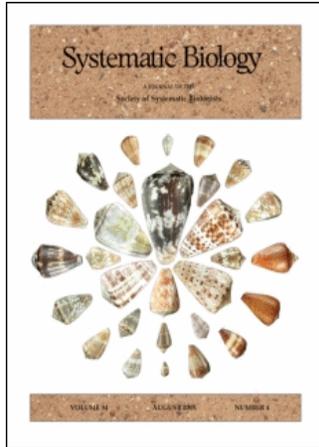


This article was downloaded by:[Max Planck Inst & Research Groups Consortium]
On: 22 March 2008
Access Details: [subscription number 769611911]
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Systematic Biology

Publication details, including instructions for authors and subscription information:
<http://www.informaworld.com/smpp/title~content=t713658732>

Curious Parallels and Curious Connections - Phylogenetic Thinking in Biology and Historical Linguistics

Quentin D. Atkinson^a; Russell D. Gray^a

^a Department of Psychology, University of Auckland, Auckland, New Zealand

First Published on: 01 August 2005

To cite this Article: Atkinson, Quentin D. and Gray, Russell D. (2005) 'Curious
Parallels and Curious Connections - Phylogenetic Thinking in Biology and Historical
Linguistics', Systematic Biology, 54:4, 513 - 526

To link to this article: DOI: 10.1080/10635150590950317

URL: <http://dx.doi.org/10.1080/10635150590950317>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Invited Historical Essay

Syst. Biol. 54(4):513–526, 2005
 Copyright © Society of Systematic Biologists
 ISSN: 1063-5157 print / 1076-836X online
 DOI: 10.1080/10635150590950317

Curious Parallels and Curious Connections—Phylogenetic Thinking in Biology and Historical Linguistics

QUENTIN D. ATKINSON AND RUSSELL D. GRAY

Department of Psychology, University of Auckland, Private Bag 92019, Auckland 1020, New Zealand; E-mail: rd.gray@auckland.ac.nz (R.D.G.)

Abstract.—In *The Descent of Man* (1871), Darwin observed “curious parallels” between the processes of biological and linguistic evolution. These parallels mean that evolutionary biologists and historical linguists seek answers to similar questions and face similar problems. As a result, the theory and methodology of the two disciplines have evolved in remarkably similar ways. In addition to Darwin’s curious parallels of process, there are a number of equally curious parallels and connections between the development of methods in biology and historical linguistics. Here we briefly review the parallels between biological and linguistic evolution and contrast the historical development of phylogenetic methods in the two disciplines. We then look at a number of recent studies that have applied phylogenetic methods to language data and outline some current problems shared by the two fields. [Comparative method; Darwin; evolution; historical linguistics; phylogeny; Schleicher.]

CURIOUS PARALLELS IN THE DOCUMENTS OF EVOLUTIONARY HISTORY

In *The Descent of Man*, Darwin (1871) noted that the process of evolution is not limited to just the biological realm.

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel . . . We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation.” (pp. 89–90)

Many of the fundamental features of biological and linguistic evolution are demonstrably analogous (see Table 1 and Croft, 2000). Just as DNA sequences contain discrete heritable units, so too do languages in their grammatical and phonological structures and their vocabularies (lexicons). These may differ from language to language and can be inherited as the languages are learned by subsequent generations. With knowledge of the processes of linguistic change, it is possible to identify homologous linguistic characters that, like homologous biological structures, indicate inheritance from a common ancestor. For example, homologous words, or *cognates*, meaning “water” exist in English (*water*), German (*wasser*), Swedish (*vatten*), and Gothic (*wato*), reflecting descent from proto-Germanic (**water* [in historical linguistics, inferred *proto forms* are denoted with a ‘*’]). Cognates are words of similar meaning with systematic sound correspondences indicating they were related due to common ancestry. Processes of mutation and random drift can operate on linguistic characters, just as they do on genes. An example of lexical mutation, or *innovation* as it is known in linguistics, is the word *boy*, which arose at some point after English split off from the other

west Germanic languages (Campbell, 2004). A phonological example is the unconditional sound change of /t/ to /k/ in Hawaiian. So the ancestral Polynesian word **tapu*, “forbidden,” changed to *kapu*, and **tolu*, “three,” changed to *kolu*, and so on (Crowley, 1992). As well as these “point mutations,” words, like gene sequences, can show insertions (e.g., Old Swedish **bökr*, “books,” to *böker*; Campbell, 2004), deletions (e.g., Proto-Oceanic **tupa*, “derris root,” to Selau *tua*; Blust, 2003) and reversals, or *metathesis* (e.g., Old English *brid* to modern English *bird*). Linguistic changes, like changes in biological form, are also sometimes structurally and/or functionally linked. Terms for “five,” for instance, tend to be correlated with terms for “hand” for obvious reasons. As in biology, mutation and drift create variation that may be subject to selection. For example, Pawley and Syder (1983) found evidence that Darwinian selection pressures have acted on English syntax. Specifically, they identified differences between vernacular and literary English grammar that they argue were “adaptive to the particular conditions imposed by the mode of language use” (p. 577). The fundamental process of language formation involves cladogenesis, where a single lineage splits to form two new languages. Often, as in biology, this is due to geographic separation or migration events. Horizontal gene transfer and hybridization also have a linguistic equivalent in borrowing between languages. For example, the English word *mountain* is borrowed from French, *montagne*. Borrowing between languages can produce reticulated evolution in a similar fashion to horizontal gene transfer in plants or bacteria. Extreme cases of contact between languages can produce a form of language “hybrid,” as in the case of some *Creoles*. For example, Sranan, a Creole spoken in Surinam,

TABLE 1. Conceptual parallels between biological and linguistic evolution

Biological evolution	Linguistic evolution
Discrete characters	Lexicon, syntax, and phonology
Homologies	Cognates
Mutation	Innovation
Drift	Drift
Natural selection	Social selection
Cladogenesis	Lineage splits
Horizontal gene transfer	Borrowing
Plant hybrids	Language Creoles
Correlated genotypes/phenotypes	Correlated cultural terms
Geographic clines	Dialects/dialect chains
Fossils	Ancient texts
Extinction	Language death

has elements of English, Dutch, Portuguese, and a number of African and Indian languages, although it is essentially English based. Many Creoles involve less mixing, and are instead cut-down versions of the language upon which they are based. One current view is that this is partly the result of a linguistic "founder effect," where the complexity of the original (usually plantation) language declines due to the small initial population and is revived with the arrival in greater numbers of speakers from other linguistic backgrounds. Lastly, like species, languages can become extinct and can even be "fossilized" in the form of ancient texts. For example, we have archaic manuscripts of ancient languages like Hittite, Homeric Greek, Sanskrit, and Mayan. A more detailed discussion of the parallels between biological and linguistic evolution, which builds on Hull's (1988) general account of evolutionary processes, can be found in Croft (2000).

In 1965 Zuckerkandl and Pauling characterized molecules as "documents of evolutionary history":

"Of all natural systems, living matter is the one which, in the face of great transformations, preserves inscribed in its organization the largest amount of its own past history." (p. 357)

Languages are also documents of evolutionary history. By comparing features between languages, linguists can make inferences about their historical relationships and thus gain insight into human prehistory. In attempting to make historical inferences, linguists and biologists must ask similar questions: What is the most reliable type of data? What is the probability of different types of change? Does a tree accurately reflect the evolutionary history or is some other representation more appropriate? Where phylogenies are of interest, which trees are best, and how confident can we be in a particular result? It is perhaps not surprising then that biologists and linguists have developed similar phylogenetic methods to answer these questions (see, for example, Platnick and Cameron, 1977, and O'Hara, 1996). In fact, the theory and methodology of evolutionary biology and historical linguistics have evolved along related paths throughout their history. In the following discussion, we will examine the curious parallels and connections between the history of phylogenetic thinking in Western biology and linguistics.

We note that this article is not intended to be a comprehensive description of the history of either field—a large amount of literature already exists on the development of historical linguistics (e.g., Pedersen, 1931; Robins, 1997; Campbell, 2000) and evolutionary biology (e.g., Gould, 2002; Mayr, 1982). Instead, we present a brief comparative history highlighting some of the intriguing interdisciplinary relationships between the two fields.

IN THE BEGINNING—TWO ANCIENT GREEK OBSESSIONS

The Ancient Greek philosophers were obsessed with archetypal form, especially the definition and classification of plants and animals according to these archetypes. Plato's, perhaps offhand, definition of man as a featherless biped was famously rebuffed when the cynic Diogenes presented him with a plucked chicken. Plato's student, Aristotle, was more methodical and was the first to use a hierarchical system of animal classification based on discrete characters (Thompson, 1913). Although he did not outline his classification system formally, Aristotle is generally considered the founder of the comparative method in biology and many of his groupings still hold today (Mayr, 1982; Thompson, 1913). He identified the birds, amphibians, fish, and mammals (with the exception of whales, which he placed in their own group) as distinct groups and part of a broader group of animals "with blood." He also identified "bloodless" cephalopods, higher crustaceans, insects, and the "lower animals." The Greek preoccupation with classification facilitated a hierarchical conception of the natural world, which was a precursor to modern phylogenetic classification. However, these groupings were presented as immutable archetypes—they had always existed and always would exist. Aristotle's classification was thus ahistorical. Although the Greeks identified biological "kinds," they had no concept of species existing through time. The prevailing view in Aristotle's time, a view that persisted up until the 17th century, was that organisms could arise through spontaneous generation from nonliving matter. As a result, questions about species lineages and historical relationships never arose.

In contrast, when the Ancient Greek philosophers turned to language, they were obsessed, not with hierarchical relationships, but with explaining the process of change in linguistic structures. The writings of Homer (ca. 730 B.C.) were of fundamental importance in Greek schooling. By the time of Socrates (469–399 B.C.) it was evident that the Greek language had changed considerably since Homer. The study of language was largely oriented towards keeping tabs on these changes in Greek (Campbell, 2000). Socrates sought to relate contemporary words to *prota onomata*, or "first words," the primordial words of some ancient Greek tongue (Percival, 1987). Socrates believed that change was a process of decay and claimed that contemporary words were the product of semantic shift and phonetic degeneration from the *prota onomata*. In Plato's *Republic*, Socrates argued that

"... the primeval words have already been buried by people who wanted to embellish them by adding and removing letters to make

them sound better, and disfiguring them totally, either for aesthetic considerations or as a result of the passage of time." 414C

As well as laying the foundations for modern biology, Aristotle advanced the Greek understanding of grammatical categories and developed ideas about the nature of linguistic change. He identified four key types of linguistic change that will be familiar to any biologist today—insertion, deletion, transposition, and substitution (*Categories* 15a13 and *On Coming-to-be and Passing-away* 314b27, cited in Householder, 1981).

Unfortunately, the Greek inquiry into language change was very Greco-centric—the Greeks were chiefly concerned with Greek (Percival, 1987). Ideas about how Greek etymology and grammar may have been related to other languages were thus not put forward. Change in the Greek language was merely a process of decay, as the language shifted from some ancestral ideal. Thus, in biology the Ancient Greeks had a hierarchical classification system but no notion of change, while in linguistics they understood something of the process of change (albeit involving decay from an ideal) but were not interested in language relationships.

It is worth noting that linguistic traditions in the East, including China, Mesopotamia, and India, stretch back far earlier than the Greek tradition, and were in many respects more advanced at this time (Robins, 1997). Eastern linguistics had a greater focus on grammar and phonetics than etymology (Pedersen, 1931). For example, the work of the Indian linguist, Panini, circa 5th century B.C., comprised a detailed grammar of Sanskrit. However, the Eastern scholars, like the Greeks, were chiefly interested in describing the particulars of their own language and its change from an archaic form. This, combined with very little contact between Eastern and Western linguists meant that Eastern linguistics had little impact on the early development of the Western linguistic tradition (Robins, 1997).

BEFORE THERE WERE TREES

During the early Christian era and the Middle Ages, the still unchallenged Ancient Greek conception of constantly regenerating "natural kinds," combined with creationist accounts of the origin of life, made positing historical species relationships both illogical and heretical. In linguistics, creationism promoted some historical thinking—for example, how to get peoples/languages aligned with the descendants of Noah. Unfortunately, the Old Testament account of the creation of all languages following the destruction of the Tower of Babel undermined any attempt to accurately infer the historical relationships between languages. St. Augustine (354–430 A.D.; *City of God*, 413–426/427 A.D.) was the first in a long tradition of intellectuals who attempted, sometimes quite creatively, to integrate science and scripture by relating all languages to Hebrew (Percival, 1987)—something akin to relating all species to the woolly mammoth. It was not until the 17th century and the "Age of Reason" that scholars in both disciplines began to look critically at the accounts offered by scripture.

In linguistics, the effects of this revolution were realized more quickly than in biology. One interesting reason for this was the invention and increasing use of the printing press. The printing press greatly increased the accessibility and quantity of raw material describing foreign languages (Pedersen, 1931). One might even draw an analogy between this boom and the current proliferation of sequence data in genetics. Where previously, linguists may have only been exposed to their own and neighboring languages, Latin and perhaps some Greek, by the 17th century most scholars had access to Greek texts and many of the languages of Europe and the near East, as well as the newly discovered languages of the Americas (Pedersen, 1931). This proliferation of material increased interest in language comparison. In short, linguistics became oriented towards classification.

One of the first to challenge the idea of a Hebraic root to the languages of Europe was J. J. Scaliger (1540–1609). Scaliger was able to identify Greek, Germanic, Romance, and Slavic language groups by comparing the word for *God* between a number of European languages (Pedersen, 1931). He understood homologous characters as reflecting descent from parent to daughter languages and recognized their importance in reconstructing language relationships. Scaliger failed to find (or chose to ignore) any relationships between the main groups and so his explanations were still essentially ahistorical (Pedersen, 1931). However, his work, combined with the existing knowledge of linguistic structure and processes of change, provided the raw ingredients for the comparative method in linguistics.

During this time, the biological comparative method also continued to build on the work of Aristotle. Leonardo da Vinci's (1452–1519) detailed anatomical studies compared humans to other species and recognized structural homologies (although most scholars maintained a purely anthropocentric interest in anatomy until the 18th century). Carolus Linnaeus (1707–1778), "the father of modern taxonomy," introduced a hierarchical classification system using precise species descriptions and Georges Cuvier (1769–1832) led the shift from anthropocentric anatomy to comparing anatomical structures between species (Mayr, 1982). Another significant milestone was the work of English naturalist John Ray (1628–1705). He was one of the first to suggest that "species" existed through time and were not simply the result of spontaneously generating organisms of various kinds. This discovery, which was initially rejected as heretical, made historical explanations of species diversity possible for the first time.

An important figure in the development of both linguistic and biological theory before the 19th century was the philosopher Gottfried W. von Leibniz (1646–1716). Leibniz was parodied as Dr. Pangloss in Voltaire's *Candide* for his belief that the world must be the best of all possible worlds because God had created it. More recently, the "Panglossian paradigm" was revisited in the famous Gould and Lewontin (1979) critique of adaptationism in biology. However, despite his theological idealism, Leibniz advocated a dynamic conceptualization

his notebook depicting an evolutionary tree. Although Darwin was not the first to use an evolutionary tree (Lamarck, for example, included a rudimentary tree in his 1809 *Philosophie Zoologique*), the *Origin of Species* elegantly linked affinity between species with proximity of descent, making tree diagrams historically meaningful and useful as explanations of the natural world (Mayr, 1982).

In the spirit of the Enlightenment, Darwin did not restrict himself to speculation about biological evolution. In the *Origin of Species* (1859) he muses on the topic of linguistic evolution:

"If we possessed a perfect pedigree of mankind, a genealogical arrangement of the races of man would afford the best classification of the various languages now spoken throughout the world; and if all extinct languages, and all intermediate and slowly changing dialects, were to be included, such an arrangement would be the only possible one. Yet it might be that some ancient languages had altered very little and had given rise to few new languages, whilst others had altered much owing to the spreading, isolation, and state of civilisation of the several co-descended races, and had thus given rise to many new dialects and languages. The various degrees of difference between the languages of the same stock, would have to be expressed by groups subordinate to groups; but the proper or even the only possible arrangement would still be genealogical; and this would be strictly natural, as it would connect together all languages, extinct and recent, by the closest affinities, and would give the filiation and origin of each tongue." (Chap. 13, p. 422)

It is tempting to credit Darwin with the subsequent proliferation of language trees in linguistics. Certainly, it seems likely that some borrowing of ideas may have occurred between biology and linguistics at this time. During the first half of the 19th century, a string of influential linguists made reference to botany, comparative anatomy, and physiology, including Franz Bopp, Jacob Grimm, Rasmus Rask, and Friedrich Schlegel (Koerner, 1983). In 1863, just 4 years after the *Origin of Species* was first published, the linguist August Schleicher (1821–1868) published a paper depicting an Indo-European language tree entitled *Die Darwinische Theorie und die Sprachwissenschaft* (see Fig. 2). The English translation was published in 1869 under the title *Darwinism Tested by the Science of Language*. Schleicher wrote the paper as an open letter to a friend, the biologist and committed Darwinian, Ernst Haeckel (1834–1919). Haeckel introduced Schleicher to the *Origin of Species* in 1863 and the linguist was evidently well informed in biology and the discourse surrounding Darwinian theory (Maher, 1983). However, in his paper, Schleicher (1863) pointed out that a "family tree" approach had been part of linguistics since well before Darwin:

"First, as regards Darwin's assertion that species change in course of time, a process repeated time and again which results in one form arising from another, this same process has long been generally assumed for linguistic organisms. . . . We set up family trees of languages known to us in precisely the same way as Darwin has attempted to do for plant and animal species." (p. 7)

Actually, Schleicher had used language trees, or *stammbaum*, in two 1853 publications, some 6 years before the *Origin of Species* was first published (Koerner, 1983). Koerner credits the idea of the genealogical language tree to Friedrich Schlegel (1772–1829), who introduced

a "stammbaum" approach in an 1808 publication on comparative grammar (Schlegel, 1808). Schlegel, however, drew ideas from biology and comparative anatomy, where tree diagrams had been used to represent classification systems, and this may have contributed to his choosing a family tree approach to represent language relationships. The matter is further complicated by the contemporaneous publication of tree diagrams, or *stemma*, by philologists studying manuscript evolution. The first published manuscript phylogeny (see Fig. 3) was drawn by Carl Johan Schlyter (1747–1805) in 1827 and, as O'Hara (1996) points out, the idea of establishing the most authentic version of a text by reconstructing its ancestry may have been part of the monastic tradition for much longer. Although there could well have been some interdisciplinary cross-fertilization, Darwinian ideas of descent with modification were less revolutionary in linguistics than they were in biology. Phylogenetic understanding and methodology in linguistics had already developed rapidly before Darwin, and this continued throughout the 19th century.

In 1822, Jacob Grimm (1785–1863), perhaps more widely known as half of the *brothers Grimm* of fairytale fame, introduced *Grimm's Law*, a series of sound changes in Germanic. Grimm formulated rules for these sound changes seen in cognate words of related languages (for example a change of original *p* to *f* in Germanic languages, as illustrated, for example, by Latin *pater*, "father," and English *father*, where Latin, not being a Germanic language, did not change the *p*, but English, a Germanic language, did change it to *f*). Later, the *Neogrammarians* identified a host of similar patterns and argued that all sound change was governed by regular sound laws. A passionate debate ensued about whether sound change was absolutely regular. Although initially controversial, this form of linguistic uniformitarianism was generally accepted by the end of the 19th century. This debate in linguistics occurred contemporaneously with debates over the relative merits of uniformitarianism in geology and biology. One of the most lively of these debates was between the geologist Charles Lyell (1797–1875), the physicist Lord Kelvin (1824–1907), and Charles Darwin over the age of the earth. In a letter written in 1837 to his sister, Darwin notes a linguistic argument put forward for the age of the earth by John F. W. Herschel (1792–1871) (Smith and Burkhardt, 1985). Herschel expresses his ideas in an 1836 letter to Charles Lyell (published in Cannon, 1961):

"[W]hen we see what amount of change 2000 years has been able to produce in the languages of Greece and Italy or 1000 in those of Germany, France and Spain we naturally begin to ask how long a period must have lapsed since the Chinese, the Hebrew, the Delaware and the Malesass [Malagasy] had a point in common with the German and Italian and each other.—Time! Time! Time!—we must not impugn the Scripture Chronology, but we *must* interpret it in accordance with whatever shall appear on fair enquiry to be the truth for there cannot be two truths."

The Neogrammarians' principle that "sound laws suffer no exceptions" eventually won out with the publication in 1878 of the "neogrammarian manifesto" by Karl Brugmann (1849–1919) and Hermann Osthoff

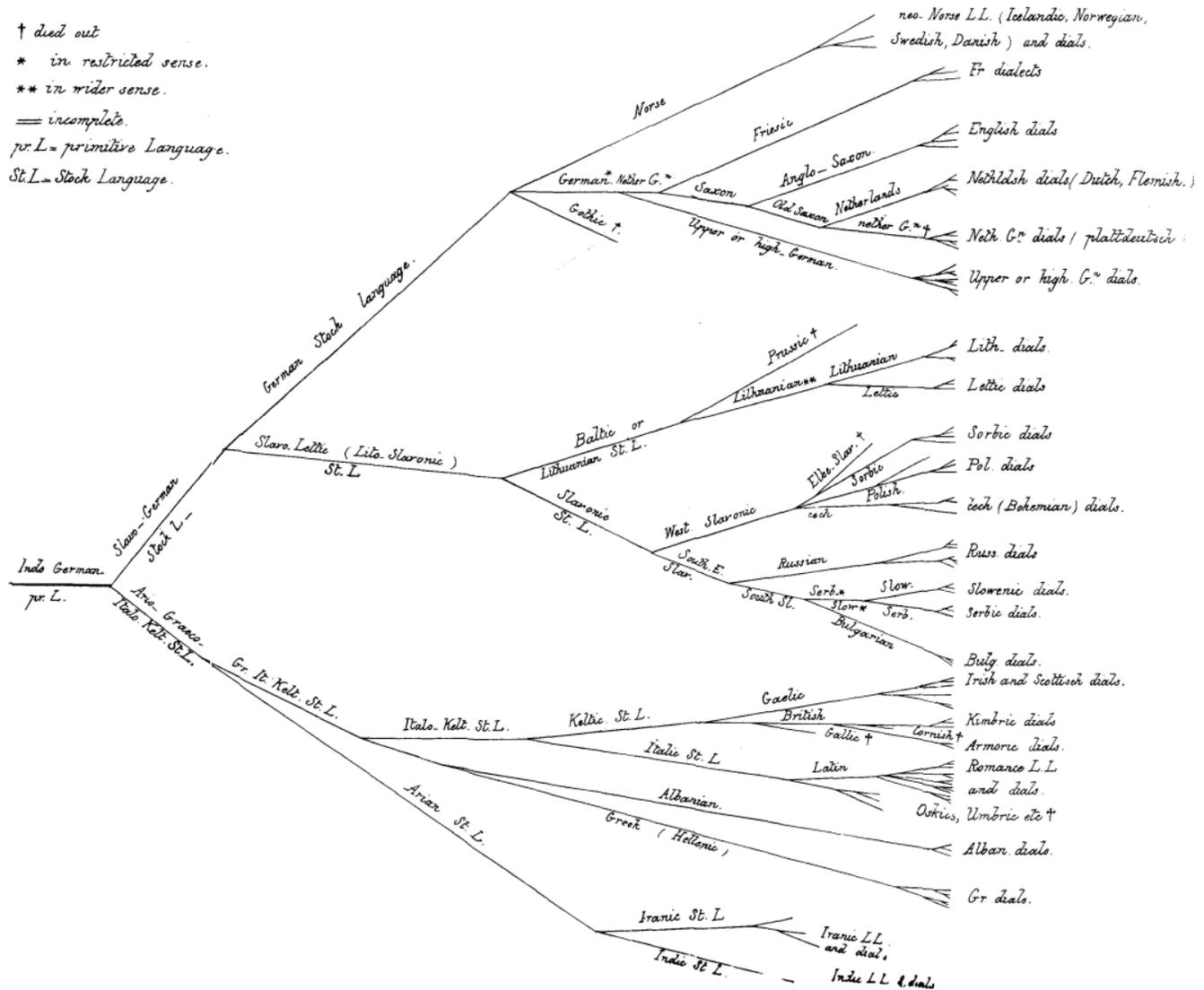


FIGURE 2. August Schleicher's (1863) Indo-European *Stammbaum*, or language family tree, from *Die Darwinische Theorie und die Sprachwissenschaft* (*The Darwinian Theory and the Science of Language*). (Schleicher, A. 1983. Darwinism tested by the science of language [A. Bikkers, trans.]. John Benjamins Publishing Co., Amsterdam [original work published in 1863], by permission of John Benjamins Publishing Co.)

(1847–1909). The comparative method in linguistics developed gradually from the end of the 17th century and was perfected with the Neogrammarians. It is a method designed to compare related languages and, on the basis of shared materials, to postulate, or “reconstruct,” the sounds, words, and structures of the parent language from which the related languages descend. It is the most commonly used method for inferring language relationships.

In 1884, the comparative method was further advanced when Brugmann made the important distinction between “innovations” and “retentions” (Hoenigswald, 1990). Innovations are shared characters that were not present in the ancestral form, while retentions are shared characters inherited from a common ancestor. Brugmann realized that shared innovations are much more informative for phylogenetic classification than shared reten-

tions. The criterion of shared innovations is now central to working out the family-tree classification of related languages using the comparative method. This distinction was made in biology some 70 years later, in 1950, when Willi Hennig (1913–1976) differentiated symplesiomorphies (shared retentions) from synapomorphies (shared innovations). Despite the methodological similarities between the comparative method in linguistics and biological cladistics, up until very recently, historical linguists did not typically use computer algorithms to search for the best language tree(s). This is surprising given that the task of finding optimal trees for even a moderate number of languages is one of considerable computational complexity (Swofford et al., 1996). Some attempts have been made, however, to formalize the criterion implicit in the method (e.g., Gleason, 1959; Hoenigswald, 1960). Thomason and Kaufman (1988)

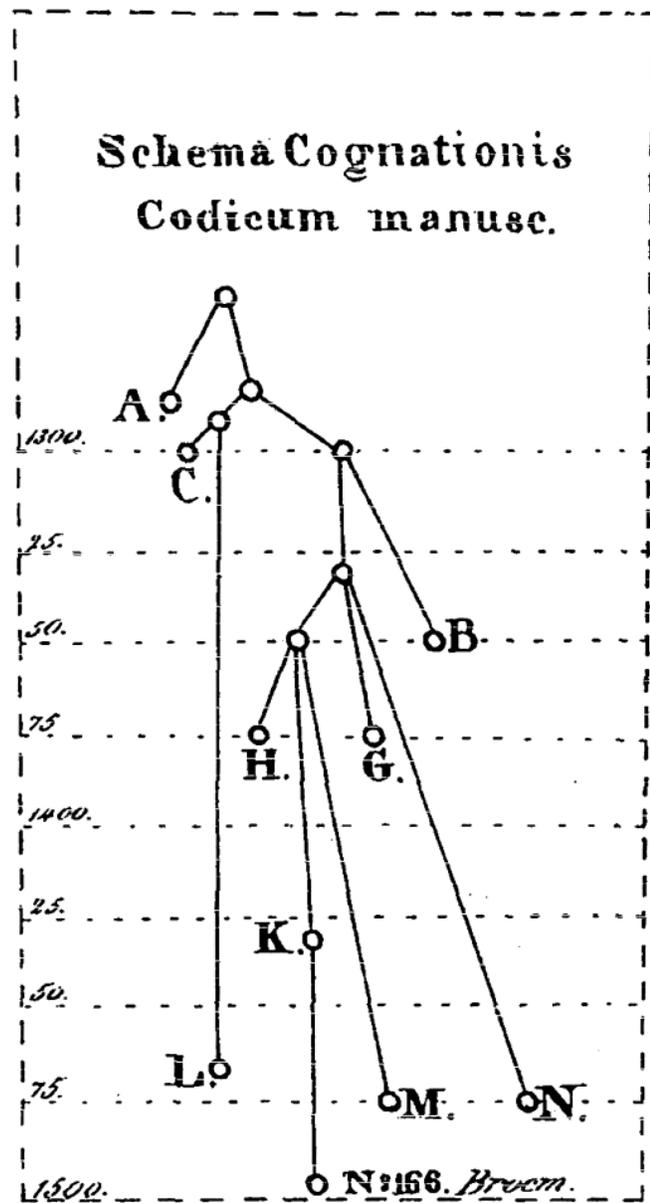


FIGURE 3. Carl Schlyter's 1827 manuscript phylogeny, or *stemma*, showing the relationships between copies of the Västgöta Law (Collin and Schlyter, 1827).

identified six steps used in the comparative method to demonstrate genetic relationships between languages: (1) determining phonological (sound) correspondences in words of the same or related meaning; (2) establishing phonological systems; (3) establishing grammatical correspondences; (4) reconstructing grammatical systems; (5) identifying subgroups of languages; and (6) producing a model of diversification.

Central to the comparative method is the determination of systematic sound correspondences among related languages in order to reconstruct ancestral sounds and hence the most likely series of phonological innovations that led to the attested sounds in the related languages.

Exclusively shared innovations are used to infer historical relationships and construct a language family tree. To return to Grimm's Law as an example, the Germanic language family is characterized by a sound change of initial **p* in the ancestral Proto-Indo-European to *f* in Proto-Germanic. So, the *p*, preserved in its unchanged form in Latin *pater*, Greek *pater*, and Sanskrit *pitár*, was replaced by an *f* in Proto-Germanic, giving rise to English *father* and Gothic *fadar*, and so on, in the other Germanic languages (Campbell, 2004). This pattern occurs repeatedly between cognate words in the Germanic languages and their cognates in non-Germanic Indo-European languages (e.g., English *foot* and Greek *podos*). The comparative method also involves the reconstruction of ancestral words in reconstructed (hypothetical) proto-languages (analogous to ancestral character states in biology). For example, based on the types of patterns highlighted above, linguists can reconstruct with relative certainty the Proto-Indo-European word **ped* for "foot" (Campbell, 2004). Ancestral reconstructions can sometimes be checked against historically attested ancient languages. For instance, reconstructions based on the comparison of Romance languages (descended from Latin) can often be checked against attested forms in Latin documents.

Some 19th century scholars opposed the Stammbaum (family tree) model of language evolution. The most often cited (although he was not the first) is Schleicher's student, Johannes Schmidt (1843–1901), who proposed a wave theory of language change (Schmidt, 1872). Schmidt studied dialects and thus became aware of the extent of borrowing between neighboring populations. The wave theory proposed that language change spread in waves emanating from some epicenter. This process could happen repeatedly and from varying epicenters. As a result, in any given language different words could have different histories. The wave theory was therefore often seen as challenging the neogrammarian conception of language change, though today the two are seen to complement one another, with both needed to get the full history of language (i.e., determine what is inherited and what is diffused). Schmidt's model bears an obvious resemblance to the demic diffusion models used more recently in biology, for example, by Menozzi et al. (1978).

AND THEN THERE WERE ALGORITHMS ...

During the first half of the 20th century, another major methodological revolution occurred in biology. The work of Ronald Fischer (1930) and Theodosius Dobzhansky (1937), and later Sir Julian Huxley, Ernst Mayr, and George G. Simpson, allowed Darwinian evolutionary theory to be explained in terms of Mendelian characters of inheritance and population genetics, thus giving rise to the modern synthesis. In 1953, Watson and Crick announced the structure of DNA and by the beginning of the 1960s the first protein sequences had been published. This produced much more data than could be analyzed by inspection, and biologists began working on numerical, algorithmic methods of inferring phylogenies.

Felsenstein (2004) provides an excellent overview of the development of numerical phylogenetic methods, beginning with the distance methods of Sokal and Sneath in the late 1950s (Michener and Sokal, 1957; Sneath 1957).

Five years earlier, however, Morris Swadesh introduced similar distance methods into linguistics. Swadesh's (1952, 1955) approach, known as lexicostatistics, used the percentage of shared cognates between languages to produce a pairwise distance matrix. These distance matrices were then analyzed using clustering algorithms to infer tree topologies. Swadesh (1952) also introduced *glottochronology*, based on the idea of a *glotto-clock*, or constant rate of lexical replacement. He realized that, under the assumption of constant rates, one could also infer divergence times from the distance data. A decade later, Zuckerkandl and Pauling (1962) introduced the idea of the molecular clock to infer species divergence times in biology. This is, however, almost certainly an example of convergent evolution, not borrowing. Zuckerkandl and Pauling's molecular clock proposal evolved from earlier work in biology linking the distance between species to variation found in hemoglobin (e.g., Reichert and Brown, 1909). Scholars of the time did, however, recognize parallels between modern evolutionary biology and historical linguistics. Stevick (1963) presents "... an extended adaptation to linguistic interests of a biologist's statement about classification" (p. 162). Below is an extract in which Stevick paraphrases several pages of Dobzhansky's (1941) *Genetics and the Origin of Species* "by substituting linguistic for biological terms and examples":

"The conclusion that is forced on us is that the discontinuous variation encountered in natural speech, except that based on single feature differences, is maintained by means of preventing the intercommunication of representatives of the now discrete language groups. This conclusion is evidently applicable to discrete groups of any rank whatever, beginning with languages and up to and including branches and families. The development of isolating mechanisms is therefore a *conditio sine qua non* for emergence of discrete groups of forms in linguistic development... This conclusion is certainly not vitiated by the well-known fact that the isolation between groups may be complete or only partial. An occasional exchange of language materials, not attaining to the frequency of random interchange, results in the production of some intergrades, without, however, entirely swamping the differences between groups." (p. 163)

While biologists have embraced computational phylogenetic methods, the same cannot be said for historical linguists. Despite some initial enthusiasm, the numerical approaches introduced by Swadesh were heavily criticized and are now largely discredited (Bergsland and Vogt, 1962; Blust, 2000; Campbell, 2004). Criticisms of lexicostatistics and glottochronology tend to fall into four main categories that will be familiar to evolutionary biologists. First, the conversion of lexical character data to distance scores between languages results in the loss of information, reducing the power of the method to reconstruct evolutionary history accurately (Steel et al., 1988). Using distance data also makes it difficult to deal with polymorphisms (i.e., multiple terms in a language for a given meaning). Second, the clustering methods employed produced inaccurate trees, grouping together

languages that evolve slowly rather than languages that share a recent common ancestor (Blust, 2000). Third, language contact and borrowing of lexical items between languages make purely tree-based methods inappropriate (Bateman et al., 1990; Hjelmslev, 1958). And fourth, the assumption of constant rates of lexical replacement through time and across all meaning categories does not hold for linguistic data, making date estimates unreliable (Bergsland and Vogt, 1962). With the possibility of such errors, and no way of quantifying uncertainty, lexicostatistics fell out of favor and methods in biology and linguistics drifted apart.

Two notable exceptions to this trend are the mathematicians David Sankoff and Joseph Kruskal. Sankoff is well known in biology for his dynamic programming algorithm for counting character state changes on a phylogeny (Sankoff, 1975) and his work on rate variation and invariant sites (e.g., Sankoff, 1990). However, his early work was in lexicostatistical methods and rates of lexical evolution (Sankoff, 1969, 1970, 1973). Sankoff (1973) introduced the gamma distribution to linguistics as a means of modeling rate variation between words shortly after Uzzell and Corbin (1971) used a gamma distribution to model rate variation in molecular evolution. More recently, Sankoff and Kruskal have outlined how similar algorithms can be used to solve computational problems in different fields, including biology and linguistics (Sankoff and Kruskal, 1983). Kruskal was also involved with a number of lexicostatistical studies of Indo-European languages in the 1970s that elaborated on Swadesh's methods (Dyen et al., 1992; Kruskal et al., 1971, 1973).

During the last 50 years, computational phylogenetic methods and statistical inference have revolutionized evolutionary biology. A burgeoning of sequence data has produced enormous databases that can only be investigated using computational techniques. Conversely, the field of linguistics, haunted perhaps by the "ghost of glottochronology past," has remained curiously averse to computational phylogenetic methods.

THE NEW SYNTHESIS OF BIOLOGY AND LINGUISTICS

"The poets made all the words and therefore language is the archives of history."

The Poet (1844), Ralph Waldo Emerson (1906)

In a landmark paper published in 1988, Cavalli-Sforza et al. reported a figure directly comparing a human genetic and linguistic tree. Although extremely controversial (Bateman et al., 1990; O'Grady et al., 1989; Penny, Watson, and Steel, 1993), the Cavalli-Sforza et al. paper highlighted the similarities between processes of historical inference in biology and linguistics, as well as the potential importance of linguistic data for inferences about human population history. In the wake of this paper, there has been a proliferation of studies attempting to test hypotheses about human population history (see Bellwood and Renfrew, 2002; Cavalli-Sforza et al., 1994; Chikhi et al., 2002; Diamond and Bellwood,

2003; Hurlles et al., 2003; Richards et al., 2000; Semino et al., 2000), and something of a resurgence of interest in computational phylogenetic methods in historical linguistics. This “new synthesis” of biology and linguistics (McMahon and McMahon, 2003) has provided solutions to many of the problems that plagued lexicostatistics and glottochronology. For example, character-based tree-building techniques retain individual character state information, thus avoiding the problem of information loss associated with distance-based methods. Ringe et al. (1997, 2002) used compatibility methods to infer an Indo-European language tree from discrete grammatical, phonological, and lexical characters. Gray and Jordan (2000) conducted a parsimony analysis of over 5000 discrete lexical characters to find an optimal tree for 77 Austronesian languages. They then used this tree to test competing scenarios for the settlement of the Pacific. Holden (2002) applied similar methods to test migration scenarios in the Bantu language family and Rexová et al. (2003) constructed an Indo-European language tree, also using parsimony methods.

In biological phylogenetics over the last 15 years, there has been a gradual move away from parsimony analysis to likelihood models and Bayesian inference of phylogeny (see Huelsenbeck et al., 2001; Swofford et al., 1996). Explicitly modeling the process of evolution makes the assumptions of the method clear, makes it easy to implement more complex and realistic models of sequence evolution, and allows different models to be compared easily (Page and Holmes, 1998; Pagel, 2000). Moreover, statistical modeling techniques make it easier to quantify uncertainty in results and to test between competing hypotheses (Swofford et al., 1996). The process of character evolution can also be modeled in linguistics. Pagel (2000) used an explicit likelihood model of lexical replacement to make inferences about different rates of word evolution. More recently, Pagel and Meade (in press) compared rates of change between meanings in Indo-European and Bantu languages. They not only found a relationship between rates of meaning evolution between language families, but also, quite remarkably, that rates of word replacement in these languages are correlated with rates of word use in English today (Pagel and Meade, in press). Gray and Atkinson (2003) combined a likelihood model of lexical evolution with Bayesian inference of phylogeny (Huelsenbeck and Ronquist, 2001) to construct a distribution of the most probable trees for the Indo-European language family. We then used a penalized likelihood rate-smoothing algorithm (Sanderson, 2002a, 2002b) to infer the age of the Indo-European language family. Sanderson developed the rate-smoothing approach to allow biologists to infer divergence times without having to assume a constant molecular clock. By applying this algorithm to linguistic data, we were able to overcome one of the fundamental problems of glottochronology (the glottoclock is not constant), and thus test between two competing hypotheses for the age of the Indo-European language family. By analyzing linguistic and genetic data in a common analytical framework, much more precise infer-

ences about human history should be possible (see Gray and Jordan, 2000, and Hurlles et al., 2003, for attempts to synthesise linguistic, genetic and archaeological inferences about Pacific settlement). Interestingly, quantitative phylogenetic methods have also been used in the study of manuscript evolution to produce a phylogeny of the *Canterbury Tales* (Barbrook et al., 1998).

FUTURE CHALLENGES

Today, researchers using computational methods in evolutionary biology and historical linguistics aim to answer similar questions and hence face similar challenges. One emerging challenge in computational historical linguistics lies in developing algorithms to determine the probability that lexical characters are cognate (Heeringa et al., 2000; Kondrak, 2001; Covington, 1996). The sounds comprising each word must be compared across large sets of data to determine cognacy. Accurate comparisons between words must allow for insertions, deletions, and metathesis (reversals) and incorporate complex models of phonological change. These comparisons are fundamentally similar to reconstructions of character change on a phylogeny. Hence, when historical linguists make cognacy judgements using the comparative method, they quite rightly consider prior knowledge of the relationships between these languages. Unfortunately, because there is no explicit optimality criterion used to make the cognacy judgements, it is difficult to evaluate the evidence supporting a relationship objectively. Identifying cognates between languages has obvious parallels with the problem of sequence alignment in biology. Biologists must also deal with insertions, deletions, and reversals, and are beginning to consider phylogeny (Felsenstein, 2004). As a result, biologists have proposed methods that simultaneously perform alignment and reconstruction of phylogeny (e.g., Clustal W; Thompson et al., 1994). These methods require further development and should be applicable to historical linguistics.

Model fitting and comparison is another challenge jointly faced by phylogenetic methods in biology and linguistics. Burnham and Anderson (1998) describe model choice as a balance between under- and overfitting parameters. A model that is too simple may produce biased results if it fails to capture important parts of the evolutionary process. Conversely, adding extra parameters may improve the apparent fit of a model to data but at the cost of increasing sampling error and computational complexity as there are more parameters to estimate. Traditional lexicostatistics and glottochronological methods implicitly assumed a very simple model of constant rates of change between different meanings and over time. Divergence dates between languages were estimated using the formula,

$$t = \log c / 2 \log r$$

where t is time, c is the percentage of shared cognates, and r is the retention rate per thousand years. The retention rate was assumed to be roughly constant at around 81% per thousand years for the Swadesh 200 word list (a

list of basic vocabulary terms). However, as mentioned above, this approach produced some obviously incorrect results. Gray and Atkinson (2003) attempted to overcome these problems by inferring an Indo-European language phylogeny from discrete rather than distance data, by using a gamma distribution to model rate-variation between cognate sets, and by using rate-smoothing to allow for rate variation through time. The likelihood model used by Gray and Atkinson assumed that gains and losses of cognates were equally likely. This is not particularly realistic. Assuming that the effects of borrowing can be excluded, it is much more probable that cognates would evolve only once but could be lost multiple times. Atkinson et al. (2005) have implemented a "Dollo" likelihood model of cognate evolution as a move in the direction of greater realism. Interestingly, the tree topologies and divergence times inferred for Indo-European languages using this model are congruent with those reported by Gray and Atkinson (2003).

The models of lexical evolution discussed so far all make the standard "rates across sites" assumption. "Rates across sites" models (see Yang, 1993, 1994) essentially modify the independent and identically distributed (IID) rates assumption to allow for rate variation between sites. This is usually achieved by treating the rate of evolution at each site as a variable drawn from some distribution (usually a gamma distribution) and/or by allowing for a proportion of invariant sites. In other words, these models assume that, while historical, social, and cultural contingencies can undoubtedly influence the process of linguistic change, fundamental factors such as similarities in the way humans acquire language, and the need to communicate in an expressive and intelligible way, mean that there are sufficient commonalities in the way different words will evolve to justify the "rates across sites" assumption as a useful starting point. In the words of

Ringe et al. (2002: 61),

"Languages replicate themselves (and thus 'survive' from generation to generation) through a process of native-language acquisition by children. Importantly for historical linguistics, that process is tightly constrained."

Warnow et al. (in press) reject the "rates across sites" assumption. Instead, they advocate a "no common mechanism" model (see Steel and Penny, 2000) of language evolution in which rates of change are unrelated between meanings and across branches. Such a model does not allow branch lengths or divergence times to be inferred. More work is needed to determine if the costs associated with this more complex set of models are outweighed by sufficient benefits.

In studies of biological evolution, investigations of factors such as generation time, body size, temperature, metabolic rate, and population size that may affect the rate of molecular evolution are a major area of current inquiry (Bromham and Penny, 2003). A similar list of factors has been proposed to affect the rate of linguistic evolution. Nettle (1999), for example, found that in a plausible computer simulation, smaller speech communities will have higher rates of change and a greater probability of borrowing. In an explicit analogy with biological evolution, Green (1966, 1987) suggested that the successive "founder events" that occurred in the settlement of the Pacific led to accelerated rates of linguistic change (see Fig. 4). Other researchers have suggested that factors like contact between languages and the population mobility might also affect language evolution rates (e.g., Blust, 2000; Pawley, 2002). One benefit that could follow from the use of explicit likelihood models in studies of language evolution is the possibility of rigorous comparative tests of hypotheses about factors that affect the rate of linguistic evolution.

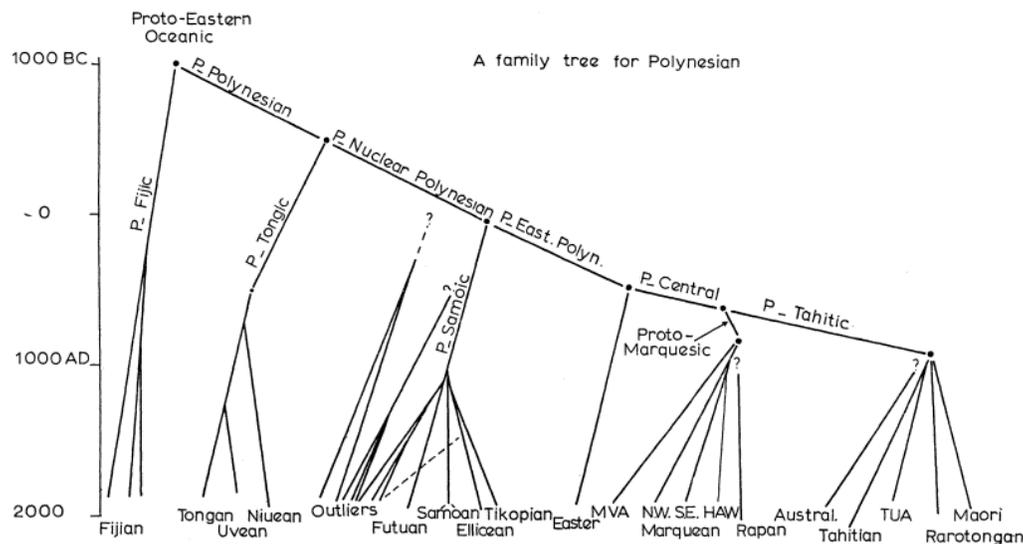


FIGURE 4. A phylogenetic tree for Polynesian languages from Green (1966). The tree is unusual for a language tree in that it depicts divergence times and suggests an increase in the rate of linguistic change in eastern Polynesia (note the flattening of the branch across the base of eastern Polynesia). (Courtesy of the *Journal of the Polynesian Society*.)

Another important challenge in both computational biology and historical linguistics lies in developing methods to investigate reticulate evolution. Tree models of evolution dominate historical linguistics, just like evolutionary biology. A persistent criticism of this approach is that human history is far from tree-like (Moore 1994; Terrell 1988; Terrell et al., 2001). Not only might patterns of genetic, linguistic, and cultural diversity reflect different histories (Bateman et al., 1990), each of these histories might be strikingly reticulate. Casual consideration of the history of the English language would lead one to believe that language evolution is anything but tree-like. English is a veritable fruit salad of a language with chunks of vocabulary from the Celts, Romans, Angles, Saxons, Jutes, Vikings, Normans, and slices of Latin, French, Greek, and Italian tossed with some more recent garnishes from Arabic, Persian, Turkish, and Hindi. There is even the odd Polynesian borrowing like "tattoo." Ninety-nine percent of words in the *Oxford English Dictionary* are, in fact, borrowings from other languages (McWhorter, 2001), and over 50% of the total English lexicon comes from Romance languages post the Norman conquest. This figure, however, falls to around 6% for basic vocabulary such as the Swadesh 200 word list (Embleton, 1986). In the case of the Indo-European language family, a number of extinct languages are attested in ancient texts and, for many subgroups, linguists have been able to reconstruct a detailed account of the sound changes that occurred since Proto-Indo-European. This makes it possible to identify many, but not all, of the borrowed terms, which do not fit into the regular pattern of sound change. However, most language families are not so well understood. In such cases, the existence of dialect chains, borrowing of cultural terms, and contact between languages can therefore pose major problems for attempts to infer language trees. Similarly, biologists studying plants and prokaryotic evolution must deal with hybridization and horizontal gene transfer. Hence, an important challenge for both biologists and historical linguistics is the development of methods to investigate reticulate evo-

lution. Minett and Wang (2003) developed a method to detect borrowing between languages through identifying incompatible characters on a phylogeny. They propose that in some circumstances even the direction of borrowing may be determined. In biology, a number of new methods have recently been proposed for visualising conflicting signals in the data (e.g., split decomposition and NeighborNet analysis, see Huson, 1998; and Bryant and Moulton, 2002). Bryant et al. (2005) have outlined how these methods can be used to investigate conflicting signal caused by lexical borrowing. One test of these methods is their ability to recover evidence of known reticulation. Sranan is a Creole language developed by African slaves in Surinam. The English established Surinam on the northern coast of South America in 1651 as a slave colony. However, Dutch has been the official language since 1667 and hence Sranan's lexicon contains words derived from both English and Dutch (McWhorter, 2001). Figure 5 shows the split decomposition and NeighborNet graphs reported by Bryant et al. (in press) for some Germanic languages, including Sranan. Both analyses recovered the conflicting signal generated by Sranan's hybrid history. A further challenge in both biology and linguistics is to explicitly model reticulate evolution. Atkinson et al. (2005) have taken a step along this road, using an evolutionary model that incorporates word borrowing between random languages to synthesize data and test the robustness of their results to borrowing.

Finally, in historical linguistics, as in biology, there is the question of whether it is possible to infer distant genetic relationships reliably. Campbell (2004) identifies a number of controversial attempts to establish linguistic superfamilies, including Nostratic (comprising Indo-European, Uralic, Altaic, Dravidian, Kartvelian, and Afroasiatic), Amerind (comprising all of the languages of the Americas except Eskimo-Aleut and Na-Dene), and even Proto-World, the global mother tongue (Shevoroshkin, 1990). The problem with making inferences about these very deep relationships is

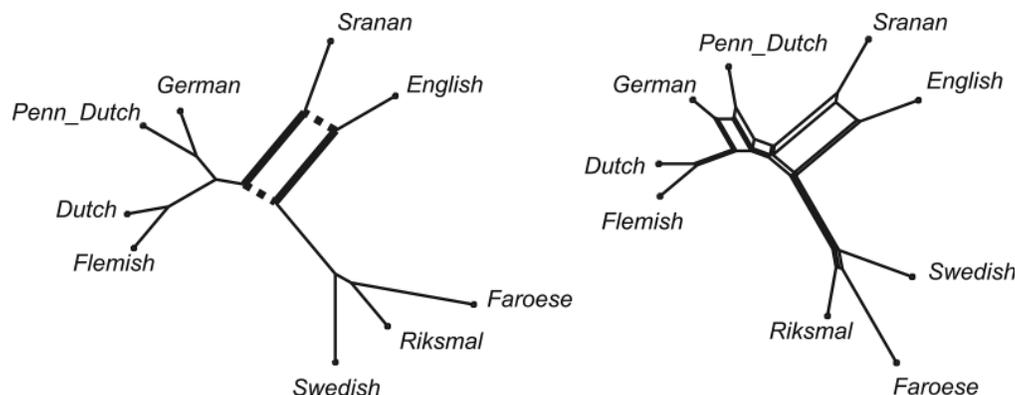


FIGURE 5. The networks produced by split decomposition (left) and NeighborNet analyses (right) for a selection of Germanic languages (from Bryant et al., in press). Both networks display the conflicting signal introduced by the Creole Sranan. In the split decomposition graph (left), the split grouping Sranan and English is shown in bold, whereas the conflicting split grouping Sranan with German, Penn. Dutch, Dutch, and Flemish is shown as a dotted line. (From Mace, R., C. Holden, and S. Shennan (eds.). *The evolution of cultural diversity: A phylogenetic approach*, Chapter 5 [© UCL Press, 2005].)

that languages change much more quickly than gene sequences. Most linguists believe that after about 8,000 to 10,000 years, the comparative method breaks down as it becomes impossible to differentiate between homology and chance resemblances or borrowings (see Nichols, 1992). For instance, although Maori *mata*, "eye," and modern Greek *mati*, "eye," appear cognate, the resemblance is actually due to chance. Linguists are thus highly sceptical of arguments for ancient language relationships, especially when cognacy judgements are made with less than the normal standard of rigor. For example, Joseph Greenberg (1987) and Merritt Ruhlen (1994) proposed a 12,000-year-old "Amerind" family of Native American languages based on a technique called mass-comparison. They examined large numbers of words between languages in search of words with a similar form and related meanings. For example, Ruhlen (1994, p. 168) offered as evidence for Amerind, words ostensibly related to a hypothetical Proto-Amerind term **t'ana*, "child, sibling." As Campbell (2003) points out, the semantic variation Ruhlen allowed (meanings including small, woman, cousin, son-in-law, old man, friend, and some 15 other terms) coupled with relatively loose phonetic matches (Ruhlen treats *tsuh-ki* and *u-tse-kwa* as related to **t'ana*) make chance resemblance highly likely. Campbell (2003) goes on to cite examples of words from English (son), German (tante, "aunt") and Maori (tiena, "younger sibling") that would be misidentified as related by Ruhlen's criteria. On a different time scale, in biology the much less controversial, but still highly contentious, tree of life continues to provoke debate (Gribaldo and Commarano, 1998). One challenge is to push the depth at which it is feasible to reconstruct a phylogeny back further and to develop criteria for accepting or rejecting genetic relationships. Pagel (2000) has shown that some words evolve slowly enough to make it possible to, at least in principle, resolve 20,000-year-old language relationships. The practical challenge of discriminating these deep homologies from more recent borrowings and chance similarities still remains, however. One possible solution is to analyze the kind of abstract grammatical characters that are claimed to have slower rates of evolutionary change (Nichols, 1992). Deep relationships may then be able to be resolved by combining different forms of linguistic data in a single analysis, each with a different model. Biologists have developed methods to combine nucleotide data with protein, restriction site, gene order and morphological data (e.g., MrBayes; Huelsenbeck and Ronquist, 2001). Linguists may be able to achieve the same benefit by treating lexical, grammatical and phonological data simultaneously. Finally, by analyzing synthetic data on a phylogeny under a given model, we can measure the ability of a given method to reconstruct deep relationships from different data types. Alternatively, Mossel and Steel (2005) have proposed analytical methods for assessing the extent to which deep phylogenetic relationships can be inferred from a number of biological data types. Again, this has obvious applications to linguistics.

These common challenges are a reflection not only of the curious parallels of process that exist between bio-

logical and linguistic evolution, but they also reflect over two millennia of coevolution between research in biology and historical linguistics. In the light of such parallels, it seems likely that biology and linguistics will remain curiously, and let's hope productively, connected.

ACKNOWLEDGMENTS

We thank Bob Blust, Lyle Campbell, Penny Gray, Roger Green, Simon Greenhill, Mark Liberman, Geoff Nicholls, Mark Pagel, and David Penny for useful advice and/or comments on the manuscript.

REFERENCES

- Atkinson, Q. D., G. Nicholls, D. Welch, and R. D. Gray. 2005. From words to dates: Water into wine, mathemagic or phylogenetic inference? *Transactions of the Philological Society* 103(2):193–219.
- Barbrook, A. C., C. J. Howe, N. Blake, and P. Robinson. 1998. The phylogeny of the Canterbury Tales. *Nature* 394:839.
- Bateman, R., I. Goddard, R. O'Grady, V. Funk, R. Mooi, W. Kress, and P. Cannell. 1990. Speaking of forked tongues: The feasibility of reconciling human phylogeny and the history of language. *Curr. Anthropol.* 31:1–24.
- Bellwood, P., and C. Renfrew (eds.). 2002. Examining the farming/language hypothesis. MacDonald Institute for Archaeological Research, Cambridge, UK.
- Bergsland, K., and H. Vogt. 1962. On the validity of glottochronology. *Curr. Anthropol.* 3:115–153.
- Blust, R. 2000. Why lexicostatistics doesn't work: The 'universal constant' hypothesis and the Austronesian languages. Pages 311–332 in *Time depth in historical linguistics* (C. Renfrew, A. McMahon, and L. Trask, eds.). The McDonald Institute for Archaeological Research, Cambridge.
- Blust, R. 2003. Vowelless words in Selau. Pages 143–152 in *Issues in Austronesian historical phonology* (J. Lynch, ed.). Pacific Linguistics, The Australian National University, Canberra.
- Bromham, L., and D. Penny. 2003. The modern molecular clock. *Nat. Rev. Genet.* 4:216–224.
- Brugmann, K., and H. Osthoff. 1878. Preface to *Morphologische Untersuchungen auf dem Gebiet der indogermanischen Sprachen*, volume 1 (W. Lehmann, trans.). In *A reader in Nineteenth-Century historical Indo-European linguistics* (W. Lehmann, ed.). Indiana University Press, Bloomington, Indiana, 1963.
- Bryant, D., F. Filimon, and R. D. Gray. 2005. Untangling our past: Pacific settlement, phylogenetic trees and Austronesian languages. Pages 69–85 in *The evolution of cultural diversity: Phylogenetic approaches* (R. Mace, C. Holden, and S. Shennan, eds.). UCL Press, London.
- Bryant, D., and V. Moulton. 2002. NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. Workshop in Algorithms for Bioinformatics, Proceedings 2002:375–391.
- Burnham, K. P., and D. R. Anderson. 1998. *Model selection and inference: A practical information-theoretic approach*. Springer, New York.
- Campbell, L. 2000. The history of linguistics. Pages 81–104 in *The handbook of linguistics*. (M. Aronoff and J. Rees-Miller, eds.). Blackwell Publishing, Oxford, UK.
- Campbell, L. 2004. How to show languages are related: Methods for distant genetic relationship. Pages 262–282 in *The handbook of historical linguistics* (B. D. Joseph and R. D. Janda, eds.). Blackwell Publishing, Malden, Massachusetts.
- Campbell, L. 2004. *Historical linguistics: An introduction*, 2nd edition. Edinburgh University Press, Edinburgh.
- Cannon, W. 1961. The impact of uniformitarianism: Two letters from John Herschel to Charles Lyell. *Proc. Am. Phil. Soc.* 105:301–14, 1836–1837.
- Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The history and geography of human genes*. Princeton University Press, Princeton, New Jersey.
- Cavalli-Sforza, L. L., A. Piazza, P. Menozzi, and J. Mountain. 1988. Reconstruction of human evolution: Bringing together genetic, archaeological, and linguistic data. *Proc. Nat. Acad. Sci. USA* 85:6002–6006.

- Chikhi, L., R. A. Nichols, G. Barbujani, and M. A. Beaumont. 2002. Y genetic data support the neolithic demic diffusion model. *Proc. Nat. Acad. Sci.* 99:11008–11013.
- Collin, H. S., and C. J. Schlyter. 1827. *Corpus Iuris Sueo-Goto-rum Antiqui*. Z. Haeggstrom, Stockholm, 1.
- Covington, M. 1996. An algorithm to align words for historical comparison. *Comput. Ling.* 22:481–496.
- Croft, W. 2000. *Explaining language change: An evolutionary approach*. Pearson Education, Harlow.
- Crowley, T. 1992. *An introduction to historical linguistics*, 3rd edition. Oxford University Press, Oxford, UK.
- Darwin, C. 1859. *The origin of species by means of natural selection*. Oxford University Press, Oxford, UK.
- Darwin, C. 1871. *The descent of man*. Murray, London.
- Diamond, J., and P. Bellwood. 2003. Farmers and their languages: The first expansions. *Science* 300:597.
- Dobzhansky, T. 1937. *Genetics and the origin of species*. Columbia University Press, New York.
- Dobzhansky, T. 1941. *Genetics and the origin of species*, 2nd edition. Columbia University Press, New York.
- Dyen, I., J. B. Kruskal, and P. Black. 1992. An Indo-European classification: A lexicostatistical experiment. *Trans. Am. Phil. Soc.* 82:1–132.
- Embleton, S. 1986. *Statistics in historical linguistics*. Brockmeyer, Bochum.
- Emerson, R. 1906. *Essays*, 1st and 2nd series. Dent, London.
- Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- Fisher, R. A. 1930. *The genetical theory of natural selection*. Oxford University Press, Oxford.
- Gleason, H. A. 1959. Counting and calculating for historical reconstruction. *Anthropol. Ling.* 1:22–32.
- Gould, S. 2002. *The structure of evolutionary theory*. Belknap Press, Cambridge.
- Gould, S. J., and R. Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc. R. Soc. Lond. B. Biol. Sci.* 205:581–598.
- Gray, R., and Q. Atkinson. 2003. Language tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 405:435–439.
- Gray, R., and F. Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 405:1052–1055.
- Green, R. C. 1966. Linguistic subgrouping within Polynesia: The implications for prehistoric settlement. *J. Polynesian Soc.* 75:6–38.
- Greenberg, J. H. 1987. *Language in the Americas*. Stanford University Press, Stanford, California.
- Gribaldo, S., and P. Commarano. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. *J. Mol. Evol.* 47:508–516.
- Grimm, J. 1822. *Deutsche grammatik*. Part I: Zweite ausgabe. Dieterich, Göttingen.
- Heeringa, W., J. Nerbonne, and P. Kleiweg. 2002. Validating dialect comparison methods. Pages 445–452 in *Classification, automation, and new media* (W. Gaul, and G. Ritter, eds.). Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e. V., University of Passau, March 15–17, 2000, Springer, Berlin, Heidelberg, and New York.
- Hennig, W. 1950. *Grundzüge einer theorie der phylogenetischen systematik*. Deutscher Zentralverlag, Berlin.
- Hjelmlev, L. 1958. *Essai d'une critique de la methode dite glottochronologique*. Proceedings of the Thirty-Second International Congress of Americanists, Copenhagen, 1956. Munksgaard, Copenhagen.
- Hoenigswald, H. 1960. *Language change and linguistic reconstruction*. The University of Chicago Press, Chicago.
- Hoenigswald, H. M. 1990. Language families and subgroupings, tree model and wave theory, and reconstruction of protolanguages. Pages 441–454 in *Research guide on language change* (E. C. Polome, ed.). Trends in Linguistics, Studies and Monographs, 48. Mouton de Gruyter, Berlin and New York.
- Holden, C. J. 2002. Bantu language trees reflect the spread of farming across Sub-Saharan Africa: A maximum-parsimony analysis. *Proc. R. Soc. Lond. B* 269:793–799.
- Householder, F. 1981. *The syntax of Apollonius dyscolus*. John Benjamins B. V., Amsterdam.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* 17:754–755.
- Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310.
- Hull, D. 1988. *Science as process: An evolutionary account of the social and conceptual development of science*. University of Chicago Press, Chicago.
- Hurles, M. E., L. Matisoo-Smith, R. D. Gray, and D. Penny. 2003. Untangling oceanic settlement: The edge of the knowable. *Trends Ecol. Evol.* 18:531–540.
- Huson, D. H. 1998. SplitsTree: Analyzing and visualizing evolutionary data. *Bioinformatics* 14:68–73.
- Jefferson, T. 1984. Reprint of *Notes on the State of Virginia* (M. D. Peterson, ed.). Library of America, Literary classics of the United States, New York. 1781–1782.
- Jones, Sir W. 1786. Third anniversary discourse: 'On the Hindus', reprinted in *The collected works of Sir William Jones*. V III, 1807. Stockdale, London.
- Koerner, K. (ed.) 1983. Preface. In *August Schleicher: Die Sprachen Europaas in systematischer Übersicht* (K. Koerner, ed.). John Benjamins Publishing Co., Amsterdam.
- Kondrak, G. 2001. Identifying cognates by phonetic and semantic similarity. Pages 103–110 in *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2001)*. Pittsburgh, June, 2001.
- Kruskal, J., I. Dyen, and P. Black. 1971. The vocabulary method of reconstructing language trees: Innovations and large-scale applications. Pages 361–380 in *Mathematics in the archeological and historical sciences* (F. R. Hodson, D. G. Kendall, and P. Tautu, eds.). Edinburgh University Press, Edinburgh.
- Kruskal, J. B., I. Dyen, and P. Black. 1973. Some results from the vocabulary method of reconstructing language trees. Pages 30–55 in *Lexicostatistics in genetic linguistics* (I. Dyen, ed.). Mouton, The Hague.
- Lamarck, J. B. 1809. *Philosophie zoologique* trans. by Hugh Elliot as *Zoological philosophy: An exposition with regard to the natural history of animals, with introductory essays* by David L. Hull and Richard W. Burkhardt Jr. Chicago, 1984.
- Leibniz, G. W. 1710. *Miscellanea berolinensia* in *Berlin memoirs*. Berlin Academy, Berlin.
- Leibniz, G. W. 1712. *Monadology*. In G. W. Leibniz: *Philosophical papers*, 2nd edition, 1969 (L. E. Loemker, ed.). Reidel, Dordrecht.
- Maher, J. 1983. Introduction. In *Linguistics and evolutionary theory* (K. Koerner, ed.). John Benjamins Publishing Co., Amsterdam.
- Mayr, E. 1982. *The growth of biological thought*. Harvard University Press, Cambridge, Massachusetts.
- McMahon, A., and R. McMahon. 2003. Finding families: Qualitative methods in language classification. *Trans. Philol. Soc.* 101:7–55.
- McWhorter, J. H. 2001. *The power of Babel*. Arrow Books, London.
- Menozzi, P., A. Piazza, and L. L. Cavalli-Sforza. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201:786–792.
- Michener, C., and R. Sokal. 1957. A quantitative approach to a problem in classification. *Evolution* 11:130–162.
- Minett, J., and W. Wang. 2003. On detecting borrowing: Distance-based and character-based approaches. *Diachronica* 20:289–330.
- Moore, J. H. 1994. Putting anthropology back together again: The ethnogenetic critique of cladistic theory. *Am. Anthropologist* 96:925–948.
- Mossel, E., and M. Steel. 2005. How much can evolved characters tell us about the tree that generated them? Pages 384–412 in *Mathematics of evolution and biology* (O. Gascuel ed.). Oxford University Press, Oxford.
- Nettle, D. 1999. Is the rate of linguistic change constant? *Lingua* 108:119–136.
- Nichols, J. 1992. *Linguistic diversity in space and time*. University of Chicago Press, Chicago.
- O'Grady, R., I. Goddard, R. M. Bateman, W. A. DiMicheal, V. A. Funk, W. J. Kress, R. Mooi, and P. F. Cannell. 1989. *Genes and tongues*. *Science* 243:1651.
- O'Hara, R. 1996. *Trees of history in systematics and philology*. *Memorie della Società Italiana di Scienze Naturali e del Museo Civico di Storia Naturale di Milano* 27:81–88.
- Page, R. D. M., and E. C. Holmes. 1998. *Molecular evolution: Phylogenetic approach*. University Press, Cambridge.

- Pagel, M. 2000. Maximum-likelihood models for glottochronology and for reconstructing linguistic phylogenies. Pages 189–207 in *Time depth in historical linguistics* (C. Renfrew, A. McMahon, and L. Trask, eds.). McDonald Institute for Archaeological Research, Cambridge.
- Pagel, M., and A. Meade. In press. Estimating rates of meaning evolution on phylogenetic trees of languages. In *Phylogenetic methods and the prehistory of languages*. (J. Clackson, P. Forster and C. Renfrew, eds.). McDonald Institute for Archaeological Research, Cambridge.
- Pawley, A. 2002. The Austronesian dispersal: Languages, technologies and people. Pages 251–274 in *Examining the farming/language hypothesis* (Bellwood, P. and C. Renfrew, eds.). MacDonal Institute for Archaeological Research, Cambridge.
- Pawley, A., and F. Syder. 1983. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *J. Pragmatics* 7:551–579.
- Pedersen, H. 1931. *The discovery of language—Linguistic science in the nineteenth century*. Indiana University Press, Bloomington.
- Penny, D., E. Watson, and M. Steel. 1993. Trees from languages and genes are very similar. *Syst. Biol.* 42:382–384.
- Percival, K. 1987. Biological analogy in the study of language before the advent of comparative grammar. Pages 3–38 in *Biological metaphor and cladistic classification* (H. Hoenigswald and L. Wiener, eds.). University of Pennsylvania Press, Philadelphia.
- Platnick, N., and D. Cameron. 1977. Cladistic methods in textual, linguistic and phylogenetic analysis. *Syst. Zool.* 26:380–385.
- Reichert, E., and A. Brown. 1909. The differentiation and specificity of corresponding proteins and other vital substances in relation to biological classification and organic evolution. Carnegie Institute, Washington, DC. Pub. No. 116.
- Rexová, K., D. Frynta, and J. Zrzavy. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19:120–127.
- Richards, M., V. Macaulay, E. Hickey, E. Vega, B. Sykes, V. Guida, C. Rengo, D. Sellitto, F. Cruciani, T. Kivisild, R. Villems, M. Thomas, S. Rychkov, O. Rychkov, Y. Rychkov, M. Golge, D. Dimitrov, E. Hill, D. Bradley, V. Romano, F. Cali, G. Vona, A. Demaine, S. Papiha, C. Triantaphyllidis, G. Stefanescu, J. Hatina, M. Belleli, A. Di Rienzo, A. Novelletto, A. Oppenheim, S. Norby, N. Al-Zaheri, S. Santachiara-Benerecetti, R. Scozari, A. Torroni, and H. J. Bandelt. 2000. Tracing European founder lineages in the near eastern mtDNA pool. *Am. J. Hum. Genet.* 67:1251–1276.
- Ringe, D., T. Warnow, and A. Taylor. 2002. Indo-European and computational cladistics. *Trans. Phil. Soc.* 100:59–129.
- Ringe, D., T. Warnow, A. Taylor, A. Michailov, and L. Levison. 1997. Computational cladistics and the position of Tocharian. Pages 391–414 of *Monograph 26 in The bronze age and early iron age peoples of eastern central Asia* (V. Mair, ed.). A special volume of the *Journal of Indoeuropean Studies*.
- Robins, R. 1997. *A short history of linguistics*. Longmans, London.
- Ruhlen, M. 1994. *The origin of language: Tracing the origin of the mother tongue*. John Wiley and Sons, New York.
- Sanderson, M. 2002a. Estimating absolute rates of evolution and divergence times: A penalized likelihood approach. *Mol. Biol. Evol.* 19:101–109.
- Sanderson, M. 2002b. R8s, analysis of rates of evolution, version 1.50. <http://ginger.ucdavis.edu/r8s/>
- Sankoff, D. 1969. *Historical linguistics as a stochastic process*. PhD thesis. McGill University, Montreal.
- Sankoff, D. 1970. On the rate of replacement of word-meaning relationships. *Language* 46:564–569.
- Sankoff, D. 1973. Mathematical developments in lexicostatistical theory. Pages 93–112 in *current trends in linguistics 11: Diachronic, areal and typological linguistics* (T. A. Sebeok, ed.). Mouton, The Hague.
- Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM J. Appl. Math.* 28:35–42.
- Sankoff, D. 1990. Designer invariants for large phylogenies. *Mol. Biol. Evol.* 7:255–269.
- Sankoff, D., and J. Kruskal. 1983. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Addison-Wesley Publishing Co., Reading, Massachusetts.
- Schlegel, F. 1808. *Über die sprache und weisheit der Indier: Ein beitrag zur begründung der alterthumskunde*. In *Amsterdam clas-*
- sics in linguistics* (S. Timpanaro, ed.). John Benjamins Publishing Co., Amsterdam.
- Schleicher, A. 1863. *Die Darwinshce theorie und die sprachwissenschaft*. Hermann Bohlau, Weimar.
- Schmidt, J. 1872. *Die verwantschaftsverhältnisse der Indogermanische sprachen*. Herman Bohlau, Weimar.
- Semino, O., G. Passarino, P. Oefner, A. Lin, S. Arbizova, L. Beckman, G. De Benedictis, P. Francalacci, S. Limborska, A. Kouvatsi, M. Marcikiae, D. Primorac, S. Santachiara-Benerecetti, L. L. Cavalli-Sforza, and P. Underhill. 2000. The genetic legacy of palaeolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. *Science* 290:1155–1159.
- Shevoroshkin, V. 1990. The mother tongue: How linguists have reconstructed the ancestor of all living languages. *The Sciences* May/June:20–27.
- Smith, S., and F. Burkhardt. 1985. *The correspondence of Charles Darwin*. Cambridge University Press, Cambridge.
- Sneath, P. 1957. The application of computers to taxonomy. *Journal of General Microbiology*. 17:201–226.
- Steel, M., M. Hendy, and D. Penny. 1988. Loss of information in genetic distances. *Nature* 333:494–495.
- Steel, M., and D. Penny. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol. Biol. Evol.* 17:839–850.
- Stevick, R. 1963. The biological model and historical linguistics. *Language* 39:159–169.
- Swadesh, M. 1952. Lexico-statistic dating of prehistoric ethnic contacts. *Am. Phil. Soc. Proc.* 96:453–463.
- Swadesh, M. 1955. Towards greater accuracy in lexicostatistic dating. *Int. J. Am. Ling.* 21:121–137.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 in *Molecular systematics*, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Marple, ed.). Sinauer, Sunderland, Massachusetts.
- Terrell, J. 1988. History as a family tree, history as an entangled bank: Constructing images and interpretations of prehistory in the South Pacific. *Antiquity* 62:642–657.
- Terrell, J., K. M. Kelly, and P. Rainbird. 2001. Foregone conclusions? *Curr. Anthropol.* 42:97–124.
- Thomason, S., and T. Kaufman. 1988. *Language contact creolization, and genetic linguistics*. University of California Press, Berkeley, California.
- Thompson, D. 1913. *On Aristotle as a biologist*. Clarendon Press, Oxford. Essay reprinted in D'Arcy Wentworth Thompson, *Science and the classics*. Oxford University Press, H. Milford, Oxford 1940.
- Thompson, J., D. Higgins, and T. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
- Uzzell, T., and K. Corbin. 1971. Fitting discrete probability distribution to evolutionary events. *Science* 172:1089–1096.
- Warnow, T., S. N. Evans, D. Ringe, and L. Nakhleh. In press. Stochastic models of language evolution and an application to the Indo-European family of languages. In *Phylogenetic methods and the prehistory of languages* (J. Clackson, P. Forster and C. Renfrew, eds.). McDonald Institute for Archaeological Research, Cambridge.
- Watson, J. D., and F. H. C. Crick. 1953. Molecular structure of nucleic acids. *Nature* 171:737–738.
- Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10:1396–1401.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39:306–314.
- Zuckerandl, E., and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189–225 in *Horizons in biochemistry*, ed. M. Kasha and B Pullman. Academic Press, New York.
- Zuckerandl, E., and L. Pauling. 1965. Molecules as documents of evolutionary history. *J. Theoret. Biol.* 8:357–366.

First submitted 20 October 2004; reviews returned 24 November 2004;

final acceptance 24 November 2004

Associate Editor: Chris Simon