Chapter 7

Population Dynamics of Grammar Acquisition

Natalia L. Komarova and Martin A. Nowak

Introduction

The most fascinating aspect of human language is grammar. Grammar is a computational system that mediates a mapping between linguistic form and meaning. Grammar is the machinery that gives rise to the unlimited expressibility of human language.

Children develop grammatical competence spontaneously without formal training. All they need is interaction with people and exposure to normal language use. The child hears a certain number of grammatical sentences and then constructs an internal representation of the rules that generate grammatical sentences. Chomsky realized that the evidence available to the child does not uniquely determine the underlying grammatical rules (Chomsky, 1965). This phenomenon is called the 'poverty of stimulus' (Wexler and Culicover, 1980). The 'paradox of language acquisition' (Jackendoff, 1997) is that children nevertheless reliably achieve correct grammatical competence. How is this possible?

As Chomsky pointed out, the child needs a pre-formed linguistic theory that can specify candidate grammars that might be compatible with the available linguistic data (Chomsky, 1965). He introduced the term Universal Grammar (UG) to denote this preformed ‘linguistic theory’, the initial pre-specification of the form of possible human grammars (Chomsky, 1972).

Hence, for language acquisition the child needs a mechanism for processing the input sentences and a search space of candidate grammars from which to choose the appropriate grammar. Chomsky’s original concept is that UG is a rule system that generates the search space. More recent views use UG to encompass both the
search space and the mechanism for evaluating input sentences. Therefore, UG has become almost synonymous with 'mechanism of language acquisition'.

The notion of an innate, genetically encoded, UG is controversial (Tomasello, 1995; Bates, 1984; Langacker, 1987; 1992). Much of the discourse, however, focuses on which specific linguistic features are innate (for example, phrase structure rules of X-bar theory, or lexical categories such as nouns and verbs) and to what extent UG is a specific syntactic module or simply uses general-purpose cognitive abilities. We do not participate in this controversy. Instead we choose a sufficiently general formulation of the process of language acquisition. Ultimately everybody agrees that human beings require some innate components for language acquisition. These innate components are what we call UG.

Ungrammatical | Set of all sentences | Grammatical
---|---|---
0 <...> 0
1
00 <...> 00
01 <...> 01
10 <...> 10
11 <...> 11
000 <...> 000
001 <...> 001
010 <...> 010
011 <...> 011
100 <...> 100
101 <...> 101
...

*Figure 7.1* Mathematically, a grammar can be seen as a rule system that divides a countably infinite number of sentences into two subsets, grammatical and ungrammatical.

Various approaches to the mathematical description of language acquisition have been formulated by Hornstein and Lightfoot (1981), Osherson, Stob and Weinstein (1986), Manzini and Wexler (1987), Lightfoot (1991), Gibson and Wexler (1994), Niyogi (1998). The sentences of all languages can be enumerated. We can say that a grammar, $G$, is a rule system that specifies which sentences are allowed and which sentences are not allowed (see Figure 7.1). Universal grammar, in turn, contains a rule system that generates a set (or a search space) of grammars, $\{G_1, G_2, \ldots, G_n\}$. These grammars can be constructed by the language learner as potential candidates for the grammar that needs to be learned. The learner cannot end up with a grammar that is not part of this search space. In this sense, UG contains the possibility to learn all human languages (and many more).

Figure 7.2 illustrates this process of language acquisition. The learner has a mechanism to evaluate input sentences and to choose one of the candidate grammars that are contained in his search space.

More generally, it is also possible to imagine that UG generates infinitely many candidate grammars, $\{G_1, G_2, \ldots\}$. In this case, the learning task can be solved if UG
also contains a prior probability distribution on the set of all grammars. This prior distribution biases the learner toward grammars that are expected to be more likely than others. A special case of a prior distribution is one where a finite number of grammars are expected with equal probability and all other grammars are expected with zero probability, which is equivalent to a finite search space.

**Universal Grammar**

![Diagram of Universal Grammar](image)

**Figure 7.2** Universal grammar specifies the search space of candidate grammars and the learning procedure for evaluating input sentences. The basic idea is that the child has an innate expectation of grammar (for example a finite number of candidate grammars) and then chooses a particular candidate grammar that is compatible with the input.

A fundamental question of linguistics and cognitive science is what are the restrictions that are imposed by UG on human language. In other words, how much is innate and how much is learned in human language. In learning theory (Vapnik, 1995; Valiant, 1984) this question is studied in the context of an ideal speaker-hearer pair. The speaker uses a certain 'target grammar'. The hearer has to learn this grammar. The question is: what is the maximum size of the search space such that a specific learning mechanism will converge (after a number of input sentences, with a certain probability) to the target grammar?

In terms of language evolution, the crucial question is what makes a population of speakers converge to a coherent grammatical system. In other words, what are the conditions that UG has to fulfill for a population of individuals to evolve coherent communication? In the following, we will discuss how to address this question (see also Nowak, Komarova and Niyogi, 2001; Komarova, Niyogi, and Nowak, 2001). The material presented here is part of a larger effort to provide a mathematical formulation of the evolution of language (Hashimoto and Ikegami, 1996; Steels, 1997; Cangelosi and Parisi, 1998; Hurford, Studdert-Kennedy and Knight, 1998; Kirby, 1999; Cangelosi, 1999; Noble, 1999).
The rest of this chapter is organized as follows. In the next section we present the dynamical system, which describes the evolution of language learning by a population of individuals. In section “Coherence Threshold” we outline the main results concerning the dynamics of grammatical coherence. We derive the coherence threshold for the language system and analyze how it is related to the complexity of universal grammar. In particular, we specify how many sampling events children need to receive during the language acquisition period in order for the entire population to maintain a coherent communication system. In the Section of “Natural selection among variants of UG” we examine the evolutionary competition among different universal grammars. Conclusions are found in the final section.

The Language Dynamics Equations

In this section, we give a rigorous definition of grammar and put it in the context of population dynamics. The more technical material is moved to section “What is grammar?” and can be skipped at the first reading.

Population dynamics of learning

First of all, let us state that ‘poverty of stimulus’ has an elegant mathematical formulation known as Gold’s theorem (Gold, 1967). Suppose there is a rule that generates a subset of all integers. A person is provided with a sample of integers that are generated by the rule. After some time the person is asked to produce other integers that are compatible with the rule. Gold’s theorem states that this task cannot be solved. Any finite number of sample integers is not enough to determine uniquely the underlying rule. The person can only solve the task if she had a preformed expectation determining which rules are possible (or likely) and which are not. The sample integers correspond to the sentences presented to the child, the rule corresponds to the grammar used by the parents (or other speakers). The preformed expectation is universal grammar. Hence, in this sense ‘poverty of stimulus’ and the necessity of an innate universal grammar are not controversial issues, but mathematical facts.

Imagine a group of individuals that all have the same UG, given by a finite search space of candidate grammars, $G_1, \ldots, G_m$, and a learning mechanism for evaluating input sentences. Let us specify the similarity between grammars by introducing the numbers $s_{ij}$ which denote the probability that a speaker who uses $G_i$ will say a sentence that is compatible with $G_j$; the rigorous way to compute the similarity matrix is given in the next section.

We assume there is a reward for mutual understanding. The payoff for someone who uses $G_i$ and communicates with someone who uses $G_j$ is given by

$$F(G_i, G_j) = (s_{ij} + s_{ji})/2$$

This is simply the average taken over the two situations when $G_i$ talks to $G_j$ and when $G_j$ talks to $G_i$. 
Denote by $x_i$ the relative abundance of individuals who use grammar $G_i$. Assume that everybody talks to everybody else with equal probability. Therefore, the average payoff for all those individuals who use grammar $G_i$ is given by

$$f_i = \sum_{j=1}^{n} x_j F(G_i, G_j).$$

We assume that the payoff derived from communication contributes to biological fitness; individuals leave offspring proportional to their payoff. These offspring inherit the UG of their parents and learn (possibly with mistakes) their grammar. In the setting of a language game with no learning mistakes, the function $F(G_i, G_j)$ defines the equilibrium states of the system. It is shown by Komarova and Niyogi (2001) that non-ambiguous languages (perhaps with some isolated synonym-like components) are the only evolutionarily stable states of the system.

To proceed, we need to include the process of language acquisition. Children receive language input (sample sentences) from their parents and develop their own grammar. In the most general approach, we do not specify a particular learning mechanism but introduce the stochastic matrix, $\{q_{ij}\}$, whose elements, $q_{ij}$, denote the probability that a child born to an individual using $G_i$ will develop $G_j$. (For simplicity, we assume here that each child receives input from one parent. Models that allow input from several individuals are also possible.) The probability that a child will develop $G_i$ if the parent uses $G_i$ is given by $q_{ii}$. The quantities $q_{ii}$ measure the accuracy of grammar acquisition. If $q_{ii} = 1$ for all $i$ (i.e., $\{q_{ij}\}$ is an identity matrix), then grammar acquisition is perfect for all candidate grammars. Some particular learning mechanisms that define the matrix $\{q_{ij}\}$ will be considered in the following section.

The population dynamics of grammar evolution are then given by the following system of ordinary differential equations, which we call the language dynamics equations:

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} f_i q_{ij} x_j - \phi x_i, \quad j = 1, \ldots, n. \tag{1}$$

The term $-\phi x_j$ ensures that the total population size remains constant: the sum over the relative abundances, $\sum x_i$, is 1 at all times. The variable $\phi = \sum_i f_i x_i$ denotes the average fitness or grammatical coherence of the population. The grammatical coherence is given by the probability that a randomly chosen sentence of one person is understood by another person. It is a measure for successful communication in a population. If $\phi = 1$, all sentences are understood and communication is perfect. In general, $\phi$ is a number between 0 and 1.

The language dynamics equation is reminiscent of the quasispecies equation of molecular evolution (Eigen and Schuster, 1979), but has frequency-dependent fitness values: the quantities $f_i$ depend on the relative abundances $x_1, \ldots, x_n$. In the limit of perfectly accurate language acquisition, $q_{ii} = 1$, we recover the replicator equation of evolutionary game theory (Hofbauer and Sigmund, 1998). Thus, our model provides a connection between two of the most fundamental equations of evolutionary biology.
What is grammar?

In this section we give a rigorous definition of grammar and specify ways to calculate the similarity matrix, \( s_{ij} \), and the payoff function, \( F(G_i, G_j) \).

Consider a finite alphabet, that is a finite set of symbols, for example, \( \{a,b,c\} \). Consider all possible strings (sentences) that can be formed with these symbols, \( \{a, b, c, aa, ab, ac, ba, ...\} \). There are a countably infinite number of such strings. Let us call the set of symbols \( \Sigma_i \) and the set of all strings \( \Sigma_i^* \). A grammar \( G_i \) is a rule system that divides the set of all strings into two subsets. One subset (the set \( G_i \)) consists of "grammatical" strings, the other subset, \( \Sigma_i^* \setminus G_i \), of "ungrammatical" strings. This is illustrated in Figure 7.1.

This framework needs to be extended if we want to define a search space of candidate grammars. We want to calculate the probability, \( s_{ij} \), that a speaker using grammar \( G_i \) says a sentence that is compatible with another grammar, \( G_j \). In this context, a grammar \( G_i \) induces a measure (or probability distribution), \( \mu_i \), on the set of strings, \( \Sigma_i^* \). The measure specifies the probabilities with which a speaker uses particular sentences. Clearly, \( \mu_i(\Sigma_i^* \setminus G_i) = 0 \) and \( \mu_i(G_i) = 1 \). The set \( G_i \) is the support of the function \( \mu_i \). For the similarity function, \( s_{ij} \), we then have \( s_{ij} = \mu_i(G_i \cap G_j) \).

So far, we have a purely syntactic formulation of grammar, but it is sometimes argued that a grammar should not only specify which sentences are correct and which are not, but also indicate what they mean. Hence, more generally, a grammar should mediate a mapping between form and meaning. Mathematically, this can be described in the following way.

Similar to the syntactic alphabet, \( \Sigma_i \), we introduce a semantic alphabet, \( \Sigma_s \), and assume that we can enumerate all possible meanings by the set of all semantic strings, \( \Sigma_s^* \). Therefore, \( \Sigma_i^* \) is the set of all possible linguistic expressions and \( \Sigma_s^* \) is the set of all possible meanings. A grammar, \( G_i \), generates a subset of \( \Sigma_i^* \times \Sigma_s^* \), which is an infinite set of sentence-meaning pairs. In this case, \( G_i \) is specified by a measure \( \mu_i \) on \( \Sigma_i^* \times \Sigma_s^* \). As before, we can define \( s_{ij} = \mu_i(G_i \cap G_j) \) to be simply the proportion of sentence meaning pairs that \( G_i \) and \( G_j \) have in common. Hence, \( s_{ij} \) is the probability that a user of \( G_i \) produces an utterance that a user of \( G_j \) can understand.

Finally we introduce a generalization that allows us to define the fitness of grammars. Each user of \( G_i \) is characterized by an encoding matrix \( P \) and a decoding matrix \( Q \). Here, \( p_{ki} = \mu(s_k, m_i)/\Sigma_i \mu(s_k, m_i) = \mu(s_k | m_i) \) which is simply the probability of using the expression \( s_k \) to convey the meaning \( m_i \). Similarly, \( q_{kl} = \mu(s_k, m_l)/\Sigma_l \mu(s_k, m_l) = \mu(m_l | s_k) \) is the probability of interpreting the expression \( s_k \) to mean \( m_l \). The need to communicate meanings is related to events in the shared world of the linguistic community. Therefore, one can define a measure \( \sigma \) on the set of possible meanings (\( \Sigma_s^* \)) that speakers and hearers might wish to communicate with each other. Given this, we can define \( s_{ij} = \text{tr}(P^{(i)} \Lambda(Q^{(j)})^T) \) (where \( \Lambda \) is a diagonal matrix such that \( \Lambda_{ii} = \sigma(m_i) \)). This is simply the probability that an event occurs and is successfully communicated from a user of \( G_i \) to a user of \( G_j \).
Note that $F(G_i, G_j)$ is the probability that users of $G_i$ will have a successful communication with each other. Communication might break down in one of two ways (i) poverty: an event happens whose meaning cannot be encoded by $G_i$, and (ii) ambiguity: an event happens whose meaning has an ambiguous encoding in $G_i$ leading to a possibility of misunderstanding. $F(G_i, G_j)$ is a number between 0 and 1 and denotes the fitness of $G_i$. Maximum fitness, $F(G_i, G_j) = 1$, is achieved by grammars that can express every possible meaning (zero poverty) and have no ambiguities.

**Coherence Threshold**

The behavior of Equation (1) can be roughly described as follows. In general, the system admits multiple (stable and unstable) equilibria. For low accuracy of grammar acquisition (low values of $q_{ij}$, when the matrix $\{q_{ij}\}$ is far from identity), all grammars, $G_i$, occur with roughly equal abundance. There is no predominating grammar in the population, and the grammatical coherence is low. As the accuracy of grammar acquisition increases (i.e., the matrix $\{q_{ij}\}$ gets closer to identity), equilibrium solutions may arise where a particular grammar is more abundant than all other grammars. We will refer to such solutions as one-grammar equilibria, or one-grammar solutions. A coherent communication system emerges. This means that if the accuracy of learning is sufficiently high, the population will converge to a stable equilibrium with one dominant grammar. In the presence of multiple one-grammar solutions, the choice of the stable equilibrium depends on the initial conditions.

The accuracy of language acquisition (the closeness of the matrix $\{q_{ij}\}$ to identity) defines the level of coherence in the population. The accuracy of language acquisition, in its turn, depends on the following factors:

- the search space, $G_1, \ldots, G_n$,
- the learning mechanism,
- the number of learning examples, $N$, available to the child during the language acquisition stage.

Obviously, the less restricted the search space of candidate grammars is, the harder it is to learn a particular grammar. Depending on the specific values of $s_{ij}$, some grammars may be much harder to learn than others. For example, if a speaker using $G_i$ has high probabilities formulating sentences that are compatible with many other grammars ($s_{ij}$ close to 1 for many different $j$) then $G_i$ will be hard to learn. In the limit $s_{ij} = 1$, $G_i$ is considered unlearnable, because no sentence can refute the hypothesis that the speaker uses $G_i$. Also, an inefficient learning mechanism or one that evaluates only a small number of input sentences will lead to a low accuracy and hence prevent the emergence of grammatical coherence. We can therefore ask the crucial question:

*Which properties must UG have such that a predominating grammar will evolve in a population of speakers?*
In other words, which UO can induce grammatical coherence in a population? We will start by presenting two examples of the search space and then proceed by looking at specific learning algorithms.

Figure 7.3 The bifurcation diagram for fully symmetrical systems, Example 1. Here, $s = 0.3$ and $n = 30$. At $q = q_c$, $n$ identical one-grammar solutions appear. At $q = \tilde{q}_c$, the uniform (low coherence) solution becomes unstable.

**Example 1: Fully symmetrical systems**

Let us impose the following symmetry condition on the similarity matrix: $s_{ij} = s$ for all $i \neq j$, where $s$ is some constant, $0 < s < 1$. We will further assume that the $\{q_{ij}\}$ matrix in this case is also symmetrical: we have $q_{ii} = q$ for all $i$, and $q_{ij} = (1-q)/(n-1)$. (Later on we will see that this assumption holds for some reasonable learning mechanisms.) When $q$ is small (low learning accuracy), there is only one stable fixed point in the system, which we call the uniform equilibrium: all the grammars are represented in the population and have an equal abundance, and the coherence is low, see Figure 7.3. As $q$ increases, a bifurcation occurs where $n$ identical one-grammar equilibria appear, each of them corresponds to a particular dominant grammar. It is possible to calculate the error threshold, i.e., the value of $q$, $q_c$, such that for $q > q_c$ the one-grammar solutions exist and are stable. It is given by Komarova et al. (2001):

$$q_c = \frac{2 \sqrt{s}}{1 + \sqrt{s}} + O(1/n)$$

What is important, is that this value tends to a constant for large values of $n$, the size of UG. This means that no matter how large the search space is, there is a fixed threshold value for the accuracy of learning. We will refer to this interesting property as universality of universal grammar.
Example 2: Random similarity matrices

In this model, let us assume that the coefficients $s_{ij}$ are independent random numbers drawn from some distribution between zero and one. In this case, we have the following results (Komarova, 2001). If the matrix $s_{ij}$ is diagonally dominant in the sense that $s_{ii} > 1/2(s_{ij} + s_{ji})$ for all $j \neq i$, and if the matrix $\{q_{ij}\}$ is close enough to identity, then there exist exactly $n$ stable one-grammar solutions in the system. In other words, if the accuracy of learning is high enough, then the population may find itself speaking any one of $n$ grammars of UG (which one it is depends on the initial condition). The condition of diagonal dominance can be easily interpreted: we require that each of the $n$ grammars understands itself better than it understands any other grammar, which is a very natural assumption.

![Bifurcation Diagram](image)

**Figure 7.4** The bifurcation diagram for a system with a random similarity matrix, Example 2. Here $n=20$. At $N=N_c$, the first one-grammar solution appears. The learning accuracy matrix, $\{q_{ij}\}$, is calculated according to the memoryless learner algorithm, and depends on $N$, the number of learning events. At $N = N_c$ the low coherence solution equivalent to the uniform solution of the fully symmetrical system, disappears.

On the other hand, if for some $k$ and $m$ the diagonal dominance is violated, so that $s_{kk} < 1/2(s_{km} + s_{mk})$ and $s_{mm} < 1/2(s_{km} + s_{mk})$, then no matter how high the learning accuracy is, grammars $G_k$ and $G_k$ cannot become dominant. Instead, a mixture of the two grammars is possible as a stable two-grammar equilibrium. Note that such situations are in principle possible: highly ambiguous grammars might have a higher communicative efficiency when communicating with each other than they do with themselves (see Komarova and Niyogi, 2001).

An example of a bifurcation diagram for a random system is presented in Figure 7.4. There, the coefficients $s_{ij}$ with $i \neq j$ are random uniformly distributed.
numbers and $s_{ii}=1$. The matrix $\{q_{ij}\}$ is parameterized with $N$, the number of learning events, according to the memoryless learner algorithm, see below. As $N$ grows, the matrix $\{q_{ij}\}$ tends to identity. We observe that the error threshold phenomenon is also present in the random system; for complete analytical results refer to Komarova (2001). One particular feature is that if we assume that the matrix $\{q_{ij}\}$ is symmetrical (as defined in Example 1), then the error threshold does not depend on the size of the system. This means that the universality property holds in this case just like it does for fully symmetrical systems. There are preliminary results that the universality property also holds for a more general class of matrices $\{q_{ij}\}$.

Now let us look at some learning mechanisms which, given the search space, define the dependence of $\{q_{ij}\}$ on $N$, the number of learning events. We present results for two learning mechanisms that represent reasonable boundaries for the actual, unknown learning mechanism employed by humans.

The memoryless learning algorithm, a favorite with learning theorists, makes little demand on the cognitive abilities of the learner. It describes the interaction between a teacher and a learner. (The ‘teacher’ can be one or several individuals or the whole population.) The learner starts with a randomly chosen hypothesis (say $G_i$) and stays with this hypothesis as long as the teacher’s sentences are compatible with this hypothesis. If a sentence arrives that is not compatible, the learner will at random pick another candidate grammar from his search space. The process stops after $N$ sentences. The algorithm is called ‘memoryless’, because the learner does not remember any of the previous sentences nor which hypotheses have already been rejected. The algorithm works, primarily because once it has the correct hypothesis it will not change anymore (this is incidentally the definition of so called ‘consistent learners’).

The other extreme is a batch learner (resembling Jorge Louis Borges’ man with infinite memory). The batch learner memorizes all the $N$ sentences and at the end chooses the candidate grammar that is most compatible with the input.

**Fully symmetrical systems**

For this example, the memoryless learner algorithm leads to a symmetrical $\{q_{ij}\}$ matrix, as was assumed before. We have $q=1-\left(1-\frac{1-s}{n-1}\right)^{n-1}$, For the memoryless learner, we can show that one-grammar equilibria exist if the number of input sentences, $N$, exceeds a constant times the number of candidate grammars,

$$N > C_1 n,$$

where $C_1=\log \frac{1+\sqrt{s}}{1-\sqrt{s}}/(1-s)$. This defines the value of $N$ corresponding to the onset of high grammatical coherence.

For a batch learner (under some natural assumptions on the configuration of grammars in the search space) we have $q=\frac{1-(1-s)^n}{s^{n}}$, and in order to maintain coherence, the number of input sentences has to exceed a constant times the logarithm of the number of candidate grammars,
\[ N > C_2 \log n, \quad (3) \]

where \( C_2 = \log^{-1} (1/s) \). Inequalities (2) and (3) define a coherence threshold, which limits the size of the search space relative to the amount of input available to the child. A UG that does not fulfill the coherence threshold does not lead to a stable, predominating grammar in a population.

**Random similarity matrices**

Let us first consider the following problem. A learner receives \( N \) sentences from the teacher and uses the memoryless learning mechanism to determine which grammar is the "correct" one. How many sample sentences does it typically take for the learner to guess the answer with the probability \( 1 - \Delta \)? The following results have been obtained by Komarova and Rivin (2001a; 2001b). Let us assume that \( s_{ii} = 1 \) for all \( i \). Then if the distribution of similarity coefficients, \( s_{ij} \), is uniform for \( i \neq j \), then it takes

\[ N > c_1 n \log n \quad (4) \]

sample sentences for the learner to converge to the truth, where \( c_1 \sim \log \Delta \). If the similarity coefficients have non-uniform distribution which favors small values, and values close to 1 almost never occur, then we have \( N > c_1 n \), i.e., learning is faster. If, on the other hand, a large fraction of grammars are very similar to each other, then it takes longer to learn: \( N > c_1 n^{1+\delta} \) for some positive \( \delta \) (the details are given in Komarova and Rivin (2001a; 2001b)).

For a batch learner, in the case of a uniform distribution of \( s_{ij} \), Rivin (2001) has shown that it takes

\[ N > c_2 n \quad (5) \]

steps for an individual learner to be on target with a finite probability. The question then arises: how can these results for an individual learner be applied to the problem of population learning? There are no simple formulas in this case which would relate the accuracy of learning to the parameters of the system, as there were in the fully symmetrical case. Some perturbation methods must be used in order to obtain the stability conditions for one-grammar solutions. This work is still in progress. A preliminary result can be described as follows: if a learning algorithm satisfies the universality property, then estimates obtained for an individual learner will hold for the entire population. Numerical simulations suggest that the memoryless learner algorithm and the batch algorithm satisfy the universality property, and therefore results (4) and (5) are true for a population of learners. For the precise statement, refer to Komarova (2001).

To summarize so far, our mathematical modeling of the population dynamics of language acquisition suggests the following:

- Depending on the search space, it may be easier or harder to maintain coherence in the population. It is harder if a lot of the candidate grammars overlap and are difficult to distinguish. It is easier if they are well separated.
- For any reasonable search space, the batch learner will perform better than the memoryless learner. The learning mechanism used by humans will do
Simulating the Evolution of Language

good than the memoryless learner and worse than the batch learner. Hence it will have a coherence threshold somewhere between the bounds given above.

Natural Selection among Variants of Universal Grammar

So far we have assumed that all individuals have the same UG. Studying the biological evolution of UG, we need variation in UG and a system that describes natural selection among variants of UG.

Denote by $x_i$ the fraction of individuals who use $G_i$ of universal grammar $U_1$; denote by $y_i$ the fraction of individuals who use $G_i$ of universal grammar $U_2$. $U_1$ and $U_2$ contain, respectively, $n_1$ and $n_2$ candidate grammars. Some of the candidate grammars can be part of both universal grammars. The universal grammars, $U_1$ and $U_2$, can also differ in the number of sample sentences, $N_1$ and $N_2$, which are being considered. Therefore, we have to take into account the rate of producing offspring with grammatical communication; this rate is given by the declining function $r(N)$. The dynamics are described by

$$x_i = r(N_1) \sum_{j=1}^{n_1} x_j f_j^{(1)} q_j^{(1)} - \phi x_i, \quad i = 1, \ldots, n_1,$$

$$y_i = r(N_2) \sum_{j=1}^{n_2} x_j f_j^{(2)} q_j^{(2)} - \phi y_i, \quad i = 1, \ldots, n_2.$$

For $m=1, 2$, we have

$$f_i^{(m)} = \sum_{j=1}^{n_1} x_j F(G_i^{(m)}, G_j^{(1)}) + \sum_{j=1}^{n_2} x_j F(G_i^{(m)}, G_j^{(2)})$$

Grammatical coherence is given by

$$\phi = \sum_{i=1}^{n_1} f_i^{(1)} x_i r(N_1) + \sum_{i=1}^{n_2} f_i^{(2)} x_i r(N_2).$$

The superscripts 1,2 refer to $U_1$ and $U_2$ respectively.

Now, it is possible to compare the stability of, say, universal grammar $U_1$ against the invasion of $U_2$, simply by performing a linear stability analysis of the solution where $x_1, \ldots, x_{n_1}$ is the one-grammar equilibrium for the users of $U_1$, and $y_1, \ldots, y_{n_1}$ for the users of $U_2$. This framework can also be extended to more than two universal grammars, thus giving conditions for a universal grammar, $U$, being stable against invasion of any other UG (in a given class). Here we present some applications of this method.

First, let us consider universal grammars with the same search space and the same learning procedure, the only difference being the number of input sentences,
Population Dynamics of Grammar Acquisition

N (Komarova and Nowak, 2001). This quantity is proportional to the length of the learning period. We find that natural selection leads to intermediate values of N. For small N, the accuracy of learning the correct grammar is too low. For large N, the learning process takes too long (and thus the rate of producing children that have acquired the correct grammar is too low). This observation can explain why there is a limited language acquisition period in humans.

Second, consider universal grammars, $U_1$ and $U_2$, which differ in the size of their search space, $n$, but have the same learning mechanism and the same value of N. In general, there is selection pressure to reduce $n$. Only if $n$ is below the coherence threshold, can the universal grammar induce grammatical communication. In addition, the smaller $n$, the larger is the accuracy of grammar acquisition. There can, however, also be selection for larger $n$: suppose universal grammar $U_1$ is larger than $U_2$ (that is $n_1 > n_2$). If all individuals use a grammar, $G_1$, that is in both $U_1$ and $U_2$, then $U_2$ is selected. Now imagine that someone invents a new advantageous grammatical concept, which leads to a modified grammar $G_2$ that is in $U_1$, but not in $U_2$. In this case, the larger universal grammar is favored. Hence there is selection both for reducing the size of the search space and for remaining open minded to be able to learn new concepts. For maximum flexibility, we expect search spaces to be as large as possible but still below the coherence threshold.

An interesting extension of the above model is obtained by assuming that UG is only very roughly defined by our genes. Randomness during the developmental process could give rise to variation in neuronal patterns in the brain and consequently to variation in UG. Hence it might be a reasonable assumption that individuals have slightly different UGs. Each individual could have a personal ‘universal’ grammar. An interesting question is how similar these UGs have to be such that a population achieves grammatical coherence. In this case, there is again selection for maintaining a large search space of candidate grammars, since the target grammar should be contained in each of the UGs.

Conclusions and Discussion

We have formulated a mathematical theory for the population dynamics of grammar acquisition. The key result here is a ‘coherence threshold’ that relates the maximum complexity of the search space to the amount of linguistic input available to the child and the performance of the learning procedure. The coherence threshold represents an evolutionary stability condition for the language acquisition device: only a universal grammar that operates above the coherence threshold can induce and maintain coherent communication in a population.

There are many ways in which the framework presented here can be extended. For instance, the language dynamics equations describe deterministic dynamics for a large population size. Smaller population sizes can play a role if we consider stochastic language dynamics. Computer simulations suggest that the equilibrium solutions of the deterministic system correspond to meta-stable states. Individual grammars will dominate for some time and then be replaced by other grammars. Such transitions are more likely to occur between similar grammars.
In a small population, the requirements imposed on UG are also slightly stronger. Grammatical coherence in a population will require a larger number of input sentences or smaller search spaces. A detailed mathematical study of the stochastic dynamics of our system is still outstanding.

Individual candidate grammars, $G_i$, can also differ in their performance. Some grammars can be less ambiguous or describe more concepts than others. In such a context, the language dynamics equation can describe a cultural evolutionary optimization of grammar within the space of grammars generated by UG (Niyogi and Berwick, 1996; 1997). It also provides a general framework for studying the dynamics of grammar change in the context of historical linguistics, where a direct comparison is possible between theoretically obtained results and the large corpus of available linguistic data.

---

**Acknowledgments**

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 NSF is gratefully acknowledged.

---

**References**


Gold EM (1967) Language identification in the limit. Information and Control, 10: 447-474


Kirby S (1999) Syntax out of learning: The cultural evolution of structured communication in a population of induction algorithms In: Floreano D, Nicoud JD, Mondada F (eds)
Population Dynamics of Grammar Acquisition

Proceedings of ECAL99 the Fifth European Conference on Artificial Life (Lecture Notes in Artificial Intelligence). Springer-Verlag, Berlin, pp 694-703
Komarova NL, Niyogi P (2001) Optimizing the mutual intelligibility of linguistic agents in a shared world. To be submitted to *Artificial Intelligence*
Rivin I (2001) Unpublished
Simulating the Evolution of Language
# Table of Contents

List of Contributors ........................................................................................................... xi

**PART I  INTRODUCTION** ............................................................................................................. 1

1 Computer Simulation: A New Scientific Approach to the Study of Language Evolution
   *Angelo Cangelosi and Domenico Parisi* .............................................................................. 3

2 An Introduction to Methods for Simulating the Evolution of Language
   *Huck Turner* ......................................................................................................................... 29

**PART II  EVOLUTION OF SIGNALING SYSTEMS** ........................................................................ 51

3 Adaptive Factors in the Evolution of Signaling Systems
   *Jason Noble, Ezequiel A. Di Paolo and Seth Bullock* ...................................................... 53

4 Evolving Sound Systems
   *Bart de Boer* ...................................................................................................................... 79

5 The Evolution of Dialect Diversity
   *Daniel Livingstone* ........................................................................................................... 99

**PART III  EVOLUTION OF SYNTAX** ......................................................................................... 119

6 The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model
   *Simon Kirby and James R. Hurford* ................................................................................... 121

7 Population Dynamics of Grammar Acquisition
   *Natalia L. Komarova and Martin A. Nowak* ..................................................................... 149

8 The Role of Sequential Learning in Language Evolution: Computational and Experimental Studies
   *Morten H. Christiansen, Rick A.C. Dale, Michelle R. Ellefson and Christopher M. Conway* .................................................................................................................................................................................. 165
PART IV GROUNDING OF LANGUAGE .......................................................... 189

9 Symbol Grounding and the Symbolic Theft Hypothesis
   Angelo Cangelosi, Alberto Greco and Stevan Harnad ........................... 191

10 Grounding Symbols through Evolutionary Language Games
   Luc Steels ......................................................................................... 211

PART V BEHAVIORAL AND NEURAL FACTORS ...................................... 227

11 Grounding the Mirror System Hypothesis for the Evolution of the Language-ready Brain
   Michael A. Arbib ............................................................................... 229

12 A Unified Simulation Scenario for Language Development, Evolution, and Historical Change
   Domenico Parisi and Angelo Cangelosi ............................................. 255

PART VI AUTO-ORGANIZATION AND DYNAMIC FACTORS ...................... 277

13 Auto-organization and Emergence of Shared Language Structure
   Edwin Hutchins and Brian Hazlehurst .............................................. 279

14 The Constructive Approach to the Dynamic View of Language
   Takashi Hashimoto .......................................................................... 307

PART VII CONCLUSION ........................................................................... 325

15 Some Facts about Primate (including Human) Communication and Social Learning
   Michael Tomasello ........................................................................... 327

Author Index ......................................................................................... 341

Subject Index ......................................................................................... 347