Population dynamics of human language: a complex system

Natalia L. Komarova

Department of Mathematics, University of California - Irvine, Irvine, CA 92697; email: komarova@math.uci.edu

1 Introduction

In the course of natural history, Evolution has come up with several great innovations, such as nucleic acids, proteins, cells, chromosomes, multi-cellular organisms, the nervous system.... The last “invention” which truly revolutionized the very rules of evolution is language. It gives us an unprecedented possibility to transmit information from generation to generation not by the “traditional” means of a genetic code, but by talking. This new mode of cross-generational information transfer has given rise to the so-called “cultural evolution”. It is responsible for a big part of being “human”. It is shaping the history and changing the rules of biology. Without exaggeration, it is one of the most fascinating traits of Homo Sapiens.

The study of language and grammar dates back to classical India and Greece. In the 18th century, the “discovery” of Indo-European led to the surprising realization that very different languages can be related to each other, which initiated the field of historical linguistics. Formal language theory emerged only in the 20th century (Chomsky, 1956, 1957; Harrison, 1978): the main goals are to describe the rules that a speaker uses to generate linguistic forms (descriptive adequacy) and to explain how language competence emerges in the human brain (explanatory adequacy). These efforts were supported by advances in the mathematical and computational analysis of the process of language acquisition, a field that became known as learning theory. Currently there are increasing attempts to bring linguistic inquiry in contact with various disciplines of biology, including neurobiology (Deacon, 1997; Vargha-Khadem et al., 1998), animal behavior (Dunbar, 1996; Hauser, 1996; Fitch, 2000), evolution (Lieberman, 1984; Pinker and Bloom, 1990; Bickerton, 1990; Lieberman, 1991; Hawkins and Gell-Mann, 1992; Batali, 1994; Maynard Smith and Szathmary, 1995; Aitchinson, 1996; Hurford et al., 1998;
Jackendoff, 1999; Knight et al., 2000) and genetics (Gopnik and Crago, 1991; Lai et al., 2001). The new aim is to study language as a product of evolution and as the extended phenotype of a species of primates.

The past decade has seen an explosion of interest in computational aspects of the evolution of language (Cangelosi and Parisi, 2001; Christiansen and Kirby, 2003). There is a lot of effort, across a wide range of disciplines, to answer such questions as: why is language the way it is and how did it become this way? Among various approaches, it has been suggested to view language as a complex adaptive system (Steels, 2000). Levin (2002) defined a general complex adaptive system by the following properties:

(i) they consist of a number of different components,

(ii) the components interact with each other with some degree of localization,

(iii) an autonomous process exists that uses the outcomes of these interactions to select a subset of components for replication and/or enhancement.

It is property (iii) that is a signature of “biology” in a complex adaptive system. It includes replication (which implies some degree of variability) and Darwinian selection. A mathematical problem posed by such systems is to find the outcome (or, more generally, describe the dynamics) of a competition where the set of players is changing, depending on the current state of affairs. New players come in, and their “strategies” (or properties) are drawn from a huge set of possibilities, $S$. The main idea in this approach to language evolution is as follows. There is a population of individuals (neural networks (Oliphant, 1999; Smith, 2002), agents (Steels and Kaplan, 1998; Steels, 2001), organisms in a foraging environment (Cangelosi, 2001)) who communicate with each other. Each individual is characterized by a set of parameters which define its phenotype. This usually amounts to the individual’s ability to speak, its vocabulary, its ability to learn and other characteristics important for communication (and sometimes other features like the life-span or the onset of the reproductive age (Hurford and Kirby, 1998)). The individuals communicate with each other, and the result of the communication is assessed in some way or another. Rounds of communication are followed by rounds of “update”, which may mimic biological reproduction (an individual is replaced by its offspring), or leaning (the individual’s vocabulary or grammatical rules

2
are changed/updated). Various numerical techniques are used to model the dynamics of reproducing and learning individuals, such as genetic algorithms. The initial condition usually assumes no common communication system in the population. After a number of rounds of update/replication, the state of communication ability is evaluated again. The questions which are often addressed are these:

Under what circumstances does a common communication system arise in a population of interacting individuals? What are the conditions under which such a communication system can be maintained?

2 What is Universal Grammar and why do we need it?

Learning is inductive inference. The learner is presented with data and has to infer the rules that generate these data. The difference between ‘learning’ and ‘memorization’ is the ability to generalize beyond one’s own experience to novel circumstances. In the context of language, the child learner will generalize to novel sentences never heard before. Any child can produce and understand sentences that are not part of his previous linguistic experience.

Children develop grammatical competence spontaneously without formal training. All they need is interaction with people and exposure to normal language use. In other words, the child hears a certain amount of grammatical sentences and then constructs an internal representation of the rules that generate grammatical sentences. Chomsky pointed out that the evidence available to the child does not uniquely determine the underlying grammatical rules (Chomsky, 1965, 1972). This phenomenon is called the ‘poverty of stimulus’ (Wexler and Culicover, 1980). The ‘paradox of language acquisition’ is that children nevertheless reliably achieve correct grammatical competence (Jackendoft, 1997, 2001). How is this possible?

The proposed solution of the paradox is that children learn the correct grammar by choosing from a restricted set of candidate grammars. The structure of this restricted set is ‘universal grammar’ (UG). Mathematical learning theory proves the “necessity” of UG. Discovering properties of UG and particular human learning algorithms requires the empirical study of neurobiological and cognitive functions of the human brain involved in language
Figure 1: 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.

acquisition. Some aspects of UG, however, might be unveiled by studying common features of existing human languages. This has been a major goal of linguistic research during the last decades.

In our modeling approach we use the concept and some properties of UG to formulate the mathematical theory of language evolution. We assume that UG 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 1 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.

Next we describe how the concept of UG enters our model of language evolution.
3 A mathematical formulation of language evolution

Our particular approach is different from many others in that we use mathematical, analytical tools to address questions of language origins and evolution (Komarova et al., 2001; Nowak et al., 2001; Nowak and Komarova, 2001; Komarova and Nowak, 2001a,b; Nowak et al., 2002; Komarova and Nowak, 2003). We assume that each individual has UG which allows to learn any language in a (finite but large) set \( S = \{G_1, \ldots, G_n\} \).

In classical learning theory, an isolated teacher-learner pair is usually considered. There is a collection of concepts (grammars), \( G_1, \ldots, G_n \), and “words” (or sample sentences, if we talk about learning a grammar) which refer to these concepts, sometimes ambiguously. The teacher generates a stream of words, referring to, say, concept \( G_2 \). This is not known to the student, but he must learn by guessing some concept \( G_i \) and checking for consistency with the teacher’s input. A typical question of interest is this: How quickly does a given method converge to the truth? Or, restated in the terminology of learning languages,

\[
\text{How many samples, } N_\delta, \text{ does a given learning algorithm typically need in order to learn the correct language with probability } 1 - \delta?
\]

Questions of this type for specific learning mechanisms are interesting mathematical problems, see for instance the treatment of the so-called memoryless learner (Komarova and Rivin, 2003). Next, let us imagine a population of learners, all equipped with a given learning algorithm. The question now becomes:

\[
\text{How many samples, } N_\delta, \text{ do individual learners in a population need, in order for the fraction } 1 - \delta \text{ of the population to converge to a common language?}
\]

The answer to this question will of course depend on the specifics of the population dynamics, which we describe next.

Borrowing the idea from population biology, we will define the fitness of speakers of different grammars. Let us denote by \( s_{ij} \) the probability that a speaker who uses grammar \( G_i \) formulates a sentence that is compatible with grammar \( G_j \). The matrix \( \{s_{ij}\} \) describes the pairwise similarity among the \( n \) grammars, \( 0 \leq s_{ij} \leq 1 \). We assume there is a reward for mutual understanding. The payoff for an individual using \( G_i \) communicating with
an individual using $G_j$ is given by $a_{ij} = (1/2)(s_{ij} + s_{ji})$. This is the average probability that $G_i$ generates a sentence that is parsed by $G_j$ and vice versa. We denote by $x_i$ the frequency of individuals who use grammar $G_i$; the vector $\mathbf{x} = \{x_1, \ldots, x_n\}$ is defined on the simplex,

$$\sum_{i=1}^{n} x_i = 1.$$  

The average payoff of each of these individuals is given by $\mathbf{f} = \hat{A}\mathbf{x}$, where $\hat{A} = \{a_{ij}\}$ is a symmetric matrix. Payoff translates into fitness: individuals with a higher payoff produce more offspring. Note that the fitness of individuals strongly depends on the current composition of the population; such is the nature of communication.

Another biological concept coming from the theory of Darwinian evolution is variability. The “mutation rates” are defined as follows: denote by $Q_{ij}$ the probability that a child learning from a parent with grammar $G_i$ will end up speaking grammar $G_j$. $\hat{Q} = \{Q_{ij}\}$ is a stochastic matrix (its rows sum up to one). Interestingly, the findings related to individual teacher-learner pairs can be incorporated in a natural way into the matrix $\hat{Q}$.

The last component of the model is the update rule for the evolutionary dynamics. The simplest rule is a deterministic equation, where each variable has a meaning of its ensemble average and the noise is neglected. This can be written by analogy with the well-known quasispecies equation (Eigen and Schuster, 1979), except it has a higher degree of nonlinearity (a consequence of the population-dependent fitness). We have,

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

Here $\phi = (\mathbf{x}, \mathbf{f})$ is the average fitness or grammatical coherence of the population; it is the probability that a sentence said by one person is understood by another person. This equation describes a mutation-selection process in a population of individuals of $n$ types.

### 4 Coherence threshold in population learning

Numerical simulations (Komarova et al., 2001; Nowak et al., 2001) and analytical estimates (Komarova, 2004) of equation (1) show the following trend.
Learning accuracy

Low coherence

No predominant grammar

Grammatical coherence increases

Coherence threshold

Many stable coherent equilibria

Each characterized by a predominant grammar

Learning accuracy

Figure 2: Coherence threshold. When the accuracy of individual learning is low, equation (1) admits a single low-coherence solution with no predominating grammar in the population. As the accuracy of grammar acquisition increases, however, equilibrium solutions arise where a particular grammar is more abundant than all other grammars. A coherent communication system emerges.

If the matrix $\hat{Q}$ is close to identity, there are many coexisting localized steady-state solutions (corresponding to stable fixed points). For each such solution, the majority of the population speaks one of the languages, and the grammatical coherence, $\phi$, takes values close to 1. As $\hat{Q}$ deviates far from the identity matrix (which means that there is a lot of “noise” in the system, that is, mistakes or learning are very likely), then this localization is lost and grammatical coherence becomes low, see figure 2. A particular, highly symmetrical case of this system has been analyzed by Komarova et al. (2001) and Mitchener (2003). It was found that the low-coherence delocalized solution undergoes a transcritical bifurcation for the value $\Delta Q = \|\hat{Q} - \hat{I}\|$ (where $\hat{I}$ is the identity matrix) defined by the entries of the matrix $\hat{A}$. A very interesting fact is that the threshold value of $\Delta Q$ does not depend on the dimension of the system, $n$.

A natural question is then to describe the phenomenon of the loss/gain of coherence for general matrices $\hat{A}$ and $\hat{Q}$. For instance, we can assume that the entries of the matrix $\hat{A}$ are taken from a distribution, and the matrix $\hat{Q}$ is a function of $\hat{A}$. Our results (Komarova, 2004) suggest that typically, the threshold value of $\Delta Q$ tends to a constant as $n \to \infty$, where $n$ is the size of the system. This finding can be termed universality property of Universal Grammars: For a (reasonable) class of learning algorithms (matrix $\hat{Q}$) and for any size of Universal Grammar ($n$), there is a finite coherence threshold in the system defined by the similarity of the grammars (matrix $\hat{A}$).

What is the significance of the coherence threshold for our understanding
of language evolution? Our analyses can help us obtain possible bounds on complexity of universal grammar that are compatible with Darwinian evolution. Indeed, if the space of all possible grammars is too large, learning would take too long: humans only have a limited time for learning before they become adults. This is where linguistics meets evolutionary biology: there is a selection pressure to make universal grammar smaller and easier to learn. On the other hand there is also an advantage to having a larger pool of grammars, which increase flexibility and facilitate innovation.

5 Discussion

There are two common misconceptions of language evolution. The first one represents the human language capacity as an undecomposable unit and states that its gradual evolution is impossible, because no part of it would have any function in the absence of other parts. For example, syntax could not have evolved without phonology or semantics and vice versa. The other misconception is that language evolution started from scratch some 5 million years ago, when humans and chimps diverged, and there are virtually no data about it.

Both views are fundamentally flawed. First, all complex biological systems consist of specific components, in a way that it is often hard to imagine the usefulness of individual parts in the absence of other parts. The usual task of evolutionary biology is to understand how complex systems can arise from simpler ones gradually by mutation and natural selection. In this sense, human language is no different from other complex traits. Second, it is clear that evolution did not build the human language faculty de novo in the last few million years, but used material that had evolved in other animals over a much longer time. Many animal species have sophisticated cognitive abilities in terms of understanding the world and interacting with one another. Furthermore, it is a well known trick of evolution to use existing structures for new and sometimes surprising purposes. Monkeys, for examples, appear to have brain areas similar to our language centers, but use them for controlling facial muscles and for analyzing auditory input. Evolution may have had an easy task here to reconnect these centers for human language. Hence the human language instinct is most likely not the result of a sudden moment of inspiration of evolution’s blind watchmaker, but rather the consequence of several hundred million years of ‘experimenting’ with animal cognition.
The goal of this paper was to show how methods of formal language theory, learning theory and evolutionary biology can be combined to improve our understanding of the origins and the properties of human language. 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.

Acknowledgments. Support from the Alfred P. Sloan foundation is gratefully acknowledged.

References


