# Human language as a culturally transmitted replicator

*Mark Pagel*

Abstract | Human languages form a distinct and largely independent class of cultural replicators with behaviour and fidelity that can rival that of genes. Parallels between biological and linguistic evolution mean that statistical methods inspired by phylogenetics and comparative biology are being increasingly applied to study language. Phylogenetic trees constructed from linguistic elements chart the history of human cultures, and comparative studies reveal surprising and general features of how languages evolve, including patterns in the rates of evolution of language elements and social factors that influence temporal trends of language evolution. For many comparative questions of anthropology and human behavioural ecology, historical processes estimated from linguistic phylogenies may be more relevant than those estimated from genes.

Languages
Linguists identify two languages as distinct when, according to various criteria, they become mutually unintelligible.

*School of Biological Sciences, University of Reading, Reading, Berkshire RG6 6AH, UK; and Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico, USA.*
*e-mail:* m.pagel@reading.ac.uk
doi:10.1038/nrg2560
Published online 7 May 2009

Here is a remarkable fact about humans: we speak approximately 7,000 mutually unintelligible languages around the world[1]. This means that a person plucked from one corner of the Earth is not able to communicate with another human in a different corner of the Earth, or often from next door. Apart from a few songbird species, and possibly some whales that learn their songs locally and show dialectical differences, this is unique among animals. For example, a chimpanzee or an elephant removed from its range and placed among any other chimpanzees or elephants will know what to do and how to communicate.

Large as the number of extant human languages is, it has probably reduced from a maximum of perhaps 12,000 to 20,000 different languages before the spread of agriculture[2], and it pales in comparison with the possibly hundreds of thousands of different languages humans have ever spoken[2]. Elsewhere I have pondered the question of why humans would evolve a system of communication that prevents them from communicating with other members of its species, and have suggested that human societies come to behave in ways that are not so different from that of biological species[3]. Whether or not that explanation is correct, the human tendency to separate into distinct societies has given human language a geographical mosaic on which to play out its evolution. My interest here is to use the phenomenon of language diversity to understand the evolution of what turns out to be a remarkable culturally transmitted replicator, one with many of the properties we have come to expect of genes, but also with many of its own.

In this Review I shall first describe how a new and expanding field of phylogenetic and comparative studies of language evolution has made use of concepts, data and statistical modelling approaches that draw inspiration from genetics to exploit the genetic-like properties of language. I shall then move on to describe recent work in four areas of language evolution in which statistical modelling approaches have begun to return results. These include the reconstruction of language phylogenies and their relationship to genetic trees; investigations of the rate, tempo and time-depth of language evolution; social influences on language; and studies of the structure of language.

My coverage of these topics will be selective, but is designed to give a flavour of what language evolution is like and of what is possible. I will not discuss the tricky and very large literatures on language origins or how we acquire it, whether our language skills are innate, or possible genetic influences on language abilities. Instead, I will treat language as evolving against what I will regard for sake of discussion as a more or less homogeneous genetic background in its human hosts.

## Descent with modification

One of the best-known theories for the diversity of human languages is a creation myth. According to the bible story of the Tower of Babel, humans developed the conceit that they could construct a tower that would take them all the way to heaven. Angered at the attempt to usurp his control, God destroyed the tower. To ensure

Table 1 | **Some analogies between biological and linguistic evolution**

| Biological evolution | Language evolution |
|---|---|
| Discrete heritable units (for example, nucleotides, amino acids and genes) | Discrete heritable units (for example, words, phonemes and syntax) |
| Mechanisms of replication | Teaching, learning and imitation |
| Mutation (for example, many mechanisms yielding genetic alterations) | Innovation (for example, formant variation, mistakes, sound changes, and introduced sounds and words) |
| Homology | Cognates |
| Natural selection | Social selection and trends |
| Drift | Drift |
| Cladogenesis* (for example, allopatric speciation (geographic separation) and sympatric speciation (ecological or reproductive separation)) | Lineage splits (for example, geographical separation and social separation) |
| Anagenesis‡ | Linguistic change without split |
| Horizontal gene transfer | Borrowing |
| Hybridization (for example, horse with zebra and wheat with strawberry) | Language Creoles§ (for example, Surinamese) |
| Correlated genotypes and phenotypes (for example, allometry‖ and pleiotropy¶) | Correlated cultural terms (for example, 'hasta' and 'spear') |
| Geographic clines# | Dialects and dialect chains |
| Fossils | Ancient texts |
| Extinction | Language death |

Darwin noted many of these parallels in *The Descent of Man*[4]. Table is modified, with permission, from *Nature* REF. 26 © (2007) Macmillan Publishers Ltd. All rights reserved. *Cladogenesis: the formation of separate groups by evolutionary splitting. ‡Anagenesis: the evolutionary process whereby one species evolves into another without any splitting of the lineage into separate groups or species. §Creole: a language that emerges in the second or later generations of the speakers of pidgins (which are the rudimentary languages that form when two language communities mix and seek a common basis for simple communication). Creoles are typically more complex than pidgins, although less so than fully developed languages. ‖Allometry: the relationship between size and shape. ¶Pleiotropy: the action of a single gene on two or more distinct phenotypic characters. #Clines: a gradual change in phenotype in a species over a given area.

that it could not be rebuilt, God confused the workers by giving them different languages, leading to the irony that language exists to stop us from communicating.

Delightful as the Babel story is, ideas taken from the theory of evolution give us the conceit that we can improve on it. Darwin[4] asserted that languages, like biological species, evolve by a process of descent with modification. If correct, we can expect human languages to form into family trees, known as phylogenies, which chart the history of their evolution in a manner analogous to that for biological species. It also means that the diversity of extant languages reflects the actions of various shared historical evolutionary processes, including features of the rate and tempo of linguistic evolution, timings and correlations, as well as the starting points or ancestral languages. This raises the possibility that, far from settling for each language being a distinct object of creation, we can use the combination of phylogenetic trees of language along with statistical models of how languages evolve to detect and characterize the signature of these historical processes. In effect, we wish to discover what the past must have been like and how it evolved given what we now see.

TABLE 1 records analogies between the ways that genes and languages evolve, giving hope that the use of phylogenetic methods will succeed. Key among these analogous features is that both systems of replicators are digital, comprising discrete heritable units: the four nucleotides in the case of genes, and words in the case of language. Without this property neither system would retain fidelity through repeated bouts of transmission from parents to offspring (genes) or from teachers to learners (language), and historical signals would quickly be lost. Other features of language evolution that might be thought to vitiate its historical signature have analogies in genetic systems. For example, languages can acquire new unrelated words by borrowing, and genes can arrive from bouts of lateral transfer. These influences often occur at lower rates in genetic systems, but do not represent a qualitative difference between genes and language.

## Data and statistical modelling

The starting point for most comparative statistical investigations of language evolution is a set of discrete characters that can be scored in each of the languages. These might include features of the syntax or structure of a language, other aspects of grammar, phonemes and, most obviously, lexical items or words. I shall confine my remarks here to the lexicon, although most of what I have to say applies to these other classes of discrete traits. Owing to pioneering work by Morris Swadesh[5] in the 1950s a common list of 200 words known as the fundamental vocabulary is available for a large number of the world's languages (see the further information box for a link to an example of a Swadesh list). This type of list contains words for things that are expected to be found in all languages, such as names for body parts, pronouns, common verbs and numerals, but excludes technological words and words related to specific ecologies or habitats. It can be thought of as like a list of universal genes. Other lists are possible, but Swadesh's has simply proven to be well chosen and widely available. His words tend to evolve slowly and are largely resistant to outside influences and borrowing[6].

**Phylogeny**
A branching diagram describing the set of ancestral–descendant relationships among a group of species or languages.

**Borrowing**
The acquisition of a new non-cognate word from another language.

**Phoneme**
Characteristically thought of as the smallest units of speech-sounds that are distinguished by the speakers of a particular language. Phonemes are not universal, but act as the fundamental building blocks to produce all of the words of a given language.

Box 1 | **A linguistic alignment and a statistical model of evolution**

**Matrix of cognates**

Whereas a gene sequence alignment identifies homologous sites in genes, a lexical 'alignment' identifies sets of cognate words, or words that descend from a common ancestral word. Let a matrix of these lexical alignments be denoted M (see also REFS. 7,9,19) to signify that it is a matrix of meanings (for example, hand, who and ear), and write M as:

$$M = \begin{array}{c} \\ \text{language 1} \\ \text{language 2} \\ \vdots \\ \vdots \\ \text{language n} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{meanings} & & & & \\ 1 & 2 & 3 & & \text{m} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & . & 0 \\ 1 & 1 & 1 & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 3 & 1 & k & . & 0 \end{pmatrix} \end{array}$$

The columns of numbers designate cognate classes, or words for a given meaning that have been identified as deriving from a common ancestral word (see text). The first column of M denotes a meaning for which four distinct cognate classes of words exist (0, 1, 2 and 3), the second column shows a meaning represented by two cognate classes, the third has $k + 1$ cognate classes, and the last column shows a meaning with a single cognate class — that is, all of the words for that particular meaning among the $n$ languages derive from a common ancestral word. This matrix is the analogue to an aligned set of gene sequences, although all gene sequences have the same four states (twenty states if considering amino acids).

**A statistical model**

The data in M can be used to infer phylogenetic trees of languages, or perhaps to investigate some feature of lexical replacement or the change from one cognate class to another. A statistical model that is widely used in phylogenetic inference from gene sequences is written as:

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ . \\ . \\ k \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & . & . & k \end{array} \\ \begin{pmatrix} - & q_{01} & . & . & q_{0k} \\ q_{10} & - & . & . & q_{1k} \\ . & . & . & . & . \\ . & . & . & . & . \\ q_{k0} & q_{k1} & . & . & - \end{pmatrix} \end{array}$$

The matrix Q is the central element of the finite-state continuous time Markov transition model (see text). Each $q_{ij}$ term in this matrix describes the instantaneous rate of change from state $i$ to state $j$ over the short interval $dt$. In gene sequence data the states are the bases A, C, G or T, and Q is always a $4 \times 4$ matrix. If protein sequences are used, the states are amino acids and Q becomes a $20 \times 20$ matrix. Using lexical data, the states are the cognate classes, and a different Q needs to be estimated for meanings with different numbers of cognate classes. For phylogenetic inference with lexical data it is convenient to rewrite M such that all of the columns have the same number of cognate classes and thus a common Q can be estimated for the entire matrix, as is common for gene sequences. This is achieved by converting M to a binary form such that the $k + 1$ cognate classes for each meaning are written as k + 1 binary vectors, each one of which identifies a different cognate class as '1' with the remainder designated '0'. Then, Q becomes a $2 \times 2$ matrix estimating a common rate at which new cognate classes occur.

The elements of the Q matrix are presumed to apply equally well to each site in a gene or protein sequence or, in the case of lexical data, to different words. To accommodate variation in the rates of evolution among sites or among words, the well-known gamma rate variability correction can be applied[40]. This correction amounts to multiplying each of the $q_{ij}$ in Q by carefully chosen constants that are either less than or greater than one to achieve an overall slower or faster rate of evolution.

Outside the linguistic context M could contain any set of cross-cultural or other comparative data, and Q could then be used to estimate their evolutionary transitions (for example, REFS 8,9).

Given a list of $m$ meanings (such as hand, tree, I, walk, run) in $n$ languages, a data set (M) can be written as a matrix, analogous to an alignment of gene sequences (BOX 1), but recording sets of cognate words. Linguists, using careful rules of sound correspondences within language families, can assign words from different languages into classes denoting words that derive from a common ancestral word, analogous to identifying homologous genes. The word 'two' in English and *dos* in Spanish are cognate. The French *fleur* and Dutch *blumen* are not. Cognacy data, by recording evolved similarities and differences among languages, can be used to infer linguistic phylogenetic trees or to study features of lexical evolution itself[7]. Phylogenies are of interest in their own right as descriptions of historical relationships, and they form the backbone of comparative studies that seek to understand the evolution of linguistic traits, and how other cultural traits have evolved and co-evolved[7–9]. For these wider cross-cultural studies, M can be broadened to include the cultural data.

Statistical approaches apply models of evolution to characterize the probability of observing the data in M under various evolutionary scenarios[10,11]. A model that will be familiar to geneticists is the finite-state continuous time Markov transition model (BOX 1). Often designated Q, it was introduced into studies of phylogenetic inference from gene sequences by Felsenstein[12] along with the sum over histories logic. Applied to lexical data, this model estimates the instantaneous rates at which the words of one cognate class evolve into another unrelated set of words[7]. Other statistical approaches to analysing M include a 'stochastic Dollo' model[13] that allows each new cognate class to arise only once on a tree of languages. The name is a conscious nod to the Belgian palaeontologist Louis Dollo (1857–1931), who suggested that identical complex forms do not arise more than once in nature. A different stochastic treatment of M (described in REF. 14) allows for borrowing while estimating the underlying tree.

Parsimony or distance-based methods can be used instead of statistical approaches. However, statistical approaches, unlike the other methods, allow one to estimate directly parameters of the models of evolution (such as the $q_{ij}$ transition rates in Q, see BOX 1) and to test among different models for the same data[15]. Whether inferring trees or studying the evolution of traits on trees, the common currency for testing models is a quantity known as the likelihood or $L$, defined as an amount proportional to the probability of the data given the model[16]. It is conventionally written as $L \propto P(\text{M} \mid \text{Q}, T)$, where M and Q are as defined here, and $T$ refers to the phylogenetic tree on which the data in M are presumed to have evolved. Likelihood methods regard the observed data as a fixed observation. This makes them particularly suited to historical inference problems such as those in linguistics, in which the observed data arise only once. Thus, the likelihood does not describe the probability that the events under study happened (they did) or that the model is true; it merely describes the 'fit' between the observed data and an inferred tree, or model (such as Q) of how a trait evolved on a tree.

The likelihood can be found by maximum likelihood methods — in this case many different trees or models are tried and the one that gives the largest value of *L* is preferred. Alternatively, Markov chain Monte Carlo (MCMC) methods[17] are increasingly being employed to infer models and trees[9,18]. Rather than seeking a single 'best' solution, MCMC methods attempt to derive a distribution of outcomes consistent with the data, called the posterior distribution. Posterior distributions can be formed for trees, for their likelihoods and for the parameters of the model of evolution. Their attraction is in providing a measure of the uncertainty in the estimates of these various components. The posterior distribution of trees also provides the logical background against which to estimate models of evolution for other traits, this being a tidy way to account for the effects of uncertainty about the past.

## Language trees and gene trees

***Features of language trees.*** An early attempt to apply a likelihood sum over histories approach to languages made use of 7 Indo-European languages, 18 meanings and the finite-state Markov transition model Q described above[19]. The analysis yielded a phylogeny with the expected monophyletic groupings of Romance languages (Spanish, French and Romanian) and Germanic languages (German, Dutch and English), and Welsh as an out-group. In the same year a tree of 77 Austronesian languages appeared, which was derived from parsimony methods[20]. Holden[21], also using parsimony and a 100-word Swadesh list, inferred a tree of 93 Bantu languages.

Later analyses of the Bantu data with likelihood models returned more or less the same tree[22]. Gray and Atkinson[23] applied the Markov transition model in a MCMC context to analyse 87 Indo-European languages using the entire Swadesh 200-word list, estimating an ancestral age for Indo-European languages of between 7,800 and 9,800 years. Trees of Papuan languages have been inferred from both typological and lexical features of language[24]. Recently, Gray and colleagues have expanded their Austronesian sample to include over 400 languages, inferring the tree using MCMC approaches and the Markov transition model[25]. Their tree supports a scenario for the origin of this group in Formosa, beginning approximately 6,000 years ago.

My interest here is less in the trees *per se* than in their characteristics. FIGURE 1 shows a consensus tree derived from the Bayesian posterior distribution of trees for 87 Indo-European languages[23,26]. The tree recovers the expected clades of Romance, Germanic, Slavic, Indo-Iranian and Celtic languages, and suggests their deeper relationships. But what is remarkable about this tree is how tree-like it is, given all of the ways that a linguistic signal can be corrupted — most obviously by borrowing. The numbers near to the nodes of this tree record the posterior support for that node, defined as the proportion of trees in the posterior distribution in which that node was found. These posterior support values rival those found for many gene trees of a similar size[27]. Comparable degrees of posterior support are reported for the Bantu and Austronesian trees[20,22], if not for the Papuan languages[24]. Techniques designed to reveal conflicting phylogenetic signals (for example, SplitsTree and Neighbour-net analyses)[28], such as would arise from borrowing, typically reveal a healthy pattern of tree-like data[29].

The Indo-European and Bantu trees reflect population expansions or radiations into new areas, riding on the back of agriculture[21,23,30], whereas the Austronesian tree records an expansion that may have been propelled in fits and starts linked to developments in sea-going boat technologies[20,25]. These population processes might contribute to the elegance of the three phylogenies by reducing the opportunities for borrowing of lexical items among the speakers of differing languages. Trees must always be carefully checked for borrowing, but unless it regularly occurs among distantly related languages, the broad structure of language phylogenies should be relatively unaffected[31]. Owing to a battle lost at Hastings, England, in 1066, English was bombarded by words of Romance origin and now approximately 50% of its vocabulary derives from such stock. Still, despite its history, English correctly appears among the Germanic languages in the I-E tree (FIG. 1), although linguists often place it closer to Frisian than the basal position it occupies in its portion of the Germanic clade.

***Comparison of gene trees and language trees.*** If languages are not the 'closed shop' to outside influences that we have come to expect of eukaryotic organisms with sequestered germ lines, the strength of descent with modification in language trees shows that the cultural processes of language teaching and learning that transmit language from one generation to the next can have a surprisingly high fidelity and can show resistance to outside effects. Although genes may only be replicated once or a few times between generations, vocabulary items are replicated by producing a sound that is copied by a listener and then produced anew in a cyclical process that may occur many tens of thousands of times (or more) per word per speaker. The opportunities for mutation and corruption of this signal, not to mention for innovation and borrowing, are great and yet these simple lists of words can reconstruct the cultural history of groups of speakers spanning thousands of years.

Trees derived from language bear a range of relationships to gene trees for the same population, and this is as we should expect. Cavalli-Sforza[32] demonstrated in the late 1980s that the major genetic groupings of people around the world conform, with few exceptions, to their language groupings. This reveals, unsurprisingly, that people divided by large geographical distances drift apart genetically and linguistically. More fine-grained analyses reveal a different picture. Sometimes language groups conform closely to genetic groups even on a small geographical scale[33] and other times they do not[34]. This does not invalidate one kind of tree or elevate another. It tells us that some trees are good for tracking the movements of genes, and others for tracking the movement of cultures.

**Cognate**
Two words are deemed cognate if they derive by a process of descent with modification from a common ancestral word.

**Sum over histories**
A mathematical technique that accounts for all possible ancestral states (that is, all possible histories) when finding the likelihood of observing the gene sequence or other data among extant species.

**Parsimony**
When applied to phylogenetic inference in a linguistic context, parsimony is a method that seeks the phylogenetic tree that implies the fewest number of changes among cognate classes.

**Distance**
As applied to phylogenetic inference in a linguistic context, distance is a set of methods that infer an underlying phylogenetic tree from a matrix of the pair-wise differences among all languages.

**Likelihood**
A statistical quantity defined as an amount that is proportional to the probability of observing some set of data given a particular model of how those data arose. In linguistic phylogenetic applications one finds the likelihood of the lexical data on the proposed tree given some model of how words evolve.

**Maximum likelihood method**
A statistical technique for finding the parameters of a model that make the observed data most likely or probable under that model.

**Markov chain Monte Carlo**
(MCMC). A statistical method for searching a complex high-dimensional space. As applied to phylogenetic inference in a linguistic context, MCMC methods return a sample of trees that are statistically representative of the trees that might arise from a given model of how words evolve.

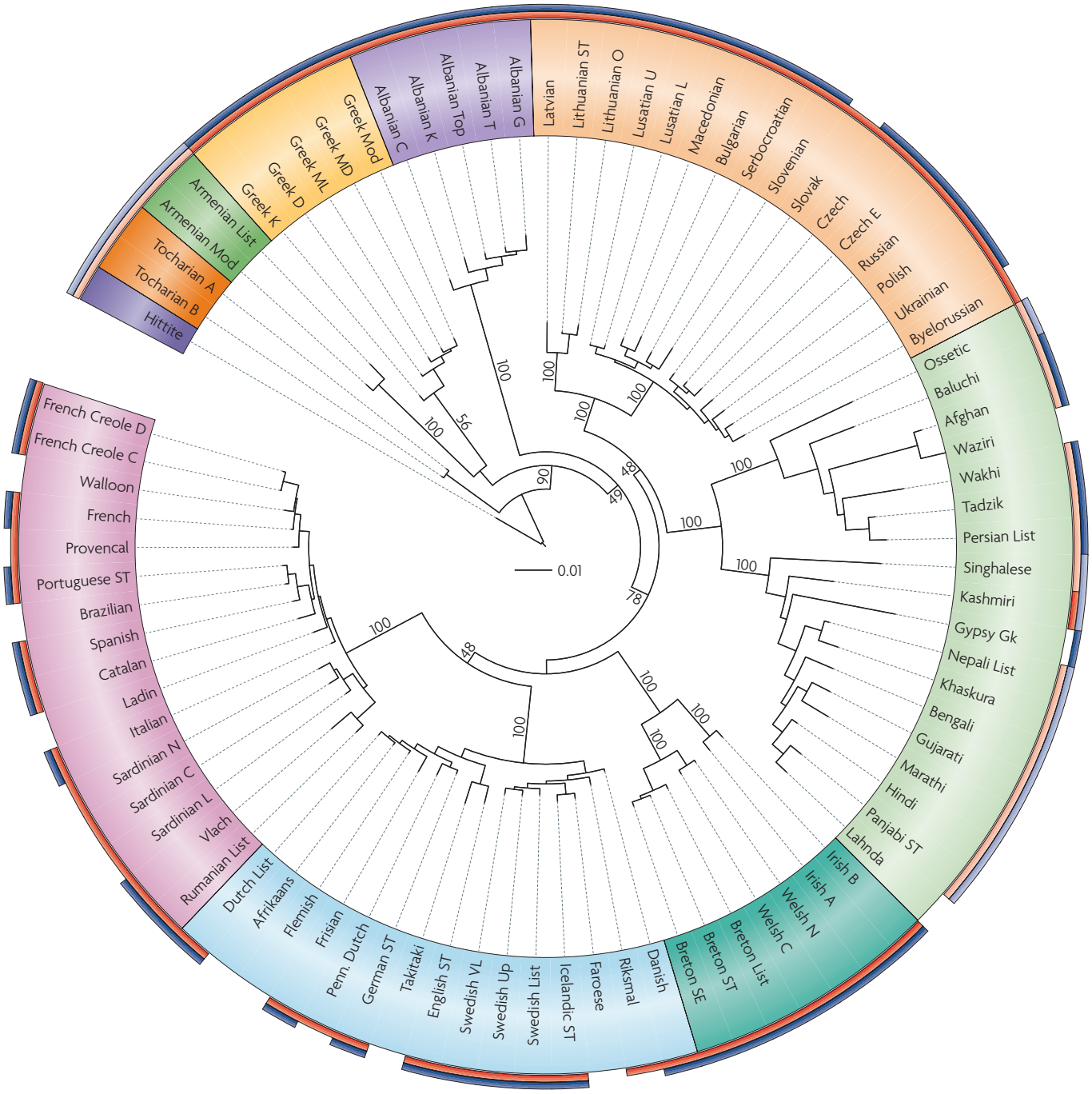


Figure 1 | **Tree of Indo-European languages.** Consensus tree of 87 Indo-European languages derived from the Swadesh list of 200 words[5,23,26]. Inner coloured ring identifies major clades as shown in the legend. The tree is rooted using ancient Hittite and Tocharian languages[23,26]. Branch lengths measure the expected number of lexical replacements (word changes) between two points on the tree. Numbers along branches are the Bayesian posterior probabilities of selected deep nodes of the tree, showing that words can resolve old relationships. Many of those nodes not labelled have high posterior support, although some are low and suffer from conflicting signals[26]. The outer colours identify a language's sentence word order in terms of subject (S) verb (V) and object (O) (red bars), and whether it employs pre or postpositional modification of sentence objects (blue bars) (see text, data from REF. 69). Red, SVO or VSO; light red, SOV; blue, prepositional; light blue, postpositional. Celtic languages are VSO; Greek, German, Dutch, Byelorussian and others are sometimes classified as no dominant word order (NDO) (here coded red). Blue–red pairs and light blue–light red pairs conform to Greenberg's[58] prediction (see text and FIG. 5)

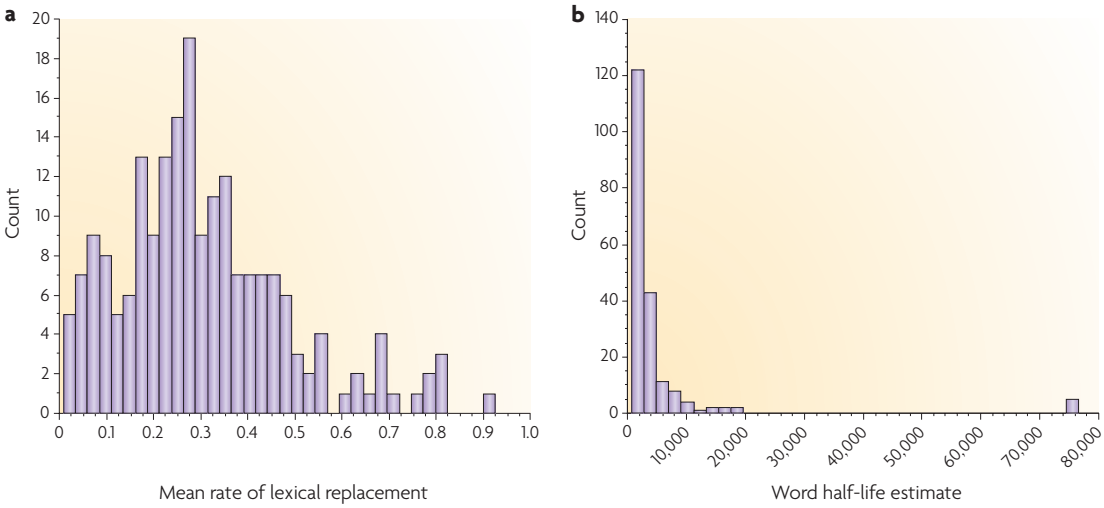


Figure 2 | **Rates of lexical replacement. a** | Histogram of mean rates of lexical replacement in Indo-European languages for the 200 words in the Swadesh list, measured in units of numbers of new cognates per 1,000 years of evolution. Fastest to slowest rate represents over 100-fold difference. Values were found as the mean of the posterior distribution of the elements of Q matrices (see text) integrated over a Bayesian posterior distribution of trees[26]. Mean = 0.3 ± 0.18 new cognates per 1,000 years, median = 0.27, range = 0.009 to 0.93. **b** | Histogram of the word half-life estimates as derived from the rates of lexical replacement, measuring the expected amount of time before a word has a 50% chance of being replaced by a new non-cognate word. A half-life of >70,000 years is indicative of a transition rate that is compatible with observing a single cognate class (that is, no changes) over the entire ~130,000 'language years' of the Indo-European tree[26] (calculated as the total obtained by adding the number of years of evolution represented by the sum of the branches of the tree in FIG. 1). Existence of at least five such classes in Indo-European lexicon lends support to this estimate. Mean = 5,300 years, median = 2,500, range = 750 to 76,000.

Indo-European languages
A family of related languages that derive from a common ancestral language that probably arose in Anatolia around 8,000 years ago and then spread throughout Europe, India, and what is now Afghanistan, Pakistan and Iran.

Monophyletic
In a phylogenetic context, a group of species (or languages) is monophyletic if they derive from a common ancestor not shared with any other species (or languages). The Germanic languages are monophyletic and are distinct from the monophyletic group of Romance languages. Monophyly implies that the group has just one origin.

Bantu languages
A group of approximately 500 languages that is part of the larger Niger-Congo language family. Bantu languages probably arose 3,000 years ago in West Africa, possibly close to present day Cameroon, and then spread east and then south eventually reaching to present day South Africa.

Clade
In the context of languages, a clade is a group of related languages.

Lexical replacement
The rate of lexical replacement is the rate at which a word is replaced by a new non-cognate word.

To understand why this is true, consider a thought experiment in which human genes flow among populations or even around the world largely invisible to the human phenotypes they inhabit (although there are some hints of genes and languages co-evolving[35]). Culture can, in principle, rest easily above this flow, as migrants adopt the local traditions, such that cultural variants and changes are independent of genetic changes. A situation similar to this has recently been reported for some Melanesian islanders[34]. Accordingly, phylogenetic trees derived from language may be preferable to gene trees in cross-cultural studies whenever the variables of interest are culturally transmitted[8]. These studies must separate the influence of common ancestry on a trait's representation among cultures from independent instances of the acquisition or evolution of that trait[8]. For cultural data, such as bride wealth and dowry, matriliny, patriliny, modes of subsistence and even sex ratio[36–38], the cultural phylogenetic tree provides the description of common ancestry that makes this separation possible. Linguists and anthropologists need not suffer from gene envy when it comes to building and using phylogenies.

In the next three sections I move away from inferring and interpreting language trees to discuss examples of how they have been used as the backbones of investigations into how features of language evolve, including rates of word evolution and the structure of languages, and to investigate social influences on the rates of lexical evolution.

## Rates of evolution and time depth

***Differing rates of word evolution.*** What English speakers call a bird, the Italians call *ucello*, the French *oiseau*, the Spanish *pajaro*, the Germans *vogel*, the Greeks *pouli*, and Caesar would have said *avis*. There are approximately 15 different cognate classes for 'bird' among the 90 or so Indo-European languages. By comparison, all Indo-European language speakers use a related form of the word 'two' (*dos*, *deux*, *due*, *zwei*; the Latin is *duo*) to describe two objects. Just as some sites in a gene sequence alignment evolve slowly and others rapidly, words in the Indo-European languages show ~100-fold variation in their rates of lexical replacement or in the acquisition of a new non-cognate form[7,26] (FIG. 2a). These rates were found from estimating the $q_{ij}$ in Q separately for each word in the Swadesh list, integrating over a Bayesian posterior sample of Indo-European trees. Slowly evolving words include 'two', 'three', 'I', 'five' and 'who', each of which has just a single cognate class among the Indo-European languages. By comparison, words such as 'bird', 'tail', 'sand', and 'belly' evolve more rapidly, with the word 'dirty' having, at 46, the largest number of cognate classes.

The rates can be expressed as word half-lives[7,19,26] (FIG. 2b) corresponding to the expected amount of time before a word has a 50% chance of being replaced by a new non-cognate word. The median half-life is 2,000–2,500 years. This may seem fast, especially when compared to genes, but it is slower than the average rate at which new languages appear (which is approximately every 500–1,000 years for Indo-European languages).

Even the most rapidly evolving words have fewer cognate classes than the number of languages. This means that, in general, words can achieve a measure of immortality by escaping into a new language before a new form replaces them.

Some words, like highly conserved genes, evolve at very slow rates. For each of the 5 most slowly evolving words there is a single cognate class in Indo-European languages. This is consistent with a half-life of over 70,000 years[26], a rate of evolution as slow as some genes[39], and shows that a culturally transmitted replicator can achieve a surprising fidelity. The sound an Indo-European language speaker makes to describe two objects is ancient, a related sound having been used by every speaker of an Indo-European language. If the unique time that each language has evolved is summed over languages this amounts to 130,000 or so language years that 'two' has remained stable. The same is true for the other words with a single cognate class. Even a word with a 6,000-year lexical half-life has a 25% chance of not changing in 11,500 years. Putting all these figures together, comparative linguists who seek evidence of very old linguistic signals are not simply chasing unicorns: there is every reason to expect that a linguistic signal exists that can identify relationships among distantly related language families.

***Reasons for rate heterogeneity.*** Heterogeneity in the rates of evolution of words can be accommodated when inferring language phylogenies in the same way as correction for differing rates of substitution in genetics (using the gamma correction)[11,40]. But at a deeper level we want to understand why this rate variation exists. We sought a general explanation for variation in rates of replacement by studying the 'expression level' of a word, that is, the frequency with which it is used in everyday speech[26]. Speech is dominated by a small number of frequently used words, the remainder being used infrequently[41,42]. We found that slowly evolving words in Indo-European are those with higher expression levels; they are used more frequently in everyday speech[26]. Within English, frequently used words are more likely to be of Old English origin[43]. For example, irregular English verbs retain their ancestral morphology[44,45] and are the more commonly used verbs.

Speakers of different Indo-European languages use the various words in the Swadesh list at similar frequencies in their everyday speech[26]. It might be that the way we use language and its structure means that some words inevitably will be used more than others; it is, for example, difficult to avoid verbs and pronouns. If so, then frequency of use has potentially been a general historical influence in the world's language families. FIGURE 3 plots the rates of lexical replacement we have reported for the Indo-European languages[26] against a list of 110 words that the late Russian comparative linguist Sergei Starostin identified as among the most stable in 14 language families from around the world[46]. The figure shows that slowly evolving words in Indo-European languages are also slowly evolving in the world's other language families, and vice versa; remarkably, this suggests that rates of evolution have been conserved throughout human history. This result attests to the generality and historical influence of the frequency effect, and gives additional support to the search for deep language relationships.

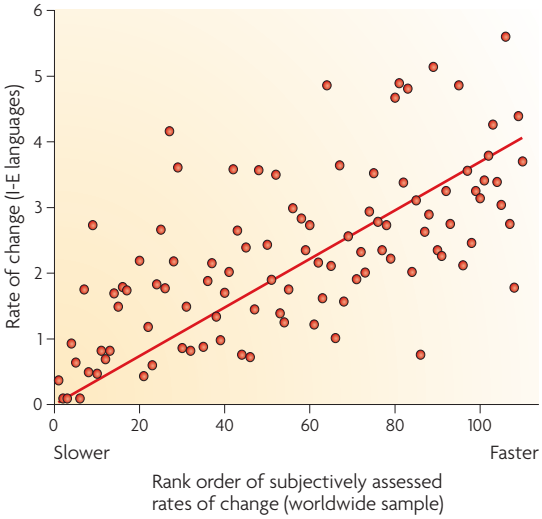


Figure 3 | **Rates of lexical replacement are stable among language families.** Statistically estimated rates of lexical replacement for 110 words from the Swadesh list in the Indo-European (I-E) languages (data from REF. 26 and FIG. 2a) correlated with rank ordering of subjectively assessed rates of change for the same words in a worldwide sample of 14 languages families, correlation ($r$) = 0.65. The 14 language families assessed were Sino-Tibetan, Austroasiatic, Altaic, Austronesian, Australian, Khoisan, North Caucasian, Dravidian, Indo-European, Kartvelian, Afroasiatic, Tai, Uralic and Yenisean. The rank order list was taken from REF. 46.

Frequency of use might affect rates of lexical replacement by altering 'production errors' — akin to the mutation rate in genetics — or by altering the rate at which a new form is adopted in a speech community (akin to selection), or both[26,47]. Word use may be under strong purifying selection within populations of speakers, if only through the rule 'speak as most others do'. It is difficult to understand how entire populations of speakers could otherwise agree on a single or small number of mostly arbitrary sounds to represent a given meaning. Such a rule would have been advantageous in our history if speakers who make mistakes are disadvantaged. If I say that the war-like tribe coming over the hill numbers two when in fact I meant two hundred, there may be consequences. Some words may acquire connections in the cognitive or semantic space[48], connections the strength or size of which may influence how rapidly words evolve. For example, *hasta* is the Sanskrit word for hand, but among Latin speakers it became the word for spear. The sound 'hasta' may have been saved by the cognitive connection between hand and spear. Questions surrounding why different words evolve at different rates are areas rich for discovery and are only just beginning to be investigated — they are likely to unlock fundamental aspects of how languages evolve.

Language year
In a phylogenetic context, each of the branches of a phylogeny represents some amount of evolution that occurs independently of the evolution in other branches. If the times in years that these branches represent are added together, the result records the total number of years of evolution that the tree represents; that is, the total number of language years.

Gamma correction
An elegant mathematical technique developed for characterizing the evolution of gene sequences that allows the nucleotides at different sites in the gene to evolve or be replaced at varying rates. The same technique can be applied to characterize the differing rates of evolution among lexical items.

Linguistic universals
A set of features of language and relationships among those features that the great comparative linguist Joseph Greenberg proposed would be found in all or nearly all languages, or which would at least show statistical evidence for being linked.

Word order
The typical order of subjects, verbs and objects in a sentence.

## Social effects and bursts of linguistic change

Are there external forces that affect linguistic change independently of the ways we use language in everyday speech? Here I briefly discuss one way in which languages, by acting as markers of social identity, may influence the rate of linguistic evolution.

Languages are not evenly distributed geographically. Cultural groups are more densely packed in coastal than inland regions[49,50]. Similarly, the density or number of different indigenous languages spoken in a given area of North America before European contact sharply increases in the more southerly regions of that continent, and is startling in its similarity to a plot of the density of different biological species from the same area[3,51] (FIG. 4). Human cultural–linguistic groups seem to partition the landscape in a manner similar to species, and perhaps for similar reasons: where in the southerly regions the landscape is richer and more ecologically diverse, a greater variety of species seems able to coexist. The puzzle is that humans are all the same species, and so their higher densities in tropical regions may suggest a tendency for cultural groups to fission whenever the environment will support it[3,51].

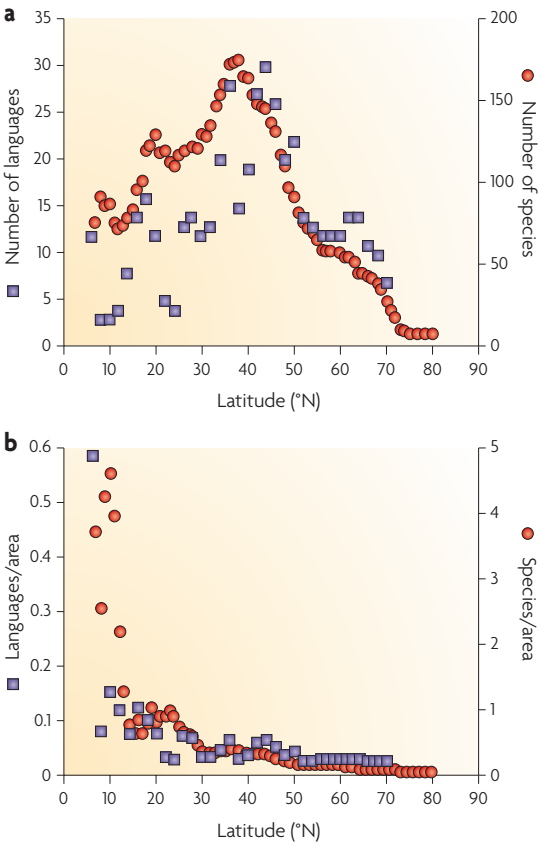


Figure 4 | **Relationships between language and species distribution.** North American human language–cultural groups (before European contact) and mammal species are distributed similarly across degrees of latitude. **a** | Numbers of languages and numbers of mammal species at each degree of north latitude in North America. The trends reflect the shape of the continent, being narrow in the south regions and growing wider at higher latitudes. Both trends peak at approximately 40°N, where North America is ~3,000 miles wide. **b** | Densities of languages and mammal species, calculated as the number of each found at the specific latitude divided by the area of the continent for a 1° latitudinal slice at each latitude. Figure and data is reproduced, with permission, from *Nature* REF. 3 © (2004) Macmillan Publishers Ltd. All rights reserved.

Gene flow is often reduced across linguistic boundaries[52], and anthropologists speculate that language may be used to advertise affinity to particular social groups[53,54]. The eighteenth century American educator Noah Webster put this view trenchantly at the time of American independence from Britain saying that "as an independent nation, our honor [*sic*] requires us to have a system of our own, in language as well as government"[55,56]. Phylogenetic trees of languages for Austronesian, Bantu and Indo-European languages all suggest that Webster was stating a general phenomenon[56]. Extant languages with a rich history of language splitting events, such as that between the speakers of American and British English, have diverged more from their ancestral languages than extant languages with fewer splitting events in their pasts. A similar pattern is observed for genetic evolution among biological species[57]. Humans seem to adjust languages at crucial times of cultural evolution, such as during the emergence of new and rival groups. Maybe there is some truth to the Babel myth after all.

## Language structure

The late eminent American comparative linguist Joseph Greenberg pioneered the study of the structural properties of languages, most famously seeking properties he called linguistic universals that could be found in all or nearly all known languages[58,59]. Languages can be classified according to structural properties of syntax, grammar and other features. Associations among these features reveal the internal structure of language and what combinations are possible. I give only the briefest treatment of this very large area, which is ripe for quantitative approaches[60,61], confining my remarks to showing how language phylogenies are fundamental to understanding how language structures evolve.

One structural feature of language is the word order of its sentences. Of the six possible orderings of subjects (S), verbs (V) and objects (O) in a sentence, two — SVO and SOV — dominate the world's languages, two others — VSO and VOS — account for ~10% of languages, and the remaining two — OSV and OVS — are rare[58,62]. From analyses of the relative frequencies of these differing orders among the world's major language families, it has been suggested that the ancestral human language was SOV[62] (M. Gell-Mann and M. Ruhlen, personal communication). One of Greenberg's best-known universals was his proposal that VSO and SVO languages use prepositional phrases to modify sentence objects, whereas SOV languages tend to use postpositional phrasing. Counts of languages support Greenberg's proposal[63].

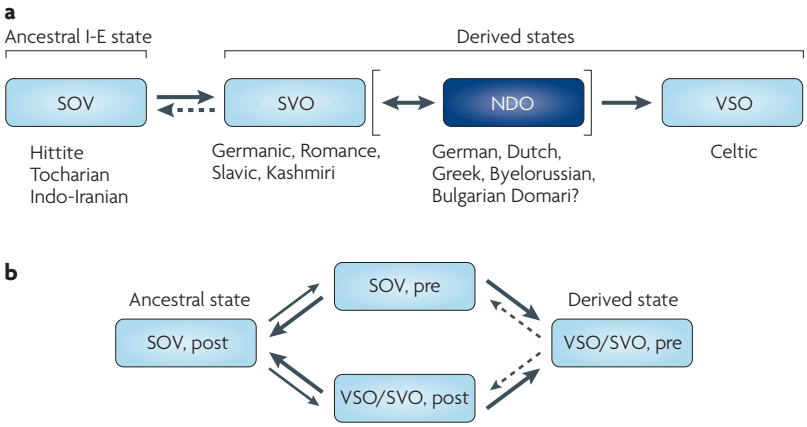


Figure 5 | **Evolution of word order changes. a** | Suggested evolution of word order changes in Indo-European (I-E) languages. Only well-supported transitions are shown. SOV, SVO and VSO refer to orderings of subject (S), verb (V) and object (O) in sentences (see FIG. 1), NDO is for languages categorized as having no dominant word order. The brackets indicate that the NDO category is questioned by some linguists, and the evolutionary relationship excluding it is represented by the light blue boxes. Statistical modelling[9,67] reconstructs ancestral Indo-European word order as SOV. SVO later evolves from SOV (dashed arrow indicates transitions back to SOV that occur in Indo-Iranian languages), and SVO gives rise to VSO. SVO and (possibly SOV) may switch to NDO in case-marked languages such as German, Latin or Greek. This broad result is predicted in REF. 62, although NDO transitions need further study. There are many possible transitions between SVO and NDO within the Indo-European tree. Uncertainty about the true Indo-European phylogeny and the models of evolution is taken into account by integrating the estimates of the model's parameters over a Bayesian sample of Indo-European trees. **b** | A diagram showing correlated evolution of word order — SOV versus VSO or SVO (VSO/SVO) including NDO — and pre versus postpositional phrasing ('pre' and 'post') to modify sentence objects (log Bayes factor test of association ~12, which indicates very strong support)[9,67]. This model reconstructs the ancestral state as SOV, post. Solid arrows indicate statistically supported evolutionary transitions, dashed arrows are not supported. Thickness conveys relative strength of the effect. Here, these arrows indicate that the derived state of VSO/SVO and prepositional phrasing evolves from the ancestral state either by adopting a different word order first, becoming SOV, pre, or by adopting a different positional phrasing first, becoming VSO/SVO, post. Examples of each evolutionary process occur in Indo-European languages (FIG. 1). These intermediate states violate Greenberg's[58] predictions but are short lived, as indicated by the thick arrows pointing to the ancestral and derived states. The derived state seems to be stable in Indo-European languages. Data is taken from the World Atlas of Linguistic Structures[71]. The evolutionary relationships shown here might change owing to a lack of consensus on the classification of some languages on these two traits (see also FIG. 1).

Do phrases such as 'I built a house' and 'I a house built' (SVO and SOV, respectively) owe their dominance to inherent properties of those systems or are they accidental winners, having ridden on the backs of people who came to dominate the globe for some other reason. Why do English speakers use the prepositional phrase in 'I built a house for you' rather than the postpositional 'I built a house you for'? Is the pairing of word order and pre versus postpositioning a chance association or does it represent co-evolution of these two structural traits? If it represents co-evolution, which feature of language changes first or can either change? These kinds of questions have direct parallels in cross-cultural studies[8,64] and in comparative biology[65], and must be studied using phylogenies. A co-evolutionary explanation would be favoured if the relationship arose independently many times in unrelated languages.

Pre versus postpositioning
Whether a language places the phrase that modifies a sentence object before (preposition) or after (postposition) that object in the sentence.

***Phylogenetic statistical approaches.*** I illustrate a phylogenetic comparative approach to the word order and positional phrasing predictions with data for the Indo-European languages (FIG. 1). Germanic, Romance and Slavic languages are mostly SVO, many Indo-Iranian languages and the ancient Tocharian and Hittite languages are SOV, and the Celtic languages are VSO. German, Greek, Bulgarian and the Indo-Iranian language Domari are among a handful of Indo-European languages sometimes regarded as having no dominant word order (NDO). The historical evolution of these four states can be studied for the Indo-European tree using the finite-state Markov model in Q, and implemented using a technique called reversible jump MCMC[66] that allows one to explore the space of possible models[9,67].

The approach outlined above reconstructs the ancestral or proto-Indo-European language as SOV (FIG. 5a). Early in Indo-European language evolution SOV gave way to SVO (or NDO, which then later resolved to SVO) before reverting to SOV in the Indo-Iranian languages. The Celtic VSO evolved from SVO in the common ancestor to the Celtic, Romance and Germanic languages. There is an intriguing hint that languages can rapidly switch between a fixed or NDO word order, perhaps using case marking in place of order. The same methodology can then be used to test Greenberg's proposal for a relationship between word order and pre versus postpositioning. Languages are scored as VSO or SVO (VSO/SVO) or as SOV, and also as prepositional or postpositional — yielding four possible combinations of paired states. The analysis finds the correlation Greenberg predicted (FIG. 5b). The analysis also shows that languages can evolve from one of Greenberg's preferred states to the other, by changing either of the individual traits first, but suggests these 'intermediate' states are unstable. These results could only have emerged from a phylogenetic analysis and should be replicated in additional families. It should be straightforward to repeat this exercise for Bantu and Austronesian languages (R. Gray and M. Dunn, personal communication).

Once many of these structural features of language have been analysed for their correlations across languages it will be possible to construct network diagrams like those used to display protein or metabolic interaction networks, in which links between pairs of features correspond to significant evolutionary correlations across species[68]. These have the potential to reveal the structural hubs and satellites of language — that is, features that are highly connected and those that are not — and the traits that are most likely to be gained or lost over time.

## Discussion

Phylogenetic and statistical methods have only begun to be used to study language evolution, but they have already returned important insights into its evolution. Much remains to be done. Models for phylogenetic inference could be improved by allowing words to alter their rates of change throughout the tree, and it should also be possible to automate cognacy judgements in a

manner analogous to automated gene sequence alignment. Greenberg's[58] proposals for linguistic universals describe dozens of associations among pairs of structural features, and these are suitable candidates for phylogenetic testing. At the level of lexicon, rather than structure, little is known about whether some words tend to change together, and whether these potential co-evolutionary linkages affect how these words evolve.

Swadesh's original vocabulary list comprises words that are used at a higher than average frequency, and so could be expanded to include less frequently used and consequently more rapidly evolving words. In a similar vein, although historically much emphasis has been placed on how words come to be replaced by new non-cognate words, there is room to study how words come to acquire new meanings, or how words gradually change their sounds while retaining their meanings and while remaining cognate. An important aspect of this process, in turn, relates the ways that languages are learned and transmitted within communities to the rates at which existing words (or other features of language) change or new words emerge and replace old ones[69]. This is analogous to attempts in evolutionary biology to describe how within-population processes give rise to differences between populations or species[70].

Language trees provide the logical backbone on which to test these and many other anthropological questions. There is no doing comparative linguistics or comparative anthropology without them, and new linguistic or anthropological research programmes should routinely make their construction a priority. Already projects such as the World Atlas of Linguistic Structures[71] or the Austronesian Basic Vocabulary Database[72] document hundreds of thousands of observations on language and these databases need to be developed in a similar way to GenBank and other genetic databases.

For geneticists, or for anyone interested in molecular evolution, the parallels between linguistic and genetic evolution should be striking, and all the more so because language is a cultural rather than a physical replicator, without built-in error correction mechanisms and potentially subject to far greater effects of borrowing and other influences that could corrupt its signal. Like genomes, the languages we observe today are the survivors of a long process of being tried out and tested by their speakers. Like genomes, we can speculate that we have retained those languages that adapted best to our minds[73], and this may be the most obvious reason why we find them easy to learn and use.

For a language system to survive it must adapt as a coherent whole, and this governs the likely combinations of language elements, be they words, grammar, syntax or morphology. These functional restraints on languages coupled with the observation of high fidelity in the transmission of linguistic elements means that there are far fewer languages and less linguistic diversity than might otherwise be possible. Are some of these languages somehow better than others or somehow better suited to their own speakers, or do existing languages represent alternative and equally functional outcomes of the linguistic evolutionary process? It is the many differences between what we see and what is possible that reveal the ways that languages adapt. How they do it and why is an area that holds great promise for furthering our understanding of this uniquely human trait as the complex and adaptively evolving system that it is[74].


1. Gordon, R. G. *Ethnologue: Languages of the World* 15th edn (SIL International, Dallas, 2005).
2. Pagel, M. in *The Evolutionary Emergence of Language* (eds Knight, C., Studdert-Kennedy, M. & Hurford, J.) 391–416 (Cambridge Univ. Press, Cambridge 2000). **An overview of linguistic diversity and how it can be studied phylogenetically and statistically.**
3. Pagel, M. & Mace, R. The cultural wealth of nations. *Nature* **428**, 275–278 (2004).
4. Darwin, C. *The Descent of Man* (Murray, London, 1871).
5. Swadesh, M. Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Am. Phil. Soc.* **96**, 453–463 (1952).
6. Embleton, Sheila M. *Statistics in Historical Linguistics. Quantitative Linguistics* Vol. 30 (Bochum, Brockmeyer, 1986).
7. Pagel, M. & A. Meade. in *Phylogenetic Methods and the Prehistory of Languages* (eds Forster, P. & Renfrew, C.) 173–182 (McDonald Institute for Archaeological Research, Cambridge, 2006).
8. Mace, R. & Pagel, M. The comparative method in anthropology. *Curr. Anthropol.* **35**, 549–564 (1994). **This paper formally introduced use of phylogenetic trees into comparative anthropology.**
9. Pagel M, Meade A. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace R., Holden C. J. & Shennan S.) 235–256 (UCL Press, London, 2005).
10. Kruskal, J., Dyen, I. & Black, P. in *Mathematics in the Archeological and Historical Sciences* (eds Hodson, F. R., Kendall, D. G. & Tautu, P.) 361–380 (Edinburgh Univ. Press, Edinburgh, 1971).
11. Sankoff, D. in *Current Trends in Linguistics 11: Diachronic, Areal and Typological Linguistics* (ed. Sebeok, T. A.) 93–112 (Mouton, The Hague, 1973).
12. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
13. Nicholls, G. K. & Gray, R. D. in *Phylogenetic Methods and the Prehistory of Languages* (eds Forster, P. & Renfrew, C.) 161–171 (McDonald Institute for Archaeological Research, Cambridge, 2006).
14. Warnow, T., Evans, S. N., Ringe, D. & Nakhleh, L. in *Phylogenetic Methods and the Prehistory of Languages* (eds Forster, P. & Renfrew, C.) 75–87 (McDonald Institute for Archaeological Research, Cambridge, 2006).
15. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
16. Edwards, A. W. E. *Likelihood* (Cambridge Univ. Press, Cambridge, 1972).
17. Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. in *Markov Chain Monte Carlo in Practice* (eds Gilks, W. R., Richardson, S. & Spiegelhalter, D. J.) 1–19 (Chapman and Hall, 1996).
18. Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314 (2001).
19. Pagel, M. in *Time-Depth in Historical Linguistics* (eds Renfrew, C., MacMahon, A. & Trask L.) 189–207 (The McDonald Institute of Archaeology, Cambridge, 2000).
20. Gray, R. & Jordan, F. Language trees support the express-train sequence of Austronesian expansion. *Nature* **405**, 1052–1055 (2000).
21. Holden, C. J. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond., B* **269**, 793–799 (2002). **This paper describes an early application of phylogenetic methods in linguistics.**
22. Holden, C. J., Meade, A. & Pagel, M. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace R., Holden C. J. & Shennan S.) 53–65 (UCL Press, London, 2005).
23. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439 (2003). **This study used language phylogeny to test a historical hypothesis for the timing of the origin of Indo-European languages.**
24. Dunn, M., Terrill, A., Reesink, G., Foley, R. A. & Levinson, S. C. Structural phylogenetics and the reconstruction of ancient language history. *Science* **309**, 2072–2075 (2005).
25. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009). **This paper describes the use of a language phylogeny to test a historical hypothesis for the timing of the origin of Austronesian languages.**
26. Pagel, M., Atkinson, Q. D. & Meade, A. Frequency of word use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–719 (2007). **A statistical phylogenetic study that proposed a general explanation for variation in rates of lexical replacement.**
27. Sanderson, M. J. & Donoghue, M. J. Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795 (1989).
28. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
29. Bryant, D., Filimon, F. & Gray, R. D. in *The Evolution of Cultural Diversity: Phylogenetic Approaches* (eds Mace, R., Holden, C. & Shennan, S.) 69–85 (UCL Press, London, 2005).
30. Renfrew, C. *Archaeology and Language: the Puzzle of Indo-European Origins* (Cape, London, 1987). **Classic text on the origin of the Indo-European language family.**
31. Greenhill, S., Currie, T. & Gray. R. Does horizontal transmission invalidate cultural phylogenies? *Proc. R. Soc. Lond., B* 18 Mar 2009 (doi:rspb.2008.1944).

32. Cavalli-Sforza, L. L., Piazza, A., Menozzi, P. & Mountain, J. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc. Natl Acad. Sci. USA* **85**, 6002–6006 (1988).
**This paper is a widely cited early attempt to link genetic and linguistic diversity.**
33. Lansing, J. S. *et al.* Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl Acad. Sci. USA* **104**,16022–16026 (2007).
34. Hunley, K. *et al.* Genetic and linguistic coevolution in Northern Island Melanesia. *PLoS Genet*. **4**, 1–14 (2008).
35. Dediu, D. & Ladd, D. R. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and microcephalin. *Proc. Natl Acad. Sci. USA* **104**, 10944–10949 (2007).
36. Holden, C. J. & Mace, R. Spread of cattle led to the loss of matriliny in Africa: a co-evolutionary analysis. *Proc. R. Soc. Lond., B* **270**, 2425–2433 (2003).
**A good example of the use of language trees to study cultural evolution.**
37. Fortunato, L., Holden, C. J. & Mace, R. From bridewealth to dowry? A Bayesian estimation of ancestral states of marriage transfers in Indo-European groups. *Human Nature* **17**, 355–376 (2006).
38. Mace, R. & Jordan, F. in *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace, R., Holden, C. & Shennan, S.) 207–216 (UCL Press, London, 2005).
39. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc. Natl Acad. Sci. USA* **104**, 3736–3741 (2007).
40. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* **39**, 306–314 (1994).
41. Zipf, G. K. *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Reading, Massachusetts, 1949).
42. Leech, G., Rayson, P. & Wilson, A. *Word Frequencies in Written and Spoken English: Based on the British National Corpus* (Longman, London, 2001).
43. Zipf, G. K. Prehistoric 'cultural strata' in the evolution of Germanic: the case of Gothic. *Mod. Lang. Notes* **62**, 522–530 (1947).
44. Francis, W. N., Kuçera, H. & Mackie, A. W. *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, Boston, 1982).
45. Lieberman, E., Michel, J.-B., Jackson, J., Tang, T. & Nowak, M. A. Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716 (2007).
46. Starostin, S. A. in *Works on Linguistics* (ed. Starostin, S. A.) 827–839 (Languages of the Slavic Culture, Moscow, 2007).
47. Ellis, N. C. Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Stud. Second Lang. Acquisit.* **24**, 143–188 (2002).
48. Huettig, F., Quinlan, P. T., McDonald, S. A. & Altmann, G. T. M. Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol.* **121**, 65–80 (2006).
49. Birdsell, J. B. Some environmental and cultural factors influencing the structuring of Australian Aboriginal populations. *Am. Nat.* **87**, 171–207 (1953).
50. Nichols, J. *Linguistic Diversity in Space and Time* (Univ. of Chicago Press, Chicago, 1992).
51. Mace, R. & Pagel, M. A latitudinal gradient in the density of human languages in North America. *Proc. Roy. Soc. Lond., B* **261**, 117–121 (1995).
52. Barbujani, G. & Sokal, R. R. Zones of sharp genetic change in Europe are also linguistic boundaries. *Proc. Natl Acad. Sci. USA* **87**, 1816–1819 (1990).
53. Labov, W. *Principles of Linguistic Change: Social Factors* (Blackwell, Oxford, 2001).
54. Milroy, J. & Milroy, L. Linguistic change, social network and speaker innovation. *J. Linguist* **21**, 229–284 (1985).
55. Webster, N. *Dissertations on the English Language* (Isaiah Thomas, Boston, 1789).
56. Atkinson, Q., Meade, A., Venditti, C., Greenhill, S. & Pagel, M. Languages evolve in punctuational bursts. *Science* **319**, 588 (2008).
57. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).
58. Greenberg, J. H. (ed.) *Universals of Languages* (MIT Press, Cambridge, Massachusetts, 1963).
59. Kirby, S. *Function, Selection, and Innateness: the Emergence of Language Universals* (Oxford Univ. Press, Oxford, 1999).
60. Cysouw, M. in *Quantitative Linguistics: An International Handbook* (eds Altmann, G., Köhler, R. & Piotrowski, R.) 554–578 (Mouton de Gruyter, Berlin, 2005).
61. Croft, W. *Explaining Language Change: an Evolutionary Approach* (Longman, Harlow, 2000).
**This text provides a good overview of evolutionary thinking about language.**
62. Newmeyer, F. J. in *The Evolutionary Emergence of Language* (eds Knight, C., Studdert-Kennedy, M. & Hurford, J.) 372–388 (Cambridge Univ. Press, Cambridge, 2000).
63. Haspelmath, M. & Siegmund, S. Simulating the replication of some of Greenberg's word order predictions. *Linguistic Typology* **10**, 74–82 (2006).
64. Mace, R. *The Evolution of Cultural Diversity: a Phylogenetic Approach* (eds Mace R., Holden C. J. & Shennan S.) 1–10 (UCL Press, London, 2005).
65. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, Oxford, 1991).
66. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).
67. Pagel, M. & Meade, A. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* **167**, 808–825 (2006).
68. Pagel, M., Meade, A. & Scott, D. Assembly rules for protein interaction networks. *BMC Evol. Biol.* **7** (Suppl. 1), S16 (2007).
69. Niyogi, P. *The Computational Nature of Language Learning and Evolution* (MIT Press, Cambridge, Massachusetts, 2006).
70. Mangel, M. & Clark, C. W. *Dynamic Modeling in Behavioral Ecology* (Princeton Univ. Press, Princeton, New Jersey, 1988).
71. Haspelmath, M., Dryer M. S., Gil, D. & Comrie, B. (eds) The World Atlas of Linguistic Structures *Max Planck Digital Library* [online], www.wals.info (2008).
72. Greenhill, S. J., Blust, R. & Gray, R. D. The Austronesian Basic Vocabulary Database: from bioinformatics to lexomics. *Evol. Bioinform. Online* **4**, 271–283 (2008).
73. Christiansen, M. H. & Chater, N. Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–558 (2008).
74. Gell-Mann, M. *The Quark and the Jaguar: Adventures in the Simple and Complex* (W.H. Freeman New York, 1994).



### Acknowledgements
I thank C. Venditti, A. Calude, I. Peiros, A. Meade, Q. Atkinson, M. Ruhlen, M. Cysouw and M. Haspelmath for help, comments and suggestions. Grants to M.P. from the Leverhulme Trust and the Natural Environment Research Council supported this work.


### FURTHER INFORMATION
**Mark Pagel's homepage:** www.evolution.reading.ac.uk
**Austronesian Basic Vocabulary Database:** http://language.psy.auckland.ac.nz/austronesian
**SplitsTree:** http://www.splitstree.org
**Swadesh list:** http://en.wiktionary.org/wiki/Appendix:Swadesh_list
**World Atlas of Linguistic Structures:** http://www.wals.info

**ALL LINKS ARE ACTIVE IN THE ONLINE PDF**