions for b, the number of sampling sentences per individual, which would allow the population to maintain a particular grammar. We will assume that the number n is large and use inequalities (24) and (25). In order for solution X_+ to exist, we must have

$$b \geqslant b_1 = \frac{n}{1-a} \log \frac{1}{c_1}. \tag{29}$$

The uniform solution loses stability if

$$b \geqslant b_2 = \frac{n}{1-a} \log \frac{n}{c_2}. \tag{30}$$

The constants c_1 and c_2 are defined in formulas (24) and (25). Now we turn to some examples.

4.2. BREAKING THE SYMMETRY OF THE A MATRIX

The A matrices that we have considered so far possessed such symmetries that all one-grammar solutions were identical for all grammars G_i. This is not the case in general. All non-symmetrical perturbations of a fully symmetric A matrix lead to the effect of suppressing some grammars and enhancing others. For instance, if we take a fully symmetric A matrix and replace one element $a_{ij} = a$ with $a + \xi$, we will observe the following picture. The branch of the stable asymmetric solution X_+ corresponding to the grammar G_i will split off from the other one-grammar solutions, whereas solutions with grammars G_l, $l \neq i$, $l \neq j$, will stay together (the one-grammar solution with G_j as the preferred grammar will

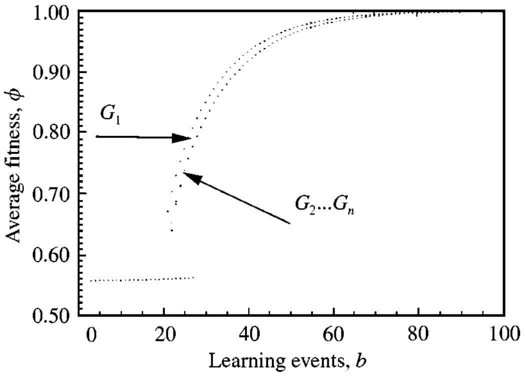

FIG. 6. The growth rates for the asymmetric matrix A with all off-diagonal $a_{ij} = 0.5$ except $a_{12} = 0.1$. The solution with the G_1 as the preferred grammar is advantageous in comparison with the rest of the one-grammar solutions, it has a higher coherence and comes into existence for smaller values of b. The grammar G_2 is slightly suppressed.

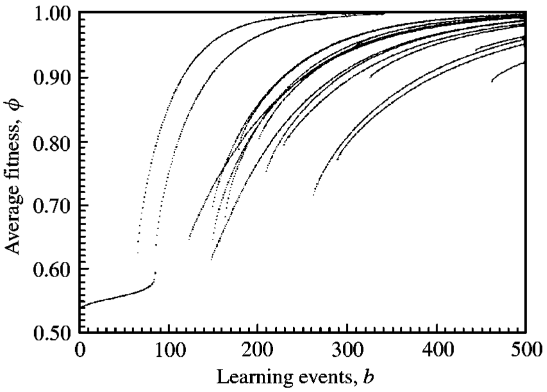

FIG. 7. The fitness of the system with the A matrix consisting of uniformly distributed random numbers $0 \leqslant a_{ij} \leqslant 1$. The number of grammars is $n = 20$, and $f_0 = 0$.

deviate ever so slightly from the rest of the grammars). It turns out that if $\xi > 0$, the grammar G_i will be suppressed (and the grammar G_j will be very slightly advantaged), and if $\xi < 0$, the grammar G_i will be enhanced (and G_j will be slightly suppressed). This means that for $\xi < 0$, the solution with grammar G_i will come into existence earlier (for smaller values for q and b) and will have a larger total fitness (see Fig. 6, where $i = 1$, $j = 2$, $a = 0.5$ and $\xi = -0.4$). This makes sense because negative (positive) values of ξ means that the grammar i has a smaller (larger) intersection with the rest of the grammars than the rest of them. When this grammar becomes preferred, it stands out more (less) than other grammars would in its place, i.e. it corresponds to higher (lower) values of X_+ and has a larger (smaller) total fitness.

4.3. RANDOM OFF-DIAGONAL ELEMENTS

Another example of a non-symmetrical system is the case of an A matrix with random elements, see Fig. 7. We take the off-diagonal elements of the A matrix to be random numbers uniformly distributed between zero and one. If the consequence, no symmetries are left in the system. If the number of learning events, b, is high, there are still n stable one-grammar solutions (17 of 20 can be seen in Fig. 8). The difference with the fully symmetric case is that here, one-grammar solutions with different dominant grammars correspond to different values of ϕ, the grammatical coherence of the population. Thus, each of the solutions is represented by a separate line. The number of stable one-grammar solutions grows with b. Some of the grammars become advantaged and have a lower threshold of existence. Some are suppressed until much higher values of b. Such behavior was already present in the example of Section 5.5. The value of b at which the first bifurcation takes place can be roughly predicted by using formula (29) with $a = 1/2$, i.e. the average value of the elements a_{ij}.

Another interesting feature that can be clearly observed in Fig. 7 is that the lowest fitness solution (which corresponds to the uniform solution of the fully symmetric case) flows smoothly into one of the one-grammar solution (the "second best" one for this particular realization of A). This effect can be predicted from the standard bifurcation theory. Namely, general perturbations of a pitchfork-like bifurcation will lead to smoothing out sharp edges and avoiding cross-sections, and might also cause the disappearance of "knees" (bistability regions) like those in Fig. 2.

We have performed computer experiments with different distributions of the elements a_{ij} of random A matrices and observed the following dependencies:

- The first bifurcation point b at which a coherent grammar emerges can be roughly estimated with formula (29) where we use the average value $a = \langle a \rangle$.

- The width of the interval over which various grammars bifurcate increases with the width of the distribution of the a_{ij} values, and also it becomes larger as $\langle a \rangle$ grows.

We conclude that systems with random A matrices behave in a predictable way, and many of the elements of the dynamics can be understood from the analysis of symmetrical systems and their perturbations. However, an extended study of a system with randomly chosen a_{ij} is still needed to describe the full bifurcation picture.

5. Batch Learning

Another example of a learning algorithm is given by the so-called *batch learner* pattern (Niyogi, 1998). According to this algorithm, the learner receives a set of b sentences from the teacher and then decides which grammar he will use. The resulting grammar has to be consistent with all of the sentences uttered by the teacher. If all the b sentences happen to be consistent with more than one grammar (say, with r grammars), then the learner can pick any of the r grammars with probability $1/r$. We will show that this mechanism is much more effective than memoryless learning.

Let us derive equations relating the learning accuracy, q and the error rate, u, with the number of sample sentences, b, a batch learner's equivalent of eqns (27–28). It turns out that the information given by the A matrix is not sufficient to give estimates for the learning accuracy. The A matrix only reflects intersections of pairs of grammars, but it does not specify the intersections of three or more grammars. The following example demonstrates that for the same A matrix, different configurations of grammars lead to very different learning accuracies for a given number of sampling sentences.

Consider a set of n grammars with a fully symmetric A matrix, $a_{ij} = a$. Let us assume that the teacher speaks grammar G_1, so that the b sentences received by the learner are all consistent with G_1. Let us denote the intersections $G_1 \cap G_i \equiv D_i$, $i \neq 1$. The size of D_i is $p[D_i] = a$. Now, let us look at two cases:

(i) All the sets D_i coincide, i.e. $D_i = D_j$ for all $2 \leqslant$ i, $j \leqslant n$,

(ii) $D_i \cap D_j = \emptyset$ for all $2 \leqslant i, j \leqslant n$.

Note that in the second case we need to assume that $a(n-1) \leqslant 1$. It is easy to see that the probability to learn the teacher's grammar in the two cases is given by

$$\text{case (i)} \quad Q_{11} = 1 - a^b \frac{n-1}{n}, \tag{31}$$

$$\text{case (ii)} \quad Q_{11} = 1 - a^b \frac{n-1}{2}. \tag{32}$$

For the same values of a, n and b, the learning accuracy in the first case is much higher than it is in the second case.

Let us now consider a random configuration of grammars. Again, we have $n-1$ sets D_i of size a inside the set G_1 of size 1. The learning accuracy, q, can be calculated as

$$q = \sum_{j=1}^{n} p(j)/j, \tag{33}$$

where $p(j)$ is the probability that all of the teacher's sentences belong to an intersection of j different grammars $G_{\alpha 1}, \ldots, G_{\alpha j}$ (and to no other grammars). All of the sample sentences by definition belong to the "correct" grammar. The probability that b sentences belong to one of the $n-1$ "wrong" grammars is a^b. The probability to belong to exactly j "wrong" grammars simultaneously is $(a^b)^j(1-a^b)^{n-1-j}$, and there are $C^j_{n-1} = (n-1)!/j!/(n-1-j)!$ ways to choose j grammars out of the lot of $n-1$ grammars. We have

$$q = \sum_{j=1}^{n} C^{j-1}_{n-1} \frac{(a^b)^{j-1}(1-a^b)^{n-j}}{j} = \frac{1-(1-a^b)^n}{a^b n}. \tag{34}$$

The error rate is given by $u = (1-q)/(n-1)$. For $a \to 0$ we have $q \to 1$, and for $a = 1$, the learning accuracy is $q = 1/n$. Note that in the limit of small values of a we recover eqn (32) for "sparse" grammars.

The random configuration of grammars is not the hardest one to learn. We can design an example where for the same values of parameters, the learning accuracy is even lower. Imagine the situation when the sets D_i form a neat N-layer tiling of the set G_1. In other words, let us assume that the number $N = a(n-1)$ is integer, and that exactly $1/a$ "wrong" grammars compose each of the layers. The layers have no gaps, i.e. within each layer, $\cup_{k=1}^{1/a} D_k = G_1$, and $D_k \cap D_j = \emptyset$. Then the learning accuracy is given by

$$q = \frac{1-(1-a^{b-1})^{N+1}}{a^{b-1}(N+1)}. \qquad (35)$$

This value of q is slightly smaller than the one defined in formula (34) for all $0 < a < 1$.

On the other hand, case (i) above describes the easiest possible configuration of grammars for batch learning. This follows from eqn (33) and the inequality $\sum_{j=2}^{n} p(j) \geqslant a$. The case when $p(j) = 0$ for all $2 \leqslant j \leqslant n-1$ and $p(n) = a$ [i.e. case (i)] minimizes the right-hand side of eqn (33) under the above restriction.

Finally, we will give an estimate for the bifurcation point, b_1, for the batch learner algorithm. The result is implicit and for $n \gg 1/(a+f_0)$ it can be written as

$$b_1 = \frac{\log(x/n)}{\log a}, \qquad (36)$$

where x is the solution of the equation $1 - xq_1 - \exp^{-x} = 0$ (for the estimate of b_2 we need to replace q_1 by q_2 in the equation for x). Since $q_1 \gg 1/n$, we have $x \ll n$, and therefore its contribution in formula (36) can be neglected. We have

$$b_1 = \frac{\log n}{\log(1/a)}. \qquad (37)$$

We can see that the number of sample sentences needed for a community of batch learners to develop a coherent language grows as $\log n$, whereas memoryless learners need $b \propto n$ sentences [formula (29)]. This is a consequence of the fact that batch learners have a perfect memory, whereas memoryless learners only remember one sentence at a time.

6. Conclusions and Discussion

We have studied an evolutionary model of grammar learning. It has been shown that grammatical coherence is possible if the learning accuracy of children is sufficiently high. This is a general result; the details of the threshold condition depend on the assumptions on the learning procedure that children use. For a memoryless learner we find that the total number of sample sentences, b, that a child receives during the language acquisition phase, must exceed a constant times n, the number of candidate grammars. For a batch learner, we obtain that b has to exceed a constant times the logarithm of n. The memoryless learning algorithm makes the minimum demands on the cognitive abilities of the individual. The batch learner represents the other extreme; it remembers (a fraction of) all sentences and then chooses a grammar that is consistent with all memorized sentences. The human strategy of grammar acquisition will be somewhere between these two limiting cases. If the threshold condition is satisfied, then a community of individuals can maintain a coherent grammatical system.

These results can also be discussed in terms of the *principles and parameters* framework proposed by Chomsky (1981). Universal grammar is specified by genetically inherited principles which limit the number of candidate grammars. The candidate grammars differ in terms of parameter settings which have to be learned. If these are k independent, binary parameters, then the number of candidate grammars is $n = 2^k$. For the memoryless learner we require that k is less than a constant times the logarithm of b. For the batch learner we need k to be less than a constant times b. The innate principles have to reduce the number of parameters to fulfill these conditions. If these conditions are not fulfilled then a population of individuals cannot maintain or evolve a consistent grammar.

The dynamics of grammar acquisition strongly depend on the parameter b. If b is below the threshold, all grammars are used in the population with a roughly similar frequency, and the resulting average fitness is low. As the number of learning events, b, increases beyond its threshold value, the system experiences a spontaneous symmetry breaking. The average fitness increases (often discontinuously) to reach a new, higher value. This means that one of the grammars becomes dominant, and as a result most of the people in the population can communicate successfully. An interesting part is that in principle, any of the given grammars that constitute the search space, can become dominant, if the number of learning events is sufficiently high. The system has multiple (up to n) stable equilibria. Which grammar becomes preferred in the population depends on the initial distribution of grammars.

An alternative approach to the problem of the evolutionary selection of coherence is stochastic modeling. Computer simulations can be carried out where in a population of individuals, children receive b sentences from their parents, and deduce the correct grammar using some learning algorithm. This process is described by the deterministic equations (2) exactly if the number of people in the population tends to infinity. For smaller population sizes, stochastic effects play a role. In fact, we expect that they may introduce some qualitatively new behavior which is suppressed in the deterministic model. From our analysis it follows that once the deterministic system has relaxed to one of the stable fixed points, no further change is possible. On the other hand, in a stochastic system, finite size effects act as perturbations to the smooth dynamics of eqns (2) and might lead to spontaneous changes in the grammatical system in time. Even if one of the grammars dominate the population for some time, it may happen that the system will be kicked out of this state and relax to a different one-grammar solution. Results of such simulations will be reported elsewhere.

Another interesting extension of the framework developed here is to look at the competition among different universal grammars. An existence of coherent solutions is clearly a necessary condition for the evolvability of universal grammar. Once universal grammar is in place, a coherent language will be maintained in the population. However, in order to demonstrate that universal grammar has come about by means of Darwinian evolution, it is important to look at competition of different types of universal grammars. Equations of type (2) can be used. The first step has been made in Komarova & Nowak (2001), where a one-parametric family of universal grammars was considered. All the universal grammars were identical (had the same search space and the same learning mechanism) except for the number, b, of sampling events available to children during the grammar learning phase. As a result, an intermediate value of b was selected which maximized the reproductive rate times learning accuracy. The next step will be to find out what evolutionary pressures act on the selection of the learning mechanism.

Finally, we note that in the current model, no spatial variations have been taken into account. Equations (2) can be easily modified to describe diversity of languages by including space-dependence. Then, at different regions, different grammars can become dominant and some interesting spatial dynamics at the grain boundaries may be observed.

Support from the Packard Foundation, the Leon Levy and Shelby White Initiatives Fund, the Florence Gould Foundation, the Ambrose Monell Foundation, the Alfred P. Sloan Foundation and the National Science Foundation is gratefully acknowledged.

REFERENCES

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

BRANDON, R. & HORNSTEIN, N. (1986). From icons to symbols: some speculations on the origins of language. *Biol. Philos.* **1,** 169–189.

CHOMSKY, N. (1959). Review of Skinner 1957. *Language* **35,** 26–58.

CHOMSKY, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

CHOMSKY, N. (1980). *Rules and Representations*. New York: Columbia University Press.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

CHOMSKY, N. (1993). A minimalist program for linguistic theory. In: *The View form Building 20* (Hale, K. & Keyser, S. J., eds), pp. 1–52. Cambridge, MA: MIT Press.

DEACON, T. (1997). *The Symbolic Species*. New York: W. W. Norton.

EIGEN, M. & SCHUSTER, P. (1979). *The Hypercycle: A Principle of Natural Self-Organization*. Berlin; New York: Springer-Verlag.

GIBSON, E. & WEXLER, K. (1994). Triggers. *Linguist. Inquiry* **25,** 407–454.

HASHIMOTO, T. & IKEGAMI, T. (1995). Evaluation of symbolic grammar systems. *Advances in Artificial Life* (Moran, F., Moreno, A., Merelo, J. J. & Chacon, P., eds), pp. 812–823. Springer-Verlag: Berlin.

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

Hauser, M. D. (1996). *The Evolution of Communication.* Cambridge, MA: Harvard University Press.

Haussler, D., Kearns, M., Seung, H. S. & Tishby, N. (1997). Rigorous learning curve bounds from statistical mechanics. *Mach. Learning* **25,** 195–236.

Hornstein, N. R. & Lightfoot, D. W. (1981). *Explanation in Linguistics.* London: Longman.

Jackendoff, R. (1997). *The Architecture of the Language Faculty.* Cambridge, MA: MIT Press.

Komarova, N. L. & Nowak, M. A. (2001). Natural selection of the critical period for grammar acquisition. *Proc. Royal Soc. B*, submitted.

Lightfoot, D. (1982). *The Language Lottery, Towards a Biology of Grammars.* Cambridge, MA: MIT Press.

Lightfoot, D. (1991). *How to Set Parameters: Arguments from Language Change.* Cambridge, MA: MIT Press.

Lightfoot, D. (1999). *The Development of Language: Acquisition, Changes and Evolution.* Malden, MA: Blackwell Publishers.

Niyogi, P. (1998). *The Informational Complexity of Learning.* Boston: Kluwer.

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

Niyogi, P. & Berwick, R. C. (1997). Evolutionary consequences of language learning. *Linguist. Philos.* **20,** 697–719.

Nowak, M. A. (2000). The basic reproductive ratio of a word, the maximum size of a lexicon. *J. theor. Biol.* **203,** 179–189.

Nowak, M. A., Komarova, N. L. & Niyogi, P. (2001). Evolution of universal grammar. *Science*, to appear.

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

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

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

Osherson, D., Stob, M. & Weinstein, S. (1986). *Systems That Learn.* Cambridge, MA: MIT Press.

Pinker, S. (1994). *The Language Instinct.* New York: W. Morrow & Company.

Pinker, S. (1999). *Words and Rules.* New York: Basic Books.

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

Pinker, S. & Prince, A. (1988). Regular and irregular morphology and the psychological status of rules of grammar. In: *Proceedings of the 17th Annual Meeting of the Berkeley Linguistics Society* (Sutton, L. A., Johnson, C. & Shields, R., eds) pp. 230–251, Berkeley: Berkeley Linguistics Society, University of CA, Berkeley.

Prince, A. & Smolensky, P. (1993). Optimality theory: constraint interaction in generative grammar. In: *Technical Report RuCCS TR-2*, pp. 234, Rutgers Center for Cognitive Science, Cambridge, MA: MIT Press.

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

Rumschitzky, D. (1987). Spectral properties of Eigen's evolution matrices. *J. Math. Biol.* **24,** 667–680.

Sorace, A., Heycock, C. & Shillcock, R., eds (1999). *Language Acquisition: Knowledge, Representation and Processing.* Amsterdam: North-Holland.

Tesar, B. & Smolensky, P. (2000). *Learnability in Optimality Theory.* MA: MIT Press.

Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM* **27,** 436–445.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

Wexler, K. & Culicover, P. (1980). *Formal Principles of Language Acquisition.* Cambridge, MA: MIT Press.

Appendix A

m-Grammar Solutions

A.1. A FULLY SYMMETRIC SYSTEM

A wider class of steady-state solutions of system (2) can be found if we assume that m of the n grammars are used with frequency $X^{(m)}$ and the other $n-m$ grammars—with frequency $(1-mX^{(m)})/(n-m)$. Obviously, solution (14) is a special case of this family with $m=1$. We can choose the m grammars out of n in $C_n^m = n!/(m!(n-m)!)$ ways which will give us C_n^m equivalent solutions. Without loss of generality we will take

$$x_l = X^{(m)}, \quad 1 \leqslant l \leqslant m \quad \text{and}$$

$$x_j = (1-mX^{(m)})/(n-m), \quad m+1 \leqslant j \leqslant n. \quad \text{(A.1)}$$

In order to find $X^{(m)}$ we can write eqn (3.1) generalized for $m \neq 1$. Obviously, for $m=0$ or n, the only solution is the uniform one, $X^{(0),(n)} = X_0 = 1/n$, and there are three solutions for $X^{(m)}$ with $1 \leqslant m \leqslant n-1$: $X_0 = 1/n$, $X_+^{(m)}$ and $X_-^{(m)}$. Solutions $X_\pm^{(1)}$ are given by eqn (8). For general m, we will describe general properties of solutions without giving explicit expressions. Solutions $X_\pm^{(m)}$ have an obvious symmetry property, namely, $X_\pm^{(n-m)} = (1-mX_\mp^{(m)})/(n-m)$. These solutions exist for $q \geqslant q_1^{(m)}$, the behavior of $q_1^{(m)}$ is shown in Fig. A1. One can see that the lowest threshold value corresponds to $m=1$ (or $m=n-1$). The asymptotic behavior of $q_1^{(m)}$ is as follows. If $n \gg 1/(a+f_0)$ and $\lim_{n\to\infty} m/n = M$, a constant between 0 and 1, we have

$$q_1 = 1 - \frac{1}{n}\frac{1-a}{4(a-f_0)M(1-M)} + O\left(\frac{1}{n^2}\right), \quad \text{(A.2)}$$

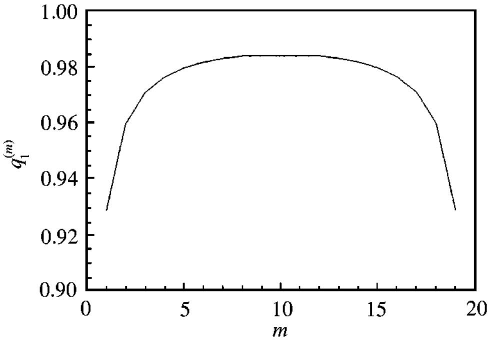

FIG. A1. The threshold value $q_1^{(m)}$ as a function of m; other parameters are $n = 20$, $a = 0.5$ and $f_0 = 1$.

and in the case when $\lim_{n\to\infty} m/n = 0$, we have eqn (11) with

$$\gamma = \frac{2}{1-a}\left(\sqrt{m}\sqrt{(a+f_0)(1+a(m-1)+mf_0)} - m(a+f_0)\right), \qquad \text{(A.3)}$$

which is a generalization of formula (12). When $q = 1$, we have $X_+^{(m)} = 1/m$ and $X_-^{(m)} = 0$. This means that the maximum fitness reached by such m-grammar solutions is $(1-a)/m + a + f_0$. Stability of solution (A.1) can be investigated. The system for the perturbations can be written as

$$Ay_l + B \sum_{\substack{1\leqslant i\leqslant m\\ i\neq l}} y_i = 0, \quad 1 \leqslant l \leqslant m, \qquad \text{(A.4)}$$

$$Cy_j + D \sum_{\substack{m+1\leqslant i\leqslant n\\ i\neq l}} y_i = 0, \quad m+1 \leqslant l \leqslant n, \qquad \text{(A.5)}$$

where coefficients A–D depend on parameters of the system and Γ, the growth rate. The above system can have non-trivial solutions if the determinant of the corresponding matrix is zero, which gives the following conditions: $C = 0$ (which results in an expression for $\Gamma_1^{(m)}$), $A + (m-1)B = 0$ (which gives $\Gamma_2^{(m)}$), $A = 0$ (which gives $\Gamma_3^{(m)}$) and $C + (n-m-1)D = 0$, the latter condition is redundant because the equations of system (A.4–A.5) are not all independent due to the conservation of the number of people.

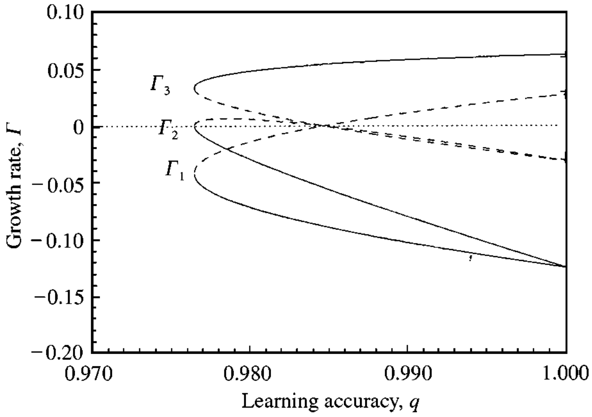

FIG. A2. The growth rates for the solutions $X_+^{(4)}$ (——) and $X_-^{(4)}$ (- - - - -), as functions of q. The other parameters are $n = 20$, $a = 0.5$, $f_0 = 1$ and $m = 4$.

The growth rates' dependence on q is shown in Fig. A2. The behavior of $\Gamma_1^{(m)}$ and $\Gamma_2^{(m)}$ is similar to the behavior of the growth rates in the case of $m = 1$, see Fig. 3. On the other hand, $\Gamma_3^{(m)}$ is a new feature. It turns out that for $m = 1$, $\Gamma_3^{(m)} = \Gamma_2^{(m)}$, but for all values of m larger than 1, $\Gamma_3^{(m)} > 0$ for the solution $X_+^{(m)}$, i.e. it is unstable. We conclude that for fully symmetric A matrices, m-grammar solutions are unstable for all $m > 1$.

A.2. A BI-DIAGONAL A MATRIX

We will now briefly discuss the example of a system where m-grammar solutions can be stable. This is a system with a bi-diagonal A-matrix. The A matrix is given by

$$a_{ii} = 1, \quad a_{ij} = a, \quad j \neq i, \quad j \neq n+1-i \quad \text{and}$$

$$a_{i,n+1-i} = a + c \qquad \text{(A.6)}$$

for all i. In other words, the main diagonal of the A matrix consists of ones, and the other diagonal has elements $a + c$, whereas the rest of the elements are all a. In Fig. A3 we can see the bifurcation diagram for such a system. We can see that a bi-diagonal system supports not only the uniform solution and the one-grammar solutions, but also solutions where two grammars are used equally often, whereas the others are distributed

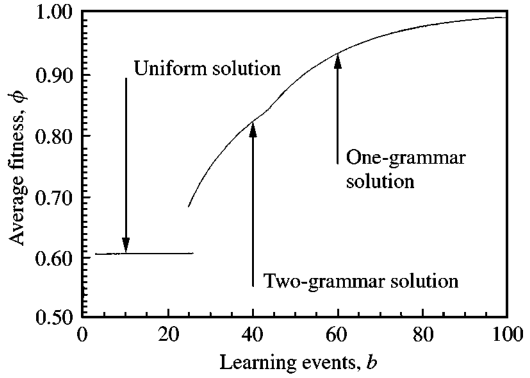

FIG. A3. The total fitness of a system with the matrix A given by eqn (A.4), as a function of b. The values are $a = 0.5$, $c = 0.35$, $n = 8$ and $f_0 = 0$.

uniformly. Such solutions are similar to solutions (A.1) of the previous section with $m = 2$. In systems with $c = 0$, these solutions were always unstable and only existed for $q_1^{(m)} > q_1$. For general values of c, two-grammar solutions can be stable. More precisely, there is a critical value c' such that if $c > c' > 0$, then $q_1^{(2)} < q_1$, i.e. two-grammar solutions appear earlier than one-grammar solutions (see Fig. A3), and they are stable. At the point where a two-grammar solution meets a one-grammar solution, it loses stability, and the system experiences the second bifurcation, to the solution of the X_+ type. There are n two-grammar solutions which look like $x_l = x_{n+1-l} = X_+^{(2)}$, $1 \leqslant l \leqslant n$.

Appendix B

n Equivalent Grammars

Here we work out an example where the A matrix is not fully symmetrical, but the n grammars are nevertheless equivalent to each other. We will show that symmetries in the configuration of the search space will lead to degeneracies in the space of equilibrium solutions.

Let us assume that the total number of sentences that each of the n grammars has is N (normally we think of grammars as being able to generate an infinite number of sentences, this is reached by taking N to infinity). Further, we suppose that each of the sentences can be formed in only two different ways, a and b. Therefore, each grammar can be represented as a sequence of N symbols chosen from the set $\{a, b\}$. There is a natural correspondence between such grammars and all binary numbers (just take $a = 0$ and $b = 1$), which gives a convenient way to order all the grammars. There are $n = 2^N$ competing grammars in total. The A matrix in this case is defined as follows: a_{ij} is the total number of positions at which grammars G_i and G_j have the same symbol, divided by N. For $N = 3$, this matrix is

$$A = \begin{pmatrix} 1 & 2/3 & 2/3 & 1/3 & 2/3 & 1/3 & 1/3 & 0 \\ 2/3 & 1 & 1/3 & 2/3 & 1/3 & 2/3 & 0 & 1/3 \\ 2/3 & 1/3 & 1 & 2/3 & 1/3 & 0 & 2/3 & 1/3 \\ 1/3 & 2/3 & 2/3 & 1 & 0 & 1/3 & 1/3 & 2/3 \\ 2/3 & 1/3 & 1/3 & 0 & 1 & 2/3 & 2/3 & 1/3 \\ 1/3 & 2/3 & 0 & 1/3 & 2/3 & 1 & 1/3 & 2/3 \\ 1/3 & 0 & 2/3 & 1/3 & 2/3 & 1/3 & 1 & 2/3 \\ 0 & 1/3 & 1/3 & 2/3 & 1/3 & 2/3 & 2/3 & 1 \end{pmatrix}. \tag{B.1}$$

One can see that A is not fully symmetric but it still contains certain symmetries (e.g. reflection with respect to both diagonals). It can be shown that the Q matrix in the case of the memoryless learner algorithm has the same structure as the A matrix. One can check that system (2) can be satisfied if we substitute

$$x_1 = \alpha, \quad x_2 = x_3 = x_5 = \beta,$$

$$x_4 = x_6 = x_7 = \gamma, \quad x_8 = \delta \tag{B.2}$$

with $\alpha + 3(\beta + \gamma) + \delta = 1$. One of the solutions is then $\alpha = \beta = \gamma = \delta = 1/8$ which corresponds to the uniform solution of Section 3. The other (asymmetric) solutions are hard to find analytically because three (instead of one) cubic equations for α, β and γ have to be solved (for linear properties of fitness landscapes like eqn (B.1) see Rumschitzky, 1987). Numerical simulations

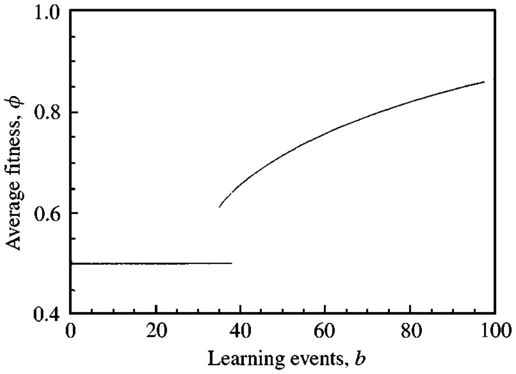

FIG. B1. The average fitness as a function of b for the system of $n = 32$ grammars consisting of $N = 5$ sentences. Each sentence can be formed in two ways.

(see Fig. B1) show that there exists a stable one-grammar solution. It resembles the X_+ solution of Section 3 in the following sense: as q becomes larger, the value of α grows and approaches unity. This means that the first grammar is the preferred one, whereas the other grammars are used less frequently. The difference is that the secondary grammars are not all used with the same frequency (i.e. $\beta \neq \gamma$, $\gamma \neq \delta$, $\beta \neq \delta$).

Next, we notice that system (2) in this case is invariant with respect to relabeling the variables in a certain way. For instance, we can set $x_2 = \alpha$, $x_1 = x_4 = x_6 = \beta$, $x_3 = x_5 = x_8 = \gamma$ and $x_7 = \delta$, and obtain exactly the same equations for α, β and γ, as we had for solution (43). There are exactly $2^3 = 8$ ways of relabeling variables which lead to the same equations, and they can all be found just by looking at the A matrix. Let us form another matrix by replacing 1 in the A matrix by α, 2/3 by β, 1/3 by γ and 0 by δ. We obtain:

$$\begin{array}{lcccccccc}
 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\
G_1: & \alpha & \beta & \beta & \gamma & \beta & \gamma & \gamma & \delta \\
G_2: & \beta & \alpha & \gamma & \beta & \gamma & \beta & \delta & \gamma \\
G_3: & \beta & \gamma & \alpha & \beta & \gamma & \delta & \beta & \gamma \\
G_4: & \gamma & \beta & \beta & \alpha & \delta & \gamma & \gamma & \beta \\
G_5: & \beta & \gamma & \gamma & \delta & \alpha & \beta & \beta & \gamma \\
G_6: & \gamma & \beta & \delta & \gamma & \beta & \alpha & \gamma & \beta \\
G_7: & \gamma & \delta & \beta & \gamma & \beta & \gamma & \alpha & \beta \\
G_8: & \delta & \gamma & \gamma & \beta & \gamma & \beta & \beta & \alpha
\end{array} \tag{B.3}$$

The first row of this matrix gives solution (B.2), where grammar G_1 is the preferred one. Each subsequent row gives another possible solution of system (2). Obviously, this is a consequence of symmetry of matrix A. Such symmetry reflects the fact that all grammars in this simple model are equivalent. Each of them shares 2/3 of its sentences with 3 other grammars, 1/3 sentences with three different grammars and does not intersect with the last grammar. Each of the grammars can become preferred, and leads to the same value of the total fitness.

The above argument can be easily extended to general values of N. In the numerical simulation presented in Fig. B1 we used the value $N = 5$. All of the $n = 2^N = 32$ equivalent one-grammar solutions are represented by the upper branch. The uniform solution is the horizontal line.