Invited Historical Essay

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.]

CURIOS 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 (*watar [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 *tola, “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 *bök, “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 breo 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

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

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...
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 (Pederson, 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...
...of the natural world that was at odds with a theistic account of biological and linguistic diversity. In biology, the Greek conception of immutable archetypes, which fitted so nicely with scripture, was beginning to be challenged. Leibniz (1712) argued that nature is constantly changing, and what is more, this change occurs gradually:

“Everything goes by degrees in nature, and nothing leaps, and this rule controlling changes is part of my law of continuity.” (p. 376)

Leibniz was influential in shifting interest during this time towards processes of change, although it was not until Jean-Baptiste Lamarck (1744–1829) that ideas of change began to be debated seriously and even then creationist accounts were still favored (Mayr, 1982). Just as historical linguists had found it difficult to integrate their historical hypotheses with accounts offered by scripture, evolutionary accounts of species diversity were hampered by the biblical chronology, which was thought to imply an age for the earth of no more than 6,000 years.

Leibniz (1710) also applied his ideas of gradualism and uniformitarianism to linguistics. He argued that languages, as natural phenomena, must change in a gradual and continuous manner. Leibniz rejected doctrinaire arguments for a Hebraic root to all languages as well as Scaliger’s proposition of a large number of unrelated language groups (Pedersen, 1931). Instead, he tried to construct a genealogy of the languages of Europe, Asia, and Egypt, arguing that all these languages had descended from some common ancestor (Pedersen, 1931). To this end, he advocated the creation of grammars and dictionaries for all of the languages of the world (Robins, 1997). Although Leibniz’s genealogical conclusions were full of errors, his ideas of gradualism and uniformitarianism remain fundamental (if somewhat controversial) in linguistics, as in biology.

Notions of gradual, continuous change were also expressed implicitly on the other side of the Atlantic by none other than Thomas Jefferson. In Notes on the State of Virginia (written 1781–1782), he suggests the possibility of using linguistic data not only to infer historical relationships, but also to infer divergence times:

“A separation into dialects may be the work of a few ages only, but for two dialects to recede from one another till they have lost all vestiges of their common origin, must require an immense course of time; perhaps not less than many people give to the age of the earth. A greater number of those radical changes of language having taken place among the red men of America, proves them of greater antiquity than those of Asia.” (p. 227)

Most modern histories of historical linguistics begin with Jefferson’s contemporary, Sir William Jones (1746–1794). In 1786, the British Orientalist and judge identified similarities between Sanskrit, Greek, Celtic, Gothic, and Latin that led him to conclude that these languages had “sprung from some common source, which perhaps no longer exists” (pp. 34–35). Jones is generally given credit for the rapid subsequent acceptance of the Indo-European language family and the proliferation of broad comparative studies of Indo-European grammar, phonology, and lexicon (Robins, 1997). In fact, Jones’s methods were not at all novel—he was not the first to suggest a link between Sanskrit and some European languages and his conclusions were full of errors (Campbell, 2000). Most notably, he mistakenly identified Pahlavi (Persian, an Indo-European language) as Semitic and argued for a genealogical connection between the Hindus, Egyptians, Phoenicians, Chinese, Japanese, and Peruvians (Campbell, 2000)! Nonetheless, Jones’s announcement did mark the beginning of a distinctly historical orientation in linguistics that would last throughout the following century. Even today, the nature, location, and timing of the “common source” or common ancestor of Indo-European is still hotly contested.

By the end of the 18th century, both biologists and linguists recognized varying degrees of relatedness and used the concept of homology. Linguists understood that diversity could be explained via descent with modification, they had linked homology with common ancestry, and they were beginning to concern themselves with genealogical language relationships. Biologists had a highly refined system of classification and were beginning to question the immutability of species.

**Tangled Trees—Evolution and the Comparative Method**

Despite the efforts of evolutionists such as Lamarck, the creationist account of the biological world was not seriously challenged until Darwin’s (1809–1882) *Origin of Species by Means of Natural Selection* (1859). Darwin’s theory of evolution by natural selection provided an alternative mechanism that could explain the diversity and complexity of nature without requiring divine influence. Like the linguists, biologists became interested in common ancestry, descent with modification, and family trees. Figure 1 shows one of Darwin’s early sketches from...
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.
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 retentions. 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)
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 pitār, 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 glottoclock, 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. Stevic (1963) presents “... an extended adaptation to linguistic interests of a biologist’s statement about classification” (p. 162). Below is an extract in which Stevic 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 not vitiating 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, 1952; 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,
scenarios in the Bantu language family and Rexov´ae ta l.

Holden (2002) applied similar methods to test migration
competting scenarios for the settlement of the Pacific.

Austronesian languages. They then used this tree to test
(2000) conducted a parsimony analysis of over 5000 dis-

cal, phonological, and lexical characters. Gray and Jordan
Indo-European language tree from discrete grammati-

formations, 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 grammatic-

cal, phonological, and lexical characters. Gray and Jordan
(2000) conducted a parsimony analysis of over 5000 dis-
crete 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 mod-
els 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 mean-
ing evolution between language families, but also, quite
remarkably, that rates of word replacement in these lan-
guages 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 devel-
oped 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 glotto-
clock 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 Hurles et al., 2003, for attempts
to synthesize linguistic, genetic and archaeological infer-
ences about Pacific settlement). Interestingly, quantita-
tive 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 an-
swer similar questions and hence face similar challenges.
One emerging challenge in computational historical lin-
guistics 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 funda-
mentally 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 rela-
tionships between these languages. Unfortunately,
because there is no explicit optimality criterion used to
make the cognacy judgements, it is difficult to evalu-
ate the evidence supporting a relationship objectively.
Identifying cognates between languages has obvious
parallels with the problem of sequence alignment in biol-
ogy. 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 re-
construction 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 lin-
guistics. Burnham and Anderson (1998) describe model
choice as a balance between under- and overfitting pa-
rameters. A model that is too simple may produce biased
results if it fails to capture important parts of the evo-
lutionary 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. Di-
vergence 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 reten-
tion rate was assumed to be roughly constant at around
81% per thousand years for the Swadesh 200 word list (a
522 SYSTEMATIC BIOLOGY VOL. 54

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 evolution. 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](image-url)
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 biological 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


Sanderson, M. 2002a. Estimating absolute rates of evolution and di-

Sanderson, M. 2002b. R8s, analysis of rates of evolution, version 1.50.

Schlegel, F. 1808. Über die sprache und weisheit der indier: Ein beitrag zur begrundung der alterthumskunde. In Amsterdam clas-


Schliecker, A. 1863. Die Darwinische theorie und die sprachwis-
schaft. Hermann Bohla, Weimar.


Shevoroshkin, V. 1990. The mother tongue: How linguists have re-

constructed the ancestor of all living languages. The Sciences 5/May:June:20–27.


Stevick, R. 1963. The biological model and historical linguistics. Lan-


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

Associate Editor: Chris Simon