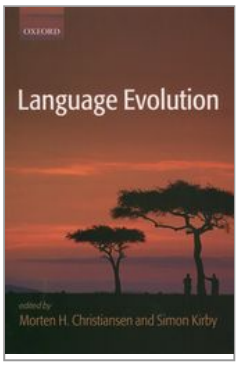

Language Evolution
Morten H. Christiansen and Simon Kirby

Print publication date: 2003
Print ISBN-13: 9780199244843
Published to Oxford Scholarship Online: January 2010
DOI: 10.1093/acprof:oso/9780199244843.001.0001

Language, Learning and Evolution

Natalia L. Komarova
Martin A. Nowak

DOI:10.1093/acprof:oso/9780199244843.003.0017

Abstract and Keywords

This chapter studies language evolution from the viewpoint of mathematical game theory. A previous investigation suggested, based on evidence from formal language theory, that innate constraints on language acquisition are a logical necessity. This chapter argues that those results do not determine whether innate constraints must be linguistic in nature. Rather, the study only demonstrates that innate constraints on learning are needed. The formal language work is combined with an evolutionary approach based on game theory in order to provide a general mathematical framework for exploring the evolution of language. Within this framework, language evolution can be studied in terms of populations of language-learning agents whose survival and ability to procreate depend on their capacity for language. The results indicate that natural selection would lend to favour systematic mechanisms for encoding grammatical knowledge. The chapter demonstrates how mathematical and computational modelling can be fruitfully applied to the study of language evolution.

Keywords: language, language evolution, learning, game theory, language acquisition, innate constraints, natural selection, computational modelling, formal language theory

Introduction

Throughout the history of life on earth, evolution has come up with several great innovations, such as nucleic acids, proteins, cells, chromosomes, multi-cellular organisms, the nervous system, and finally-language. Among these innovations, language is the only one which is (presently) confined to one species. What

exactly human language is, how we learn it, and how it evolved in its enormous complexity are some of the most fascinating questions of evolutionary biology and cognitive science.

The study of language and grammar dates back to classical India and Greece (Robins 1979). In the eighteenth 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 twentieth 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’ (Gold 1967; Vapnik and Chervonenkis 1971; 1981; Valiant 1984; Osherson et al. 1986; Vapnik 1998; Jain et al. 1998). Currently there are increasing attempts to bring linguistic inquiry into contact with various disciplines of biology, including neurobiol-ogy (Deacon 1997; Vargha-Khadem et al. 1998), animal behaviour (Smith 1977; Dunbar 1996; Hauser 1996; Fitch 2000), evolution (Lieberman 1984; **(p.318)** Brandon and Hornstein 1986; Pinker and Bloom 1990; Bickerton 1990; Lieberman 1991; Newmayer 1991; Hawkins and GellMann 1992; Batali 1994; Maynard Smith and Szathmary 1995; Aitchinson 1996; Hurford et al. 1998; Jackendoff 1999; Lightfoot 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.

There are two common misconceptions of language evolution. The first 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 five 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, so 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 example, appear to have brain areas similar to our language centres, but use them for controlling facial muscles and for analysing auditory input. Evolution may have had an easy task here to reconnect these centres 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.

We can obtain data for language evolution in two ways. We can study the evolution of cognitive abilities and communication in animals (Hauser 1996; Fitch 2000a), and we can analyse the enormous evidence provided by the existing human language instinct and its manifestation in 6,000 different languages. The data are in us, similar to genetic evolutionary history being written in our genome.

(p.319) The perspective of this chapter is 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. In the following section we discuss the key notions of 'language', grammar', and 'learning' and present rigorous definitions. We then formulate the 'paradox of language acquisition', and we show that learning theory can demonstrate in what sense Universal Grammar is a logical necessity. In the section on evolution in languages, we present a quantitative approach to questions of communication and evolution. We develop a general framework for the evolution of grammar acquisition and discuss how natural selection acts on Universal Grammar. We define grammatical coherence and find a coherence threshold which gives conditions for a population of speakers to evolve and maintain a stable language. We explore the conditions under which natural selection favours the emergence of a recursive, rule-based grammatical system. We show that our approach can be used to study problems of historical linguistics. In the concluding section we discuss cultural evolution of language as opposed to biological evolution of universal grammar, and come up with a unified description that contains both.

Formal Language Theory and Applications

Language is a mode of communication, a crucial part of human behaviour, and a cultural object defining our social identity. There is also a fundamental aspect of human language that makes it amenable to formal analysis: linguistic structures consist of smaller units that are grouped together according to certain rules.

The combinatorial sequencing of small units into bigger structures occurs at several different levels. Phonemes are concatenated into syllables and words. Sequences of words form phrases and sentences. Most crucially, the rules for such groupings are not arbitrary and are language specific. Certain word orders

are admissible in one language but not in another. In some languages, word order is relatively free but case marking is pronounced. There are always specific rules that *generate* valid or meaningful linguistic structures. Much of modern linguistic theory proceeds from this insight. The area of mathematics and computer science called formal language theory provides a mathematical machinery for dealing with such phenomena.

(p.320) Some Important Definitions

We define an *alphabet* as a set containing a finite number of symbols. Possible alphabets for natural languages would be the set of all phonemes or the set of all words of a language. For these two choices one obtains formal languages on different levels, but the mathematical principles are the same. We can also simply consider the binary alphabet {0,1} (see Fig. 17.1).

A *sentence* is defined as a string of symbols. The set of all sentences that can be generated by the binary alphabet is given by {0,1, 00,01,10,11,000, …}. Note that there are infinitely many sentences. More precisely, there are as many sentences as there are integers. Hence, the set of all sentences is 'countable'.

A *language* is a set of sentences. Among all possible sentences, some are part of the language and some are not. A finite language contains a finite number of sentences. An infinite language contains an infinite number of sentences. There are infinitely many finite languages, as many as integers. There are infinitely many infinite languages, as many as real numbers; they are not countable. Hence, the set of all languages is not countable.

A grammar is a 'theory' of a language: it is a finite list of rules that specify the language. A grammar is normally expressed in terms of 'rewrite rules', which are of the form: a certain string can be rewritten as another string.

(p.321) Strings contain elements of the alphabet together with so-called 'non-terminals', which are place holders. After iterated application of the rewrite rules the final string will only contain symbols of the alphabet. Figs. 17.1 and 17.2 give examples of grammars.

There are infinitely many grammars, but only as many as integers: any finite list of rewrite rules can be encoded by an integer. Since there are un-

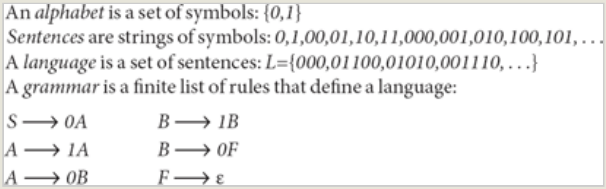

Fig. 17.1 The basic objects of formal language theory are alphabets, sentences, languages, and grammars. Grammars consist of rewrite rules: a particular string can be rewritten as another string. Such rules contain symbols of the alphabet (here: 0 and 1), and so called 'non-terminals' (here: S, A,B,F), and a null-element, ε. The grammar in this figure works as follows:

countably many languages, only a small subset of them can be described by a grammar. These languages are called ‘computable’.

Each sentence begins with the symbol S. S is rewritten as 0A. Now there are two choices: A can be rewritten as 1A or 0B. B can be rewritten as 1B or 0F. F always goes to ε. This grammar generates sentences of the form 01^m01^n0, which means every sentence begins with 0 followed by a sequence of *m* 1 s followed by a 0 followed by a sequence of n 1 s followed by 0.

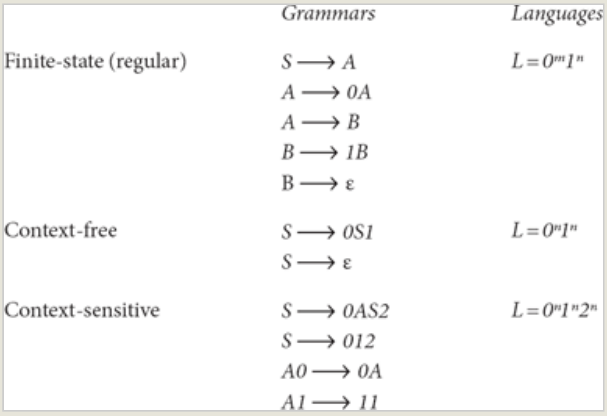

Fig. 17.2 Three grammars and their corresponding languages. Finite-state grammars have rewrite rules of the form: a single non-terminal (on the left) is rewritten as a single terminal possible followed by a non-terminal (on the right). The finite-state grammar, in this figure, generates the regular language 0^m1^n; a valid sentence is any sequence of 0s followed by any sequence of 1s. A context-free grammar admits rewrite rules of the form: a single non-terminal is rewritten as an arbitrary string of terminals and nonterminals. The context-free grammar in this figure generates the language 0^n1^n; a valid sentence is a sequence of 0s followed by the same number of 1s. There is no finite-state grammar that could generate this language. A context-sensitive grammar admits rewrite rules of the form α A β → α β γ. Here α, β, and γ are strings of

(p.322) The Chomsky Hierarchy of Languages

There is a correspondence between languages, grammars and machines. The set of all computable languages is described by 'phrase structure' grammars which are equivalent to Turing machines. A Turing machine embodies the theoretical concept of a digital computer with infinite memory (Turing 1936; 1950). For each computable language, there exists a Turing machine which can list as output all sentences of this language.

terminals and nonterminals. While α and β may be empty, γ must be non-empty. The important restriction on rewrite rules of context-sensitive grammars is that the complete string on the right must be at least as long as the complete string on the left. The context-sensitive grammar, in this figure, generates the language $0^n1^n2^n$. There is no context-free grammar that could generate this language.

A subset of computable languages are 'context-sensitive' languages, which are generated by context-sensitive grammars (Fig. 17.2). For each of these languages there exists a machine that can decide if a given sentence is part of the language or not. This can be done by a Turing machine with a finite memory, a so-called 'linear bounded automaton'.

(p.323) A subset of context-sensitive languages are 'context-free' languages, which are generated by context-free grammars (Fig. 17.2), which are equivalent to push-down automata. These are computers with a single memory stack; at any given time they have access only to the top register of their memory.

An even simpler machine can be constructed. A finite-state automaton has a start, a finite number of intermediate states, and a finish. Whenever the machine jumps from one state to the next, it emits an element of the alphabet. A particular run from start to finish produces a sentence. There are many different runs from start to finish, and hence there are

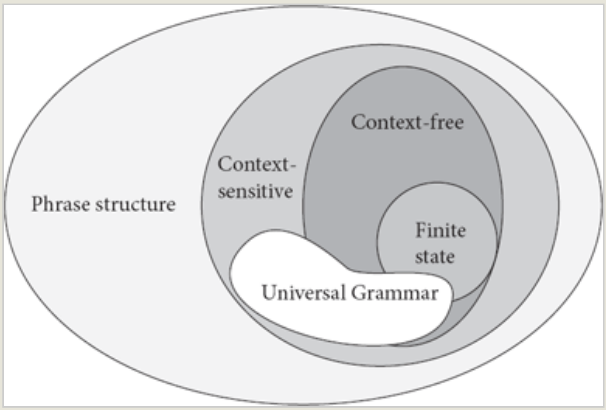

Fig. 17.3 The Chomsky hierarchy and the logical necessity of Universal Grammar. Finite-state grammars are a subset of context-free grammars, which are a subset of context-sensitive grammars, which are a subset of phrase structure grammars, which represent all possible grammars. Similarly, regular languages are a subset of context-free languages, which are a subset of context-sensitive languages, which are a subset of computable languages. Natural languages are considered to be more

many different sentences. If a finite-state automaton contains at least one loop, then it will be able to generate infinitely many sentences. Hence, finite-state automata can generate all finite languages and some infinite languages. Finite-state automata and the corresponding finite-state grammars define the set of regular languages.

powerful than regular languages. The crucial result of learning theory is that there exists no procedure that could learn an unrestricted set of languages; in most approaches even the class of regular languages is not learnable. The human brain has a procedure for learning language. This procedure can only learn a restricted set of languages. Universal Grammar is the theory of this restricted set.

Fig. 17.3 shows the Chomsky hierarchy: regular languages are a subset of context-free languages, which are a subset of context-sensitive languages, which are a subset of computable languages, which are disjunct from non-computable languages.

The Structure of Natural Languages

Natural languages are infinite: it is not possible to imagine a finite list that contains all English sentences. Furthermore, most linguists agree that finite-state grammars are inadequate for natural language. Such grammars are unable to represent long-range dependencies of the form: *If [she has an idea for how to solve the problem she has been working on for a long time] then [she will not go for lunch]*. The string of words between *If* and *then* could be arbitrarily long, and could itself contain more paired if-then constructions. Such pairings ultimately relate to rules that generate strings of the form 0^n1^n, which require context-free grammars.

The fundamental structures of natural languages are trees. The nodes represent phrases that can be composed of other phrases in a recursive manner. Finite-state automata can only generate a very limited class of trees, which is again an argument that natural grammars are more powerful than finite-state grammars. There is a continuing debate whether context-free grammars are adequate for natural languages (Pullum and Gazdar 1982; Shieber 1985), or whether context-sensitive grammars need to be evoked (Bar-Hillel 1953; Joshi et al. 1975).

(p.324) One can also define grammars that directly specify which trees are acceptable for a given language. A tree is a ‘derivation’ of a sentence within the rule system of a particular grammar. The interpretation of a sentence depends on the underlying tree structure. Ambiguity arises if more than one tree can be associated with a given sentence. Much of modern syntactic theory deals with grammars that directly act on tree structures (Chomsky 1984; Sadock 1991; Bresnan 2001; Pollard and Sag 1994). We cannot go into more detail here, but emphasize that all such grammars are ultimately placed somewhere on the

Chomsky hierarchy, and that the results of learning theory (to be discussed now) apply to them.

Learning Theory and a Logical Necessity of Universal Grammar

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.

The Paradox of Language Acquisition

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 number 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 (Jackendoff 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 'theory' of this restricted set is Universal Grammar (UG). The concept of an innate, genetically determined UG was controversial when introduced **(p.325)** some forty years ago and has remained so. The mathematical approach of learning theory, however, can explain in what sense UG is a logical necessity.

What is Learnable?

Imagine an ideal speaker-hearer pair. The speaker uses grammar G to construct sentences of language L. The hearer receives sentences and should, after some time, be able to use grammar G to construct other sentences of L. Mathematically speaking, the hearer is described by an algorithm (or more generally, a function), A, which takes a list of sentences as input and generates a language as output.

Let us introduce the notion of a 'text' as a list of sentences. Specifically, text T of language L is an infinite list of sentences of L with each sentence of L occurring at least once. Text T_N contains the first N sentences of T. We say that language L is learnable by the algorithm, A, if for each T of L there exists a number M such that for all $N>M$ we have $A(T_N)=L$. This means that, given enough sentences as input, the algorithm will provide the correct language as output.

Furthermore, a set of languages is learnable by an algorithm if each language of this set is learnable. We can imagine an algorithm that gives 'English' as output for every input. This algorithm can 'learn' English, but no other language. Hence, we are interested in the question of what set of languages, $l=(L_l, L_2.)$, can be learned by a given algorithm.

We can now present a key result of learning theory: according to Gold's theorem (Gold 1967), there exists no algorithm that can learn the set of regular languages. As a consequence, no algorithm can learn a set of languages that contains the set of all regular languages. Hence, no algorithm can learn the set of context-free languages, the set of context-sensitive languages, or the set of computable languages. Needless to say, no algorithm can learn the set of all languages.

A common criticism of Gold's framework is that the learner has to identify exactly the right language. For practical purposes, however, it might be sufficient that the learner acquires a grammar which is almost correct. There are various extensions of the Gold framework. For example, the approach of statistical learning theory contains the crucial requirement that the learner converges with a certain probability to a language that is almost the correct language. In the framework of statistical learning theory, it turns **(p.326)** out that there exists no procedure that can learn the set of all regular languages. Hence, we obtain the same necessity of an innate UG as before.

Some statistical learning models are motivated by 'informational complexity'. The question is: given a specific amount of information, is there any procedure that can in principle infer the correct language? Other learning theories include the concept of 'computational complexity'. Here the question is: given a specific amount of information, is there any computer programme that can come up with the correct language in reasonable (polynomial) time? Considerations of computational complexity lead to further restrictions on the set of languages that can be learned.

The Necessity of Innate Expectations

We can now state in what sense there has to be an innate UG. The human child is equipped with a learning algorithm, A_H, which enables the child to learn certain languages. This algorithm can learn each of the existing 6,000 human languages and presumably many more, but it is impossible that A_H could learn *any* language. Hence, there is a set of languages that can be learned by A_H, and this set must be heavily restricted compared to the set of all possible languages. UG is the rule system that describes the restricted set of languages that is learnable by A_H.

Learning theory shows there must be an innate UG, which is a consequence of the particular learning algorithm, A_H, used by humans. Discovering properties of

A_H requires the empirical study of neurobiological and cognitive functions of the human brain involved in language 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. A particular approach is the 'principle and parameter theory', which assumes that the child comes equipped with innate principles and has to set parameters that are specific for individual languages (Chomsky 1981; 1984; Gibson and Wexler 1994; Manzini and Wexler 1987). Another approach is 'optimality theory', where the child has innate constraints, and learning a specific language is ordering these constraints (Prince and Smolensky 1997).

There is some discourse as to whether the learning mechanism, A_H, is language-specific or general-purpose (Elman et al.1996). Ultimately this is a question about the particular architecture of the brain and which neurons participate in which computations; but one cannot deny that there is **(p.327)** a learning mechanism, A_H, that operates on linguistic input and enables the child to learn the rules of human language. This mechanism can learn a restricted set of languages; the theory of this set is an innate UG.

Hence, the continuing debate around an innate UG should not be whether there is one, but what form it takes (Elman et al.1996; Tomasello 1999; Sampson 1999). One can dispute individual linguistic universals (Greenberg et al.1978; Comrie 1981; Baker 2001), but one cannot generally deny their existence.

There is also some discussion about the role neural networks can play in language acquisition. Learning theory clearly shows that there exists no neural network that can learn an unrestricted set of languages. It is well understood that any particular neural network can only learn a very specific set of rules (Geman et al.1992).

Sometimes it is claimed that the logical arguments for an innate UG rest on particular mathematical assumptions of generative grammars which deal only with syntax and not with semantics. Cognitive (Lakoff 1987; Langacker 1987) and functional linguistics (Bates and MacWhinney 1982) instead are not based on formal language theory, but use psychological objects such as symbols, categories, schemas, and images. This does not, however, remove the necessity of innate restrictions. The results of learning theory apply to any learning process, where a 'rule' has to be learned from some examples. This generalization is an inherent feature of any model of language acquisition and applies to semantics, syntax, and phonetics. Any procedure for successful generalization has to pick from a restricted range of hypotheses.

Evolution of Language

Let us now formulate a mathematical description of language acquisition. The material presented here is part of a larger effort to use quantitative and computational approaches to study the evolution of language (Cavalli-Sforza and

Feldman 1981; Aoki and Feldman 1987; Hurford 1989; Hashimoto and Ikegami 1996; Steels 1996; Kirby and Hurford 1997; Hazlehurst and Hutchins 1998; Nowak and Krakauer 1999; Kirby 2000; Nowak et al. 2000; Kirby 2001; Komarova and Nowak 2001; Cangelosi and Parisi 2002; Christiansen et al. 2002; Christiansen and Dale in press). 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.

(p.328) Universal Grammar, in turn, contains a rule system that generates a set (or a search space) of grammars, *(*G_1*,* G_2*,..., Gn)*. These grammars can be constructed by the language learner as potential candidates for the grammar that needs to be learned. The learner has a mechanism to evaluate input sentences and to choose one of the candidate grammars that are contained in his search space. 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).

More generally, it is also possible to imagine that UG generates infinitely many candidate grammars, {G_1 G_2,...}. 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 is expected with equal probability and all other grammars are expected with zero probability, which is equivalent to a finite search space.

A fundamental question of linguistics and cognitive science is what restrictions 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, 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 fulfil for a population of individuals to evolve coherent communication? In the following, we will discuss how to address this question.

Language Learning: From Individuals to Populations

Evolution takes place in heterogeneous populations. We have to envisage a population of speakers using slightly different languages. They communicate with each other, which affects their performance, survival, and reproduction.

Whether they use language for exchanging information, making plans, cooperative hunting, social bonding, deception, or cooperation, language affects the fitness of individuals. Those that are better at the language **(p.329)** game leave more offspring. Otherwise natural selection, which we use here inclusive of sexual selection, could not operate, and the alternative hypothesis would be that language is the product of neutral evolution (or the byproduct of some other process), which is unlikely given the extreme complexity of this trait (Pinker 1994).

Imagine a group of individuals that all have the same UG, given by a finite search space of candidate grammars, $G_1,\ldots, G_n$, and a learning mechanism for evaluating input sentences. Let us specify the similarity between grammars by introducing the numbers s; which denote the probability that a speaker who uses G_i will say a sentence that is compatible with Gj.

We assume there is a reward for mutual understanding. The pay-off for someone who uses G_i and communicates with someone who uses Gj 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 and when Gj 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 pay-off 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 pay-off derived from communication contributes to biological fitness: individuals leave offspring proportional to their pay-off. These offspring inherit the UG of their parents. They receive language input (sample sentences) from their parents and develop their own grammar. At first, we will not specify a particular learning mechanism but introduce the stochastic matrix, Q, whose elements, qij, denote the probability that a child born to an individual using G_i will develop Gj. (In this model, we assume that each child receives input from one parent. It is possible to extend this approach to allow input from several individuals.) The probabilities 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, then grammar acquisition is perfect for all candidate grammars.

The population dynamics of grammar evolution are then given by the following system of ordinary differential equations, which we call the ‘language dynamical equations’ (Nowak et al.2001): **(p.330)**

$$(1) \qquad \frac{dx_j}{dt}=\sum_{i=1}^{n} f_i q_{ij} x_i-\phi x_j, \quad j=1,\ldots,n.$$

The term $-\varphi x_j$ ensures that the total population size remains constant: the sum over the relative abundances, $\Sigma_i x_i$, is 1 at all times. The variable

$$\phi=\sum_{i=1}^{n} f_i x_i$$

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 φ=1, all sentences are understood and communication is perfect. In general, φ is a number between 0 and 1.

The language dynamical equation is reminiscent of the *quasi-species equation* of molecular evolution, but has frequency dependent fitness values: the quantitiesfi 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. Thus, our model provides a connection between two of the most fundamental equations of evolutionary biology.

Evolution of Grammatical Coherence

In general, equation (1) above admits multiple (stable and unstable) equilibria (Fig. 17.4). For low accuracy of grammar acquisition (low values q_{ii}), all grammars, G_i, occur with roughly equal abundance. There is no predominating grammar in the population. Grammatical coherence is low. 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. This means that, if the accuracy of learning is sufficiently high, the population will converge to a stable equilibrium with one dominant grammar. Which one of the stable equilibria is chosen depends on the initial condition.

The accuracy of language acquisition depends on UG. 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 sij 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 **(p.331)** other grammars (s_ij close to 1 for many different;'), then G_i will be hard to learn. In the limit $s_{;}=1$, G_i is considered unlearnable, because no sentence can refute the hypothesis that the speaker uses *Gj*.

The accuracy of language acquisition also depends on the learning mechanism specified by UG. An inefficient learning mechanism or one that evaluates only a small number of input sentences will lead to a low accuracy and hence prevent

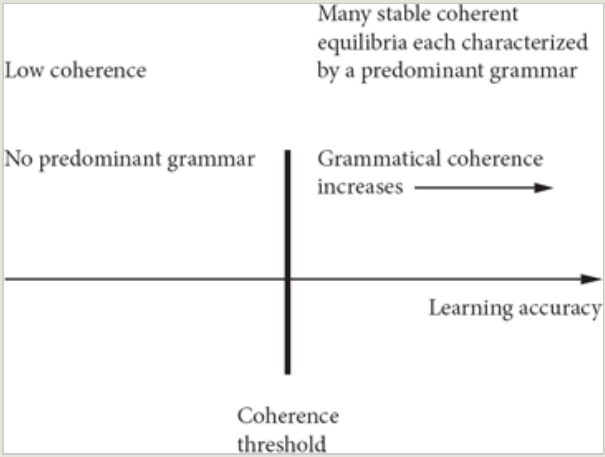

Fig. 17.4 Coherence threshold. When the accuracy of individual learning is low, equation (1) admits a single low-

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 UG can induce grammatical coherence in a population?*

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.

As outlined above, the answer will depend on the learning mechanism and the search space. We can derive results for two learning mechanisms that represent reasonable boundaries for the actual, unknown learning mechanism employed by humans:

(p.332)

> The memoryless learning algorithm, a favourite with learning theorists, makes few demands 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 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 a certain number of sentences. The algorithm is called 'memoryless' be cause 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 any more (this is, incidentally, the definition of so called 'consistent learners').
>
> The other extreme is a batch learner. The batch learner memorizes all sentences and at the end chooses the candidate grammar that is most com patible with the input.

Let us first make a simplifying assumption that all the n grammars are in some sense equally distant from each other. Mathematically speaking this amounts to setting $sij=s$, some constant, for $i \neq j$, and $s_i i = 1$. For the memoryless learner, we can show that grammatical coherence is possible if the number of input sentences, N, exceeds a constant times the number of candidate grammars, $N >$

Q n. For the batch learner, the number of input sentences has to exceed a constant times the logarithm of the number of candidate grammars, $N>C_2 \log n$.

In the more general case, when the similarity coefficients are taken from a uniform distribution, we can prove that for the memoryless learner, the number of sample sentences has to exceed $c1n \log n$ in order for the population to maintain grammatical coherence. For the batch learner, this result is $N > c_2n$. These inequalities 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 fulfil the coherence threshold does not lead to a stable, predominating grammar in a population.

The learning mechanism used by humans will perform better than the memoryless learner and worse than the batch learner; hence it will have a coherence threshold somewhere between $N>c_1n \log n$ and $N>c_2n$. The coherence threshold relates a life history parameter of humans, N, to the maximum size of the search space, n, of Universal Grammar.

(p.333) List Makers and Rule Finders

Next, let us explore the conditions under which natural selection favours the emergence of a rule-based, recursive grammatical system with infinite expressibility. In contrast to such rule-based grammars, one might consider list-based grammars that consist only of a finite number of sentences. Such list-based grammars can be seen as very primitive evolutionary precursors (or alternatives) to rule-based grammars. Individuals would acquire their mental grammar not by searching for underlying rules, but simply by memorizing sentences and their meaning (similar to memorizing the arbitrary meaning of words). List-based grammars do not allow for creativity on the level of syntax. Nevertheless, whether or not natural selection favours the more complicated rule-based grammars depends on circumstances that we need to explore.

Current human grammars can generate infinitely many sentences, but for the purpose of transmitting information only a finite number of them can be relevant. Natural selection cannot directly reward the theoretical ability to construct infinitely long sentences. Let us therefore consider a group of individuals that use M. different sentences (or syntactic structures). Note that M. specifies the number of sentences that are relevant from the perspective of biological fitness.

Now imagine individuals that learn their mental grammar by memorizing lists of sentences. We can ask how many sample sentences, N, a child must hear for the whole population to maintain M. sentences. If all sentences occur equally often, we simply obtain $N> M.$.

We can compare the performance of individuals using list-based versus rule-based grammars. Let us use the result for batch learners, which have comparable memory requirements to the list learners, and assume that grammar similarity coefficients, s_{ij}, are distributed uniformly between zero and one. Then we obtain that the number of relevant sentences, M., has to exceed a constant times the number of the candidate grammars, *n*. We have

$$M > c_3 n.$$

If this condition did *not* hold, it would be more efficient to *memorize* sentences associated with arbitrary meaning. In this case, language would have remained a rather dull communication system without any creative ability on the level of syntax. If, on the other hand, the condition above is satisfied, then rule-based grammars are more efficient than list-based grammars and **(p.334)** will have a fitness advantage. Furthermore, if rule-based grammars are selected, then the potential for 'making infinite use of finite means' comes as a by-product.

Applications for Historical Linguistics

The language dynamic equation can be used to study language change in the context of historical linguistics (Lightfoot 1991; Kroch 1989; Wang 1998; Niyogi and Berwick 1997; de Graff 1999). Here, a good assumption is that minor language changes are selectively neutral. Hence we can use a neutral version of our approach possibly in conjunction with stochastic and spatial population dynamics. It is possible to show that coherence threshold phenomena calculated for deterministic dynamics of infinite populations carry over to stochastic dynamics of finite populations.

Of special interest is that, for neutral language dynamics, we find linguistic coherence for

$$u < 1/M,$$

where *u* is the error rate of language acquisition and M is the effective population size. (This condition was derived under the symmetry assumption *sij=s* so that $q_i i=1-u$ for all i.) This condition relates the accuracy of language learning by individuals and the size of the linguistic community. If the linguistic coherence threshold is satisfied, then the language is passed down the generations with a high degree of accuracy; if the condition is violated, then a language change is likely to occur.

Neutral coherence is an important finding because it explains how linguistic features that do not contribute to communicative fitness (efHcacy) can be fairly homogeneous in a population. Neutral language dynamics provide an appropriate description for many language changes studied in historical linguistics, where fitness effects can probably be neglected.

Cultural Evolution of Language vs. Biological Evolution of Universal Grammar

Evolution of UG requires variation of UG. Thus UG is neither a grammar nor universal. Imagine a population of individuals using universal grammars U1 to U_M. Each UI admits a subset of *n* grammars and determines a particular

learning matrix $Q^{(I)}$. *UI* mutates genetically to *UJ* with probability **(p.335)** W_{IJ}. Deterministic population dynamics are given by

$$(2) \quad \frac{dx_{J_j}}{dt} = \sum_{I=1}^{m} W_{IJ} \sum_{i=1}^{n} f_{I_i} Q_{ij}^{(J)} x_{I_i} - \phi x_{J_j} \quad j = 1, .., n \quad J = 1, .., M$$

This equation describes mutation and selection among M different universal grammars. The relative abundance of individuals with Universal Grammar *UJ* speaking language Lj is given by x_{Jj}. At present, little is known about the behaviour of this system. In the limit of no mutation among universal grammars, $W_I I = 1$, we find that the selective dynamics often lead to the elimination of all but one universal grammar, but sometimes coexistence of different UGs can be observed. Equation (2) describes two processes on different time scales: the biological evolution of UG and the cultural evolution of spoken language (Fig. 17.5).

Only a UG that is sufficiently specific can lead to coherent communication in a population. The ability to induce a coherent language is a major selective criterion for UG. There is also a trade-off between learnability and adaptability: a small search space (small *n)* is more likely to lead to linguistic coherence, but might exclude languages with high communicative pay-off.

Since the necessity of a restricted search space applies to any learning task, we can use an extended concept of UG for animal communication. Therefore, during primate evolution, there was a succession of UGs that finally led to the UG of currently living humans. At some point a UG emerged that **(p.336)** allowed languages of unlimited expressibility. Such evolutionary dynamics are described by equation (2) above.

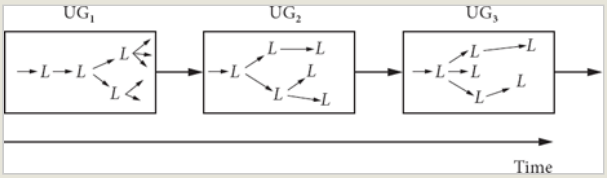

Fig. 17.5 The evolution of language takes place on two distinct time-scales. The larger scale corresponds to the biological evolution of Universal Grammar. Cultural evolution of language takes place on a much shorter time-scale. When modelling cultural evolution of language, we can assume that biological changes in the innate ‘hardware’ are negligibly small, and treat UG as constant in time and throughout the population. This brings us back to language dynamic equation (1). In order to describe changes in UG, we

need to consider biological mutations, equation (2).

Conclusions

We have outlined a connection between language, learning and evolution. Ultimately, these three fields of investigation need to be combined: ideas of language theory should be discussed in the context of acquisition, and ideas of acquisition in the context of evolution. The aim is to use evolutionary theory to understand basic design features of human language and the way it is learned. Evolutionary and game-theoretic approaches should outline the gradual emergence of various parts of human language, such as arbitrary signs, words, lexicons, and grammatical rules. Some of these problems constitute individual research projects that can be addressed separately without taking into account the whole unmeasurable complexity of human language. Fascinating questions are: How does a population settle for an arbitrary sign? How does natural selection equip a population with the cognitive machinery for grasping a (linguistic) rule? What is the interplay between the biological evolution of Universal Grammar and the cultural evolution of particular languages? Each of these questions seems to be as rich, for example, as the problem of evolution of cooperation, which is a vast field of evolutionary biology and incidentally tightly linked to language: we speak because we cooperate, we cooperate because we speak.

Cooperation is also required for advancing this line of research. Linguists need to establish more contact with biology, while biologists need to show more concern for human language, which is after all evolution’s most interesting invention ever since multicellularity-that is, within the last 500 million years.

FURTHER READING

Classical learning theory was formulated by Gold (1967). Perhaps the most significant extension of the classical framework is statistical learning theory. Here, the learner is required to converge approximately to the right language with high probability. For statistical learning theory, see Vapnik (1998). A deep result, originally due to Vapnik and Chervonenkis (1971) and elaborated since, states that a set of languages is learnable if and only if it has finite VC dimension. The VC dimension is a combinatorial measure of the complexity of a set of languages. Thus if the set **(p.337)** of possible languages is completely arbitrary (and therefore has infinite VC dimension), learning is not possible.

Considerations of computational complexity can also be added, where the learner is required to approximate the target grammar with high confidence using an efficient algorithm. Consequently, there are sets of languages that are learnable in principle (have finite VC dimension), but no algorithm can do this in polynomial time: see Valiant (1984).

For more information on the present authors’ approach to modelling learning and evolution of language, see the following papers: Nowak and Krakauer (1999)

and Nowak, Plotkin, and Krakauer (1999) for language in a game-theoretic framework, Nowak, Krakauer, and Dress (1999) for an error limit in the evolution of language, Nowak, Plotkin, and Jansen (2000) for the evolution of syntactic communication, Komarova and Nowak (2001) for the evolution of a lexical matrix, Nowak et al. (2001) and Komarova, Niyogi and Nowak (2001) for modelling of the evolution of UG, Komarova and Rivin (2003) for the convergence speed of the memoryless learner algorithm, and Komarova and Nowak (2003) for the evolution of language in finite populations.