Computational and evolutionary aspects of language

Martin A. Nowak*, Natalia L. Komarova*† & Partha Niyogi‡

** Institute for Advanced Study, Einstein Drive, Princeton, New Jersey 08540, USA*
† Department of Mathematics, University of Leeds, Leeds LS2 9JT, UK
‡ Department of Computer Science, University of Chicago, Chicago, Illinois 60637, USA

Language is our legacy. It is the main evolutionary contribution of humans, and perhaps the most interesting trait that has emerged in the past 500 million years. Understanding how darwinian evolution gives rise to human language requires the integration of formal language theory, learning theory and evolutionary dynamics. Formal language theory provides a mathematical description of language and grammar. Learning theory formalizes the task of language acquisition—it can be shown that no procedure can learn an unrestricted set of languages. Universal grammar specifies the restricted set of languages learnable by the human brain. Evolutionary dynamics can be formulated to describe the cultural evolution of language and the biological evolution of universal grammar.

Biology uses generative systems. Genomes consist of an alphabet of four nucleotides, which, together with certain rules for how to produce proteins and organize cells, generates an unlimited variety of living organisms. For more than 3 billion years, evolution of life on Earth was restricted to using this generative system. Only very recently another generative system emerged, which led to a new mode of evolution. This other system is human language. It enables us to transfer unlimited non-genetic information among individuals, and it gives rise to cultural evolution.

Currently there are many efforts to bring linguistic inquiry into contact with several areas of biology including evolution[1–11], genetics[12–14], neurobiology[15,16] and animal behaviour[17–20]. The aim of this Review is to formulate a synthesis of formal language theory[21,22], learning theory[23–28] and evolutionary dynamics in a manner that is useful for people from various disciplines. We will address the following questions: What is language? What is grammar? What is learning? How does a child learn language? What is the difference between learning language and learning other generative systems? In what sense is there a logical necessity for genetically determined components of human language, such as 'universal grammar'? Finally, we will discuss how formal language theory and learning theory can be extended to study language as a biological phenomenon, as a product of evolution.

Formal language theory

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 form syllables and words. Words form phrases and sentences. The rules for such groupings are not arbitrary. Any native English speaker recognizes that the sentence 'He ran from there with his money' obeys the rules of English, while 'He his money with there from ran' does not. In Bengali the reverse is true.

Individual languages have specific rules. 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.

What is language?

An alphabet is a set containing a finite number of symbols. Possible alphabets for natural languages are 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. Without loss of generality, we can consider the binary alphabet, {0,1}, by enumerating the actual alphabet in binary code.

A sentence is defined as a string of symbols. The set of all sentences over the binary alphabet is {0,1,00,01,10,11,000,...}. There are infinitely many sentences, as many as integers; the set of all sentences is 'countable'.

A language is a set of sentences. Among all possible sentences

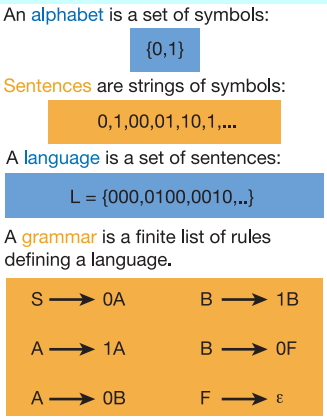


Figure 1 The basic objects of formal language theory are alphabets, sentences, languages and grammars. Grammars consist of rewrite rules: a particular string can be rewritten as another string. Such rules contain symbols of the alphabet (here 0 and 1), and so-called 'non-terminals' (here S, A, B and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

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.

What is grammar?

A grammar is a finite list of rules specifying a language. A grammar is expressed in terms of 'rewrite rules': a certain string can be rewritten as another string. Strings contain elements of the alphabet together with 'non-terminals', which are place holders. After iterated application of the rewrite rules, the final string will only contain symbols of the alphabet. Figures 1 and 2 give examples of grammars.

There are countably infinitely many grammars; any finite list of rewrite rules can be encoded by an integer. As there are uncountably many languages, only a small subset of them can be described by a grammar. These languages are called 'computable'.

Languages, grammars and machines

There is a correspondence between languages, grammars and machines. 'Regular' languages are generated by finite-state grammars, which are equivalent to finite-state automata. Finite-state automata have 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, hence there are many different sentences. If a finite-state machine contains at least one loop, then it can generate infinitely many sentences. Finite-state automata accept all finite languages

Box 1

Statistical learning theory and other approaches

Classical learning theory as formulated by Gold[23] makes a number of somewhat problematic assumptions: (1) the learner has to identify the target language exactly; (2) the learner receives only positive examples; (3) the learner has access to an arbitrarily large number of examples; and (4) the learner is not limited by any consideration of computational complexity. Assumptions (1) and (2) are restrictive: relaxing these assumptions will enable particular learners to succeed on larger sets of languages. Assumptions (3) and (4) are unrestrictive: relaxing these assumptions will reduce the set of learnable languages. In fact, each assumption has been removed in various approaches to learning, but the essential conclusion remains the same: no algorithm can learn an unrestricted set of languages.

Perhaps the most significant extension of the classical framework is statistical learning theory[24–26]. Here, the learner is required to converge approximately to the right language with high probability. Let us consider the following objects that play an important role in the basic framework of statistical learning:

- Languages and indicator functions. Let $\mathscr{L}$ be a set of languages. Every language L of this set defines an indicator function $1_L(s)$ that takes the value 1 if a sentence s is in L and 0 otherwise.
- Linguistic examples. Linguistic examples are provided to the learner according to some distribution P on the set of all sentences. The learner receives both positive and negative examples.
- Distances between languages. The probability measure P has a dual role. In addition to providing sentences, it also defines the distance between two languages, $d(L_1,L_2) = \sum_s |1_{L_1}(s) - 1_{L_2}(s)|P(s)$.
- Learnability. The criterion for learnability is specified as follows. Assume some language $L \in \mathscr{L}$ is the target language. The learner receives a collection of positive and negative example sentences according to the distribution P. On the basis of these empirical data, the learner guesses a language $L_N \in \mathscr{L}$ after N examples have been received. The target L is said to be learnable if for all $\epsilon > 0$ the probability that the distance between L_N and L is greater than ϵ converges to 0 as $N \to \infty$. As a consequence, there exists a finite number of example sentences, such that the probability is greater than $1 - \delta$ (where δ is a small number) that the learner's current estimate is within an ϵ-approximation of the target language.

A deep result, originally due to Vapnik and Chervonenkis[24] 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 of possible languages is completely arbitrary (and therefore has infinite VC dimension), learning is not possible. It can be shown that the set of all regular languages (even the set of all finite languages) has infinite VC dimension and hence cannot be learned by any procedure in the framework of statistical learning theory. Subsets of regular languages that are generated by finite-state automata with n states, however, have finite VC dimension, and one can estimate bounds on the number of sample sentences that are needed for learning.

Statistical learning theory in the VC framework removes assumption (1), (2) and (3) of the Gold framework: it does not ask for convergence to exactly the right language, the learner receives positive and negative examples, and the learning process has to end after a certain number of examples. The theory provides bounds for how many example sentences are needed to converge approximately to the right language with high probability; this is the concept of informational complexity.

Valiant[25] also added considerations of computational complexity, thereby removing assumption (4) of the Gold framework: 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. (Computer scientists consider a problem 'intractable' if no algorithm can solve the problem in polynomial time, which means in a number of time steps that is proportional to the size of the input raised to some power.)

Some other models of learning deserve mention. For example, in one form of query-based learning, the learner is allowed to ask whether a particular sentence is in the target language or not. In this model, regular languages can be learned in polynomial time[72], but context-free languages cannot[73]. Other query-based models of learning, with varying degrees of psychological plausibility, have been considered, and none permit all languages to be learnable[74].

Another model of learning follows the classical Gold framework but only requires the learner to identify the target language on almost all texts. Texts are assumed to be generated by a probabilistic process, where sentences are drawn according to some measure μ on the target language. If the learner is to identify the target grammar in a distribution free fashion (that is independent of μ), then the set of learnable languages is no larger than those that are learnable in the classical Gold framework. If constraints are put on the family of measures μ then the set of learnable languages can be enlarged. Such constraints act like a probabilistic form of universal grammar.

In summary, all extensions of learning theory underline the necessity of specific restrictions.

and some infinite languages.

'Context-free' languages are generated by context-free grammars, which can be implemented by push-down automata. These are computers with a single memory stack: at any one time they have only access to the top register of their memory.

'Context-sensitive' languages are generated by context-sensitive grammars. For each of these languages there exists a Turing machine, which can decide for every sentence whether it is part of the language or not. A Turing machine embodies the theoretical concept of a digital computer with infinite memory.

Computable languages are described by 'phrase structure' grammars that have unrestricted rewrite rules. For each computable language, there exists a Turing machine that can identify every sentence that is part of the language. If, however, the Turing machine receives as input a sentence which does not belong to the language, then it might compute forever. Hence, the Turing machine cannot always decide whether a sentence is part of the language or not.

Figure 3 shows the Chomsky hierarchy: 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 are Turing complete.

The structure of natural languages

With the introduction of the Chomsky hierarchy, there was some interest in placing natural languages within this scheme. Natural languages are infinite: it is not possible to imagine a finite list that contains all English sentences. Furthermore, finite-state grammars are inadequate for natural languages. Such grammars are unable to represent long-range dependencies of the form 'if… then'. The string of words between 'if' and 'then' could be arbitrarily long, and could itself contain more paired if–then constructions. Such pairings relate to rules that generate strings of the form 0^n1^n, which require context-free grammars (Fig. 2). There is a continuing debate whether context-free grammars are adequate for natural languages, or whether more complex grammars need to be evoked[29,30].

The fundamental structures of natural languages are trees. The nodes represent phrases that can be composed of other phrases in a recursive manner. 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. One can also define grammars that directly specify which trees are acceptable for a given language. Much of modern syntactic theory

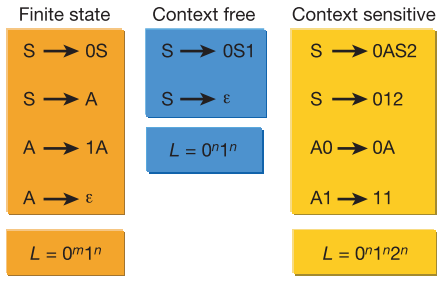


Figure 2 Three grammars and their corresponding languages. Finite-state grammars have rewrite rules of the form: a single non-terminal (on the left) is rewritten as a single terminal possibly followed by a non-terminal (on the right). The finite-state grammar, in this figure, generates the regular language 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 non-terminals. 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 $\alpha A\beta \rightarrow \alpha\gamma\beta$. Here α, β and γ are strings of terminals and non-terminals. Although 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.

Box 2
Language as mapping between sound and meaning

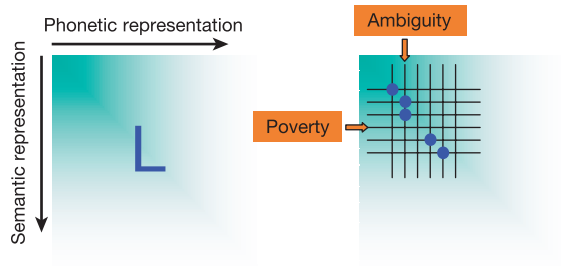
The mathematical formalism of language can be extended to include aspects of communicative behaviour and performance. A language can be seen as a (possibly infinite) matrix, L, that specifies mappings between phonetic forms and semantic forms that are 'sound' and 'meaning' insofar as they are linguistically determined (see Box 2 figure). This matrix defines linguistic competence. For evaluating communicative success, we also need to describe linguistic performance. We assume that the matrix, L, leads to a pair of matrices, P and Q, determining speaking and hearing. The element p_{ij}^l denotes the probability for a speaker of L_l to use sound j for encoding meaning i. The element q_{ij}^l denotes the probability for a hearer to decode sound j as meaning i. A language may not encode all meanings and may not make use of all possible sounds.

Next, we introduce a measure, σ, on the set of all meanings. Let us denote by σ_i the probability that communication is about meaning i. The measure, σ, depends among other things on the environment, behaviour and phenotype of the individuals. It also defines which meanings are more relevant than others.

The probability that a speaker of L_l generates a sound which is understood by a hearer using L_J is given by $a_{lJ} = \sum_{ij} \sigma_i p_{ij}^l q_{ij}^J$. The communicative pay-off between L_l and L_J can be defined as $F_{lJ} = \frac{1}{2}(a_{lJ} + a_{Jl})$. The intrinsic communicative pay-off of L_l is $F_{ll} = a_{ll}$.

In this framework, communicative pay-off is a number between 0 and 1. A pay-off of less than 1 arises as a consequence of ambiguity and poverty. Ambiguity, α_l, of language L_l is the loss of communicative capacity that arises if individual sounds are linked to more than one meaning. Poverty, β_l, is the fraction of meanings (measured by σ) that are not part of L_l. The communicative capacity of L_l can be written as $F_{ll} = (1 - \alpha_l) \times (1 - \beta_l)$.

For language acquisition we need a measure for the similarity, s_{lJ}, between languages L_l and L_J. A possibility is $s_{lJ} = \sum_{ij} \sigma_i p_{ij}^l q_{ij}^J / \sum_{ij} \sigma_i p_{ij}^l$ denoting the probability that a sound generated by a speaker of L_l is correctly interpreted by a hearer using L_J. In this context, the similarity between L_l and L_J declines because of ambiguity, but not because of poverty. Ambiguity implies that a learner holding the correct hypothesis might think he is wrong and change his hypothesis. This has consequences for the notion of consistent learning; a consistent learner does not change his hypothesis if he already holds the correct hypothesis.



deals with such grammars[31–34], which are, of course, also part of the Chomsky hierarchy, and the results of learning theory, to be discussed now, apply to them.

Learning theory

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 will generalize to novel sentences never heard before. Any person can produce and understand sentences that are not part of his previous linguistic experience. Learning theory describes the mathematics of learning with the aim of outlining conditions for successful generalization.

The paradox of language acquisition

Children learn their native language by hearing grammatical sentences from their parents or others. From this 'environmental input', children construct an internal representation of the underlying grammar. Children are not told the grammatical rules. Neither children nor adults are ever aware of the grammatical rules that specify their own language.

Chomsky pointed out that the environmental input available to the child does not uniquely specify the grammatical rules[35]. This phenomenon is known as 'poverty of stimulus'[36]. 'The paradox of language acquisition' is that children of the same speech community reliably grow up to speak the same language[37]. The proposed solution 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). Formally, UG is not a grammar, but a theory of a collection of grammars.

The concept of an innate, genetically determined UG was controversial when introduced some 40 years ago and has remained so. The mathematical approach of learning theory, however, can explain in what sense UG is a logical necessity.

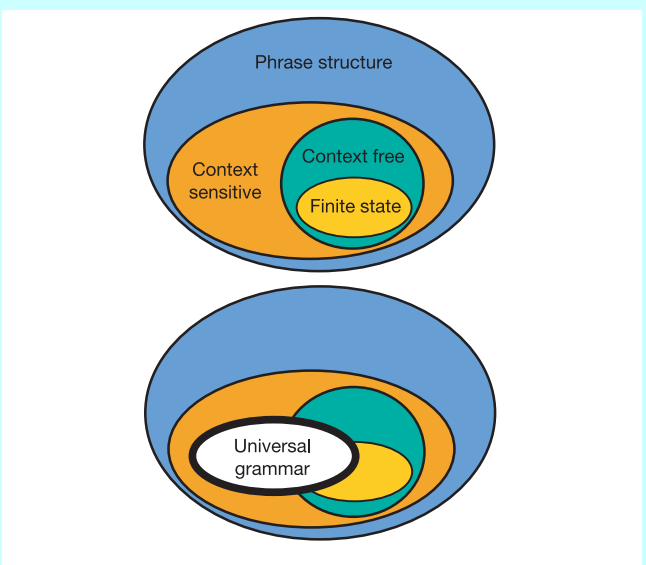


Figure 3 The Chomsky hierarchy and the logical necessity of universal grammar. Finite-state grammars are a subset of context-free grammars, which are a subset of context-sensitive grammars, which are a subset of phrase-structure grammars, which represent all possible grammars. Natural languages are considered to be more 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, but this procedure can only learn a restricted set of languages. Universal grammar is the theory of this restricted set.

Learnability

Imagine a 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 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 are interested in what set of languages, $\mathscr{L} = \{L_1, L_2, ..\}$, can be learned by a given algorithm.

A key result of learning theory, Gold's theorem[23], implies 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 regular languages, such as the set of context-free languages, context-sensitive languages or computable languages.

Gold's theorem formally states there exists no algorithm that can learn a set of 'super-finite' languages. Such a set includes all finite languages and at least one infinite language. Intuitively, if the learner infers that the target language is an infinite language, whereas the actual target language is a finite language that is contained in the infinite language, then the learner will not encounter any contradicting evidence, and will never converge onto the correct language. This result holds in greatest possible generality: 'algorithm' here includes any function from text to language.

Probably almost correct

A common criticism of Gold's framework is that the learner has to identify exactly the right language. For practical purposes, it might be sufficient that the learner acquires a grammar that is almost correct. Box 1 explains various extensions of the Gold framework, and in particular the approach of statistical learning theory. Here, the crucial requirement is that the learner converges with high probability to a language that is almost correct. Statistical learning theory also shows there is no procedure that can learn the set of all regular languages, thereby confirming the necessity of an innate UG. Some learning theories provide more information for the learner and thus allow larger classes of languages to be learnable, but no learning theory admits an unrestricted search space.

Learning finite languages

Some readers might think that the arguments of learning theory rely on subtle properties of infinite languages. Let us therefore consider finite languages. In the context of statistical learning theory, the set of all finite languages cannot be learned. In the Gold framework, the set of all finite languages can be learned, but only by memorization: the learner will identify the correct language only after having heard all sentences of this language. A learner that considers the set of all finite languages has no possibility for generalization: the learner can never extrapolate beyond the sentences he has already encountered. This is not the case for natural language acquisition: we can always say new sentences.

Let us consider a finite set of finite languages. Suppose there are 3 sentences, S_1, S_2, S_3. Hence there are 8 possible languages. Suppose learner A considers all 8 languages, while learner B considers only 2 languages, for example $L_1 = \{S_1, S_2\}$ and $L_2 = \{S_3\}$. If learner A receives sentence S_1, he has no information whether sentences S_2 or S_3 will be part of the target language or not. He can only identify the target language after having heard all sentences. If learner B receives sentence S_1 he knows that S_2 will be part of the language, whereas S_3

will not. He can extrapolate beyond his experience. The ability to search for underlying rules requires a restricted search space.

The necessity of innate expectations

We can now state in what sense there has to be an innate UG. The human brain is equipped with a learning algorithm, A_H, which enables us 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 every computable language. Hence, there is a restricted set of languages that can be learned by A_H. UG is the theory of this restricted set.

Learning theory suggests that a restricted search space has to exist before data. By 'data' we mean linguistic or other information the child uses to learn language or modify its language acquisition procedure. Therefore, in our terminology, 'before data' is equivalent to 'innate'. In this sense, 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 past decades. A particular approach is the 'principles and parameters theory', which assumes that the child comes equipped with innate principles and has to set parameters that are specific for individual languages[38,39]. Another approach is 'optimality theory', where learning a specific language is ordering innate constraints[40].

There is some discourse as to whether the learning mechanism, A_H, is language specific or general purpose[41]. 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 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 UG. The continuing debate around an innate UG should not be whether there is one, but what form it takes[41–43]. One can dispute individual linguistic universals[44,45], but one cannot generally deny their existence.

Neural networks are an important tool for modelling the neural mechanisms of language acquisition. The results of learning theory also apply to neural networks: no neural network can learn an unrestricted set of languages[46].

Sometimes it is claimed that the logical arguments for an innate UG rest on particular mathematical assumptions of generative grammars that deal only with syntax and not with semantics. Cognitive[47,48] and functional linguistics[49] are not based on formal language theory, but use psychological objects such as symbols, categories, schemas and images. This does not 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. 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 choose from a restricted range of hypotheses.

The results of learning theory also apply to learning mappings between linguistic form and meaning. If meaning is to be explicitly considered, then a language is not a set of sentences, but a set of sentence-meaning pairs (Box 2). The task of language acquisition is then to learn grammars that generate sentence-meaning pairs. Such grammars are also part of the Chomsky hierarchy, and there exists no learning procedure that can succeed on an unrestricted set of such languages.

What is special about language acquisition?

Usually when we learn the grammar of generative systems, such as chess or arithmetic, somebody tells us the rules. We do not have to guess the moves of chess by looking at chess games. In contrast, the process of language acquisition occurs without being instructed about rules; neither teachers nor learners are aware of the rules. This is an important difference: if the learner is told the grammar of a language, then the set of all computable languages is learnable by an algorithm that memorizes the rules.

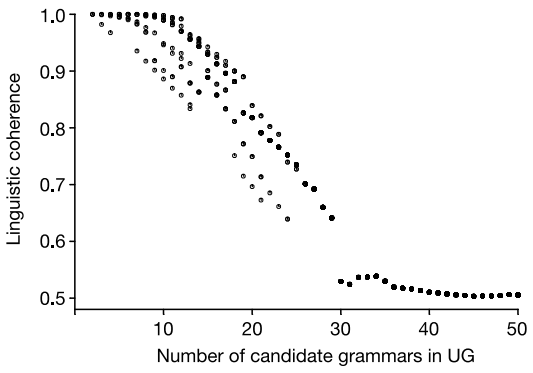


Figure 4 Linguistic coherence evolves if universal grammar (UG) is sufficiently specific. The figure shows equilibrium solutions of the language dynamical equation. Linguistic coherence, ϕ, is the probability that one individual says a sentence that is understood by another individual. UG specifies n candidate grammars. The similarity between any two candidate grammars, s_{ij}, is a random number from a uniform distribution on [0, 1]. The language acquisition device is a memoryless learner receiving $N = 100$ example sentences. For $n > 30$, all candidate grammars are represented in the population with similar frequencies; the linguistic coherence, ϕ, is about 1/2, which means complete randomness. For $n < 30$ the equilibrium is dominated by a single grammar. For each value of n there can be multiple equilibria dominated by different grammars. The equilibrium that is reached depends on the initial condition. Coherence is required for adaptation of language and selection of UG for linguistic function.

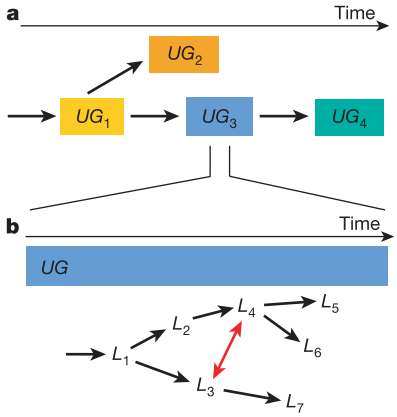


Figure 5 Two aspects of language evolution. **a**, There is a biological evolution of universal grammar (UG) via genetic modifications that affect the architecture of the human brain and the class of languages it can learn. UG can change as a consequence of (1) random variation (neutral evolution), (2) as a by-product of selection for other cognitive function or (3) under selection for language acquisition and communication. At some point in the evolutionary history of humans, a UG arose that allowed languages with infinite expressibility. **b**, On a faster timescale, there is cultural evolution of language constrained by a constant UG. Languages change by (1) random variation, (2) by contact with other languages (red arrow), (3) by hitch-hiking on other cultural inventions, or (4) by selection for increased learnability and communication. Although many language changes in historical linguistics might be neutral, a global picture of language evolution must include selection.

Evolutionary language theory

Humans and chimpanzees separated some 5 million years ago. Chimpanzees have a complex system of conceptual understanding and rich social interactions, but they do not have communication comparable to human language. The central question of the origin of human language is which genetic modifications led to changes in brain structures that were decisive for human language. Given the enormous complexity of this trait, we should expect several incremental steps guided by natural selection. In this process, evolution will have reused cognitive features that evolved long ago and for other purposes.

Understanding language evolution requires a theoretical framework explaining how darwinian dynamics lead to fundamental properties of human language such as arbitrary signs, lexicons, syntax and grammar[50–62]. Here we outline a minimalist program combining formal language theory, learning theory and evolutionary theory.

The basic approach is similar to evolutionary game theory. There is a population of individuals. Each individual uses a particular language. Individuals talk to each other. Successful communication results in a pay-off that contributes to fitness. Offspring inherit—subject to mutations—a mechanism to learn language and a UG. They use this mechanism to learn—subject to mistakes—the language of their parents or others.

Cultural evolution of language with constant universal grammar

From a biological perspective, language is not the property of an individual, but the extended phenotype of a population. Let us consider a population where all individuals have the same UG, and let us assume that UG does not change from one generation to the next. Suppose that (in accordance with principles and parameters theory or optimality theory) UG specifies a finite number of languages $L_1,..L_n$. Each individual in the population will speak one of those n languages.

We need a model of language that allows us to calculate the communicative pay-off, F_{ij}, for an individual using L_i talking to an individual using L_j. Box 2 outlines a fairly general approach, based on the assumption that a language can be seen as an infinite binary matrix linking phonetic forms to semantic forms. The languages, L_i, can differ in their intrinsic communicative pay-off, F_{ii}, which depends on ambiguity and expressive power. Some languages could be finite, others infinite.

Denote by x_i the relative abundance of speakers of L_i. The fitness of L_i is given by $f_i = \sum_{j=1}^{n} x_j F_{ij}$. Individuals leave offspring proportional to their pay-off. The probability that a child will develop L_j if the parent uses L_i is given by Q_{ij}. The 'learning matrix', Q, depends on the learning algorithm and UG. The language dynamical equation[62] is given by:

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

The term $-\phi(\mathbf{x})x_j$ ensures that $\sum_i x_i = 1$. The variable $\phi(\mathbf{x}) = \sum_i f_i(\mathbf{x}) x_i$ denotes the average fitness of the population, and is a measure for linguistic coherence. The dynamics can also be interpreted in a purely cultural sense: individuals that communicate successfully are more likely to influence language acquisition of others. Equation (1) describes selection of languages for increased communicative function, F_{ii}, and increased learnability, Q_{ii}.

For low accuracy of language acquisition, when Q is far from the identity matrix, there is no predominating language in the population, and the linguistic coherence is low. As the accuracy of language acquisition increases, and Q gets closer to the identity matrix, equilibrium solutions arise where a particular language is more abundant than others. The population has achieved linguistic coherence. The 'coherence threshold' specifies the minimum specificity of UG that is required for linguistic coherence (Fig. 4).

For certain learning mechanisms we can calculate the coherence threshold. The 'memoryless learner' starts with a randomly chosen language and stays with it as long as the input sentences are compatible with this language. If a sentence arrives that is not compatible, then the learner picks at random another language. The learner does not memorize which languages have already been rejected. The process stops after N sentences. Another mechanism is the 'batch learner', which memorizes N sentences, and at the end chooses the language that is most compatible with all N sentences.

If the similarity coefficients between languages, s_{ij} (Box 2), are constant, $s_{ij} = s$ and $s_{ii} = 1$, then the memoryless learner has a coherence threshold $N > C_1 n$, whereas the batch learner has a coherence threshold $N > C_2 \log n$. If the s_{ij} values are taken from a uniform random distribution on the interval [0, 1] and if $s_{ii} = 1$, then the memoryless learner has a coherence threshold $N > C_3 n \log n$, whereas the batch learner has a coherence threshold $N > C_4 n$ (refs 63, 64). C_1 to C_4 are some constants. These conditions provide boundaries for the actual learning mechanism used by humans, which is arguably better than the memoryless learner and worse than the batch learner. The coherence threshold relates a life-history parameter of humans, N, to the maximum size of the search space, n, of UG.

Evolution of universal grammar

Evolution of UG requires variation of UG. (UG is in fact neither a grammar nor universal.) Imagine a population of individuals using UGs U_1 to U_M. Each U_I admits a subset of n grammars and determines a particular learning matrix Q^I. U_I mutates genetically to U_J with probability W_{IJ}. Deterministic population dynamics are given by:

$$\frac{dx_{Jj}}{dt} = \sum_{I=1}^{M} W_{IJ} \sum_{i=1}^{n} f_{Ii} Q^I_{ij} x_{Ii} - \phi x_{Jj} \quad j = 1, .., n \quad J = 1, .., M \qquad (2)$$

This equation describes mutation and selection among M different universal grammars. The relative abundance of individuals with UG U_J speaking language L_j is given by x_{Jj}. At present little is known about the behaviour of this system. In the limit of no mutation among UGs, $W_{II} = 1$, we find that the selective dynamics often lead to the elimination of all but one UG, but sometimes coexistence of different UGs can be observed. Equation (2) describes two processes on different timescales: the biological evolution of UG and the cultural evolution of language (Fig. 5).

The ability to induce a coherent language is a major selective criterion for UG. A UG that induces linguistic coherence allows language adaptation and can be selected for linguistic function. 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.

Because the necessity of a restricted search space applies to any learning task, we can use an extended concept of UG for animal communication. 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 allowed languages of unlimited expressibility. Such evolutionary dynamics are described by equation (2).

Historical linguistics

The language dynamical equation (1) can be used to study language change in the context of historical linguistics[65–68]. Languages change because the transmission from one generation to the next is not perfect. UG limits the type of variation that can occur. In the context of the principles-and-parameters theory, changes in syntax arise because children acquire different parameter settings[66]. Grammaticalization[69] is the process where lexical items take on grammatical function. Creolization is the formation of a new language by

children receiving mixed input[70,71]. All such language changes can be studied mathematically. Many language changes are selectively neutral. Hence, we can use a neutral version of our approach possibly in conjunction with small population sizes and stochastic and spatial population dynamics. These are open problems.

Outlook

We have reviewed mathematical descriptions of language on three different levels: formal language theory, learning theory and evolution. These approaches need to be combined: ideas of language should be discussed in the context of acquisition, and ideas of acquisition in the context of evolution.

Some theoretical questions are: what is the interplay between the biological evolution of UG and the cultural evolution of language? What is the mechanism for adaptation among the various languages generated by a given UG? In terms of the principles-and-parameters theory, can we estimate the maximum number of parameters compatible with the coherence threshold? Some empirical questions are: what is the actual language learning algorithm used by humans? What are the restrictions imposed by UG? Can we identify genes that are crucial for linguistic or other cognitive functions? What can we say about the evolution of those genes?

The study of language as a biological phenomenon will bring together people from many disciplines including linguistics, cognitive science, psychology, genetics, animal behaviour, evolutionary biology, neurobiology and computer science. Fortunately we have language to talk to each other. □

doi:10.1038/nature00771.

1. Pinker, S. & Bloom, A. Natural language and natural selection. *Behav. Brain Sci.* **13,** 707–784 (1990).
2. Jackendoff, R. Possible stages in the evolution of the language capacity. *Trends Cogn. Sci.* **3,** 272–279 (1999).
3. Bickerton, D. *Language and Species* (Univ. Chicago Press, Chicago, 1990).
4. Lightfoot, D. *The Development of Language: Acquisition, Changes and Evolution* (Blackwell, Oxford, 1999).
5. Brandon, R. & Hornstein, N. From icon to symbol: Some speculations on the evolution of natural language. *Phil. Biol.* **1,** 169–189 (1986).
6. Hurford, J. R., Studdert-Kennedy, M. A. & Knight, C. (eds) *Approaches to the Evolution of Language* (Cambridge Univ. Press, Cambridge, UK, 1998).
7. Newmeyer, F. Functional explanation in linguistics and the origins of language. *Lang. Commun.* **11,** 3–28, 97–108 (1991).
8. Lieberman, P. *The Biology and Evolution of Language* (Harvard Univ. Press, Cambridge, Massachusetts, 1984).
9. Maynard Smith, J. & Szathmary, E. *The Major Transitions in Evolution* (Freeman Spektrum, Oxford, 1995).
10. Hawkins, J. A. & Gell-Mann, M. *The Evolution of Human Languages* (Addison-Wesley, Reading, Massachusetts, 1992).
11. Aitchinson, J. *The Seeds of Speech* (Cambridge Univ. Press, Cambridge, UK, 1996).
12. Cavalli-Sforza, L. L. Genes, peoples and languages. *Proc. Natl Acad. Sci. USA* **94,** 7719–7724 (1997).
13. Gopnik, M. & Crago, M. Familial aggregation of a developmental language disorder. *Cognition* **39,** 1–50 (1991).
14. Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F. & Monaco, A. P. A forhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413,** 519–523 (2001).
15. Deacon, T. *The Symbolic Species* (Penguin, London, 1997).
16. Vargha-Khadem, F. *et al.* Neural basis of an inherited speech and language disorder. *Proc. Natl Acad. Sci. USA* **95,** 12695–12700 (1998).
17. Smith, W. J. *The Behaviour of Communicating* (Harvard Univ. Press, Cambridge, UK, 1977).
18. Dunbar, R. *Grooming, Gossip, and the Evolution of Language* (Cambridge Univ. Press, Cambridge, UK, 1996).
19. Fitch, W. T. The evolution of speech: a comparative review. *Trends Cogn. Sci.* **4,** 258–267 (2000).
20. Hauser, M. D. *The Evolution of Communication* (Harvard Univ. Press, Cambridge, Massachusetts, 1996).
21. Chomsky, N. A. *Syntactic Structures* (Mouton, New York, 1957).
22. Harrison, M. A. *Introduction to Formal Language Theory* (Addison-Wesley, Reading, Massachusetts, 1978).
23. Gold, E. M. Language identification in the limit. *Informat. Control* **10,** 447–474 (1967).
24. Vapnik, V. N. & Chervonenkis, A. Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theor. Prob. Applicat.* **17,** 264–280 (1971).
25. Valiant, L. G. A theory of learnable. *Commun. ACM* **27,** 436–445 (1984).
26. Vapnik, V. N. *Statistical Learning Theory* (Wiley, New York, 1998).
27. Osherson, D., Stob, M. & Weinstein, S. *Systems That Learn* (MIT Press, Cambridge, Massachusetts, 1986).
28. Pinker, S. Formal models of language learning. *Cognition* **7,** 217–283 (1979).
29. Pullum, G. K. & Gazdar, G. Natural languages and context free languages. *Linguist. Phil.* **4,** 471–504 (1982).
30. Shieber, S. M. Evidence against the context-freeness of natural language. *Linguist. Phil.* **8,** 333–343 (1985).
31. Chomsky, N. A. *Lectures on Government and Binding: The Pisa Lectures* (Foris, Dordrecht, 1984).
32. Sadock, J. M. *Autolexical Syntax: A Theory of Parallel Grammatical Representations. Studies in Contemporary Linguistics* (Univ. Chicago Press, Chicago, 1991).
33. Bresnan, J. *Lexical-Functional Syntax* (Blackwells, London, 2001).
34. Pollard, C. J. & Sag, I. A. *Head-Driven Phrase Structure Grammar* (Univ. Chicago Press, Chicago, 1994).
35. Chomsky, N. *Language and Mind* (Harcourt Brace Jovanovich, New York, 1972).
36. Wexler, K. & Culicover, P. *Formal Principles of Language Acquisition* (MIT Press, Cambridge, Massachusetts, 1980).
37. Jackendoff, R. *Foundations of Language* (Oxford Univ. Press, Oxford, 2001).
38. Chomsky, N. in *Explanation in Linguistics* (eds Hornstein, N. & Lightfoot, D.) 123–146 (Longman, London, 1981).
39. Baker, M. C. *Atoms of Language* (Basic Books, New York, 2001).
40. Prince, A. & Smolensky, P. Optimality: From neural networks to universal grammar. *Science* **275,** 1604–1610 (1997).
41. Elman, J. L. *Rethinking Innateness* (MIT Press, Cambridge, Massachusetts, 1996).
42. Tomasello, M. *The Cultural Origins of Human Cognition* (Harvard Univ. Press, Cambridge, Massachusetts, 1999).
43. Sampson, G. *Educating Eve: The Language Instinct Debate* (Cassell Academic, London, 1999).
44. Greenberg, J. H., Ferguson, C. A. & Moravcsik, E. A. (eds) *Universals of Human Language* (Stanford Univ. Press, Stanford, 1978).
45. Comrie, B. *Language Universals and Linguistic Typology* (Univ. Chicago Press, Chicago, 1981).
46. Geman, S., Bienenstock, E. & Doursat, R. Neural networks and the bias/variance dilemma. *Neural Comput.* **4,** 1–58 (1992).
47. Langacker, R. *Foundations of Cognitive Linguistics* Vol. 1 (Stanford Univ. Press, Stanford, 1987).
48. Lakoff, G. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind* (Univ. Chicago Press, Chicago, 1987).
49. Bates, E. & MacWhinney, B. *Language Acquisition: The State of the Art* (Cambridge Univ. Press, Cambridge, 1982).
50. Aoki, K. & Feldman, M. W. Toward a theory for the evolution of cultural communication: Coevolution of signal transmission and reception. *Proc. Natl Acad. Sci. USA* **84,** 7164–7168 (1987).
51. Hurford, J. R. Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77,** 187–222 (1989).
52. Cangelosi, A. & Parisi, D. *Simulating the Evolution of Language* (Springer, London, 2002).
53. Kirby, S. & Hurford, J. *Proc. Fourth European Conf. on Artificial Life* (eds Husbands, P. & Harvey, I.) 493–502 (MIT Press, Cambridge, Massachusetts, 1997).
54. Steels, L. *Proc. Fifth Artificial Life Conf.* (eds Langton, C. G. & Shimohara, T.) 113–131 (MIT Press, Tokyo, 1996).
55. Nowak, M. A. & Krakauer, D. C. The evolution of language. *Proc. Natl Acad. Sci. USA* **96,** 8028–8033 (1999).
56. Nowak, M. A., Plotkin, J. B. & Jansen, V. A. A. Evolution of syntactic communication. *Nature* **404,** 495–498 (2000).
57. Komarova, N. L. & Nowak, M. A. Evolutionary dynamics of the lexical matrix. *Bull. Math. Biol.* **63,** 451–485 (2001).
58. Christiansen, M. H., Dale, R. A. C., Ellefson, M. R. & Conway, C. M. in *Simulating the Evolution of Language* (eds Cangelosi, A. & Parisi, D.) 165–187 (Springer, London, 2002).
59. Hashimoto, T. & Ikegami, T. Emergence of net-grammar in communicating agents. *Biosystems* **38,** 1–14 (1996).
60. Hazlehurst, B. & Hutchins, E. The emergence of propositions from the coordination of talk and action in a shared worlds. *Lang. Cogn. Process.* **13,** 373–424 (1998).
61. Pinker, S. *The Language Instinct* (Morrow, New York, 1994).
62. Nowak, M. A., Komarova, N. L. & Niyogi, P. Evolution of universal grammar. *Science* **291,** 114–118 (2001).
63. Komarova, N. L. & Rivin, I. Mathematics of learning. Preprint math.PR/0105235 at ⟨http://lanl.arXiv.org⟩ (2001).
64. Rivin, I. Yet another zeta function and learning. Preprint cs.LG/0107033 at ⟨http://lanl.arXiv.org⟩ (2001).
65. Lightfoot, D. *How to Set Parameters: Arguments from Language Change* (MIT Press, Cambridge, Massachusetts, 1991).
66. Kroch, A. Reflexes of grammar in patterns of language change. *Lang. Variat. Change* **1,** 199–244 (1989).
67. Wang, W. S. Y. in *The Origins and Past of Modern Humans* (eds Omoto, K. & Tobias, P. V.) 247–262 (World Scientific, Singapore, 1998).
68. Niyogi, P. & Berwick, R. C. Evolutionary consequences of language learning. *Linguist. Phil.* **20,** 697–719 (1997).
69. Hopper, P. & Traugott, E. *Grammaticalization* (Cambridge Univ. Press, Cambridge, 1993).
70. de Graff, M. *Language Creation and Language Change: Creolization, Diachrony and Development* (MIT Press, Cambridge, MA, 1999).
71. Mufwene, S. *The Ecology of Language Evolution* (Cambridge Univ. Press, Cambridge, 2001).
72. Angluin, D. Learning regular sets from queries and counterexamples. *Informat. Comput.* **75,** 87–106 (1987).
73. Angluin, D. & Kharitonov, M. When won't membership queries help? *J. Comput. Syst. Sci.* **50,** 336–355 (1995).
74. Gasarch, W. & Smith, C. Learning via queries. *J. Assoc. Comput. Machin.* **39,** 649–674 (1992).

Acknowledgements

Support from the David and Lucille Packard foundation, the Leon Levy and Shelby White initiatives fund, the Florence Gould foundation, the Ambrose Monell foundation, the National Science Foundation and J. E. Epstein is acknowledged.

Correspondence and requests for materials should be addressed to M.A.N. (e-mail: nowak@ias.edu).