15

Phylogenetic Models of Language Change

Three New Questions

Russell D. Gray, Simon J. Greenhill, and Quentin D. Atkinson

Abstract

Computational methods derived from evolutionary biology are increasingly being applied to the study of cultural evolution. This is particularly the case in studies of language evolution, where phylogenetic methods have recently been used to test hypotheses about divergence dates, rates of lexical change, borrowing, and putative language universals. This chapter outlines three new and related questions that could be productively tackled with computational phylogenetic methods: What drives language diversification? What drives differences in the rate of linguistic change (disparity)? Can we identify cultural and linguistic homelands?

Introduction

Evolutionary biology has changed remarkably over the last thirty years. Phylogenies have sprung from the margins to center stage. Open any evolutionary journal, or go to any evolutionary meeting, and you will find wall-to-wall phylogenetic trees. Tree thinking (O'Hara 1997) is now the dominant way of making inferences in evolutionary biology (see Figure 15.1). The phylogenetic revolution in biology has been driven by two main events: the development of computational methods and the deluge of molecular sequence data. Today, molecular phylogenies are used to analyze everything from Aardvarks (Seiffert 2007) to Zoogloea (Kalia et al. 2007).

Despite its apparent position on the other side of the arts/science divide, linguistics is also a discipline that requires making complex inferences from a wealth of comparative data. Moreover, as scholars dating back to at least Darwin (1871) have noted, there are numerous "curious parallels" between

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

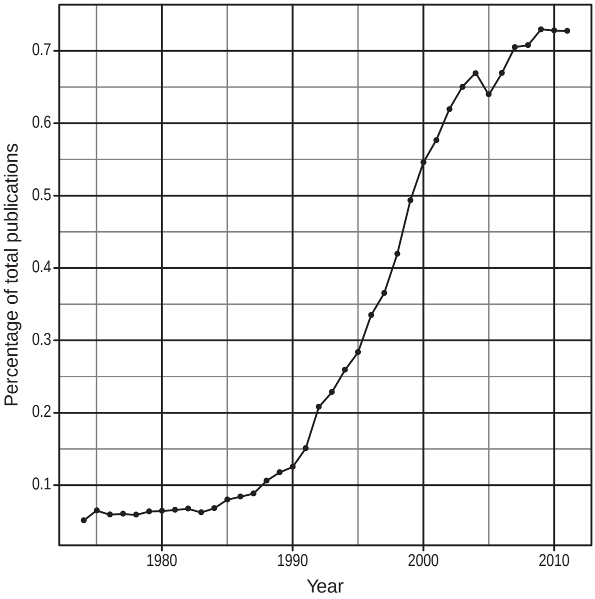

Figure 15.1 A plot showing the percentage increase in papers mentioning the keyword "phylogen*" in the Scopus publication database by year.

the processes of language change and biological evolution (see Atkinson and Gray 2005). Although early attempts to turn historical linguistics into a computational science were far from successful (Swadesh 1952; Bergsland and Vogt 1962; Greenhill and Gray 2009), we are currently witnessing a steady growth in both the use of computational methods and the development of large comparative databases (Greenhill et al. 2008; Dryer and Haspelmath 2011). Computational methods derived from evolutionary biology have been used to construct phylogenetic trees for language families including Aslian (Dunn et al. 2011a), Austronesian (Gray et al. 2009, 2011; Greenhill and Gray 2009, 2010), Bantu (Holden 2002; Holden and Gray 2006), Indo-European (Gray and Atkinson 2003), Japonic (Lee and Hasegawa 2011), Pama-Nyungan (Bowern and Atkinson 2012), Semitic (Kitchen et al. 2009), and even creoles (Bakker et al. 2011). They have been used to:

- Date language divergences and thus test hypotheses about human prehistory (e.g., Gray and Atkinson 2003; Gray et al. 2009).
- Investigate the rates of change in aspects of language (Pagel et al. 2007; Greenhill et al. 2010).
- Quantify patterns of borrowing in languages (Greenhill et al. 2010; Nelson-Sathi et al. 2011; Gray et al. 2010).
- Identify functional dependencies in language and thus test claims about language universals (Dunn et al. 2011b; Levinson et al. 2011).

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

As these approaches have recently been reviewed by Gray et al. (2011) and Levinson and Gray (2012), we will not cover the same ground here. Instead, we outline three new and related questions about language evolution that could be productively tackled with computational phylogenetic methods: What drives language diversification (cladogenesis)? What drives linguistic disparity (anagenesis)? Can we identify cultural and linguistic homelands?

What Drives Language Diversification?

Vast amounts of ink have been spilt, and millions of computer keys pressed, addressing detailed linguistic questions such as the development of Proto-Indo-European laryngeals.[1] We certainly do not wish to diminish the importance of these endeavors; however, we are surprised at how little attention linguists have given to the question of language diversity. Explaining why the human species currently has around 7,000 languages (Lewis 2009) should be a fundamental task for both linguists and theorists of cultural evolution. Moreover, the patchy distribution of this diversity cries out for explanation. According to Lewis (2009), there are 194 language families. Most of these families, 74, have a single member (i.e., are isolates). At the other extreme, Niger-Congo and Austronesian contain over one-third of the total between them (1,495 and 1,246 languages, respectively). This massive disparity between language families suggests that there has been substantial variation in the rates at which languages diversify and go extinct. The large number of isolates suggests that uneven patterns of extinction have had a major role (Nichols 1997). However, diversification rates vary strikingly as well. For example, both Mayan and Malayo-Polynesian are estimated to be around 4,000 years old (Gray et al. 2009; Atkinson et al., in preparation) and yet there are 69 Mayan and 1,226 Malayo-Polynesian languages. Thus, if we assume no extinction, Mayan gave birth to approximately one language every 58 years, whereas Malayo-Polynesian spawned one language every 40 months or so. Patterns of language diversity also vary strikingly in space. For example, the island of New Guinea, despite covering less than 0.5% of Earth's land area, supports over 900 languages (13% of all languages). Comparatively, Russia is over 20 times the size of New Guinea, but only has 105 languages.

Characterizing language diversity is not straightforward (see also Evans, this volume). Following the literature on biodiversity (see MacLaurin and Sterelny 2008), we will distinguish between three types of language diversity: alpha diversity (the number of languages at a location), phylogenetic language diversity (the sum of the path lengths between a set of languages on a phylogenetic tree), and language disparity (the overall amount of variation between

[1] This example is actually one of the triumphs of the comparative method. The brilliant reasoning involved was subsequently confirmed by the discovery of ancient Anatolian languages with two laryngeals.

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

languages). Note that alpha diversity is only the product of language-splitting events (cladogenesis), whereas phylogenetic diversity and disparity are produced by both change within lineages (anagenesis) and cladogenesis (see Figure 15.2). As Nettle (1999) pointed out, language families are not really ideal units for comparative quantitative analyses because the differing time depths of language families means they are not equivalent evolutionary units.

Our focus here is on ways in which phylogenetic methods can help us explore the causes of the drivers of alpha language diversity (the following section will focus on drivers of language disparity). First, biologists have noted that the *shape* of the tree alone provides clues to the diversification dynamics that gave rise to a phylogeny. If a set of languages are diversifying at fairly constant rate, then the tree will be *balanced*; that is, each node (protolanguage) at a given time depth on the tree will tend to have the same number of descendants in each of its daughter lineages. If, however, there are substantial differences in the rate at which some subgroup diverged, then the tree will be *unbalanced* so that one branch will have more descendants than the other (Figure 15.3). For language families that have undergone large expansions, we would expect them to be highly unbalanced.

There is a suite of tools for quantifying the shape of a tree to identify the signature of variation in diversification rates (e.g., Agapow and Purvis 2002; Fusco and Cronk 1995). To date, Holman (2010) conducted the only study to apply these tools to language trees. Holman calculated the imbalance, *Iw* (Fusco and Cronk 1995), from the trees of 19 large language families from the *Ethnologue* database (Lewis 2009), published language phylogenies, and found that almost all of the language trees were significantly more unbalanced

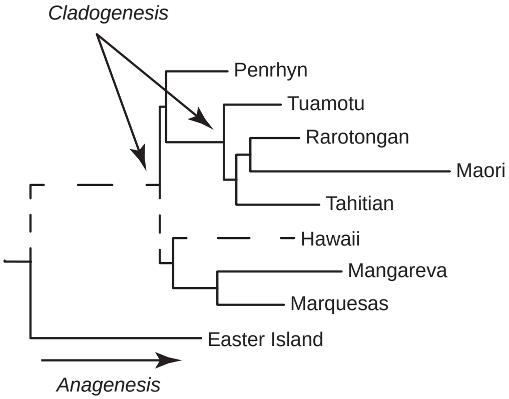

Figure 15.2 A phylogenetic tree for Polynesian languages showing cladogenesis (lineage splitting) and anagenesis (change in a lineage). In this tree the branch lengths are scaled to be proportional to the amount of change in a lineage. The dotted line shows a path from the ancestral language (root of the tree) to a tip (Hawaiian). The length of this path measures the amount of change from the root to the Hawaiian tip.

From “Cultural Evolution: Society, Technology, Language, and Religion,” edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

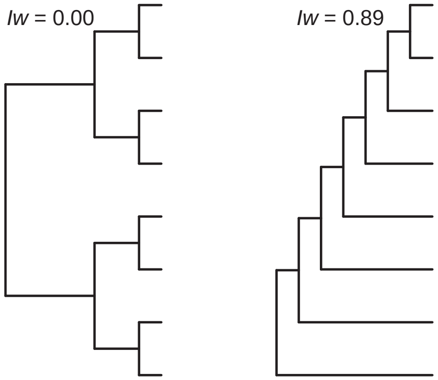

Figure 15.3 Depiction of (a) a perfectly balanced tree versus (b) an extremely unbalanced tree; *Iw* represents imbalance.

than expected by chance. These results indicate that there is substantial variation in diversification rates within families and support the notion that much of the world's linguistic diversity is a result of large-scale expansion events. This statistic, *Iw*, is robust, comparable across trees, and can accommodate unresolved subgroupings (polytomies) and incomplete phylogenies (Fusco and Cronk 1995). We have calculated the same statistic, *Iw*, on some of the major language family trees (see Table 15.1). The *Iw* score varies from 0 for balanced trees to 1 for completely unbalanced trees.

If the tree shows no evidence for differences in rates of language diversification, then the expected value of *Iw* will be 0.5. We can therefore test if the observed tree differs from 0.5 by using a null model of branching that assumes a simple Markov process, where all languages share the same birth rate (Fusco and Cronk 1995).[2] Table 15.1 shows that the most balanced families are Mayan and Austroasiatic. At the other extreme, Austronesian and Semitic are moderately imbalanced.

What factors could have caused this imbalance? One possible explanation is that imbalance is caused by the pruning of branches due to language extinction. Phylogenetic methods can help uncover periods of extinction using birth–death models (Nee 2006). If we were to plot the number of languages over time on a semilog plot, we would then recover a line with a slope proportional to the diversification rate. If the trees grew without any major extinctions (i.e., a pure birth model), we would expect this line to be straight. However, if there is extinction, this line is expected to show an uptick toward the present, as the most recently born languages have not yet had a chance to become extinct. The difference between the diversification rate slope and the rate on this uptick is the extinction rate. Using this logic, we test whether a given phylogeny is best explained by a pure birth model that assumes no extinction or a birth–death

[2] See also http://R-Forge.R-project.org/projects/caper/

Table 15.1 Mean *Iw* scores for various language families. Languages are sorted from most balanced to least balanced.

Family	Languages	*Iw*	Source
Mayan	53	0.33	Atkinson et al., in prep.
Austroasiatic	54	0.39	Sidwell et al., in prep.
Pama-Nyungan	194	0.44	Bowern and Atkinson (2012)
Indo-European	103	0.45	Bouckaert et al. (2012)
Japonic	59	0.47	Lee and Hasegawa (2011)
Semitic	25	0.51	Kitchen et al. (2009)
Austronesian	400	0.59	Gray et al. (2009)

process that allows extinction. In both the Austronesian (Gray et al. 2009) and Mayan families, the pure birth model fits the tree significantly better than a birth–death model ($p < 0.001$), suggesting that extinction has played a relatively minor role.

If the observed differences in tree topology are not caused by extinction, then they must be caused by differences in the rates at which the languages diversify. In Gray et al. (2009), we developed a Bayesian method for modeling diversification as a change-point process along a phylogeny. We applied this method to the Austronesian language phylogeny (Figure 15.4) and identified four regions with significant evidence of increases in diversification rate (i.e., expansion pulses). Our results showed that significant pulses occurred prior to the proto-Malayo-Polynesian branch, before the breakup of the Philippines languages, before the diversification of the Micronesian languages, and the branch leading to the Micronesian and Central Pacific subgroups. We suggest that these pulses could be linked to technological advances, such as the development of the outrigger canoe enabling the Austronesian peoples to cross the channel into the Philippines, and the invention of the double-hulled canoe enabling the expansion into Eastern Polynesia (cf. Pawley and Pawley 1994).

However, although there was evidence to suggest that the pulses were linked to advances in canoe technology, we did not directly test this. A new set of methods, BiSSE and QuaSSE, can directly test the effect of a binary trait (e.g., presence or absence of double-hulled canoes) or a quantitative variable on the rates of diversification (Maddison et al. 2007; FitzJohn 2010). These new methods open up exciting possibilities for comparative analyses as they provide a powerful way of testing hypotheses about the causes of cultural evolution and diversification. Many such factors have been proposed, ranging from the simple acquisition of new technological items like canoes, to social factors such as the level of political complexity (Currie and Mace 2009). One of the most prominent suggestions links the advent of farming to the expansion of language families around the world (Diamond and Bellwood 2003).

From “Cultural Evolution: Society, Technology, Language, and Religion,” edited by Peter J. Richerson and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

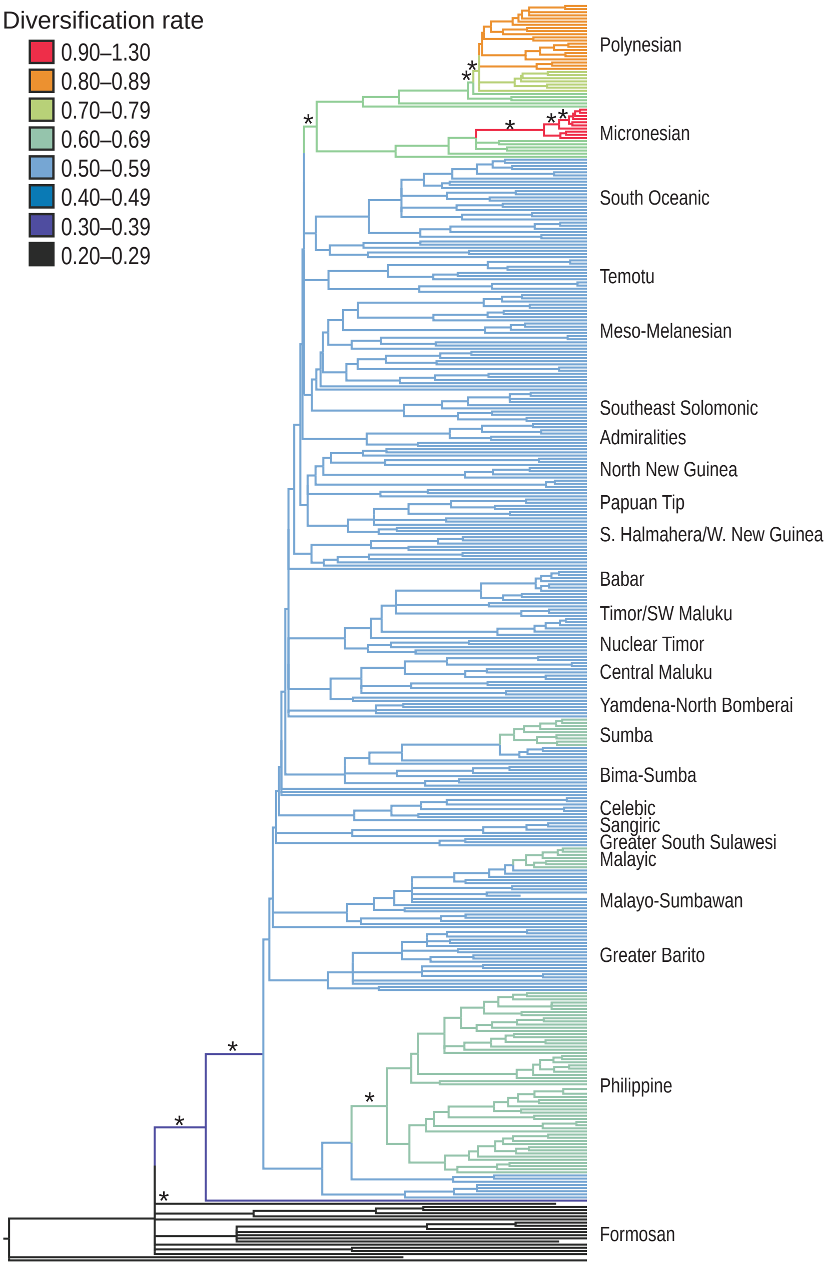

Figure 15.4 Austronesian language phylogeny showing changes in diversification rate due to expansion pulses. Branches with significant shifts in rate are marked with an asterisk. Reprinted with permission from the supplementary material in Gray et al. (2009).

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

This theory suggests that the invention of agriculture enabled the new farmers to obtain higher yields of food and reach much higher population densities. This advantage allowed farmers to outcompete existing hunter-gatherer populations and led to major population expansions out of agricultural homelands. Diamond and Bellwood (2003) claim that the signature of these farming-driven expansions is evident in the distribution of no less than 13 of the major language families: Afro-Asiatic, Austro-Asiatic, Austronesian, Bantu, Dravidian, Indo-European, Japanese, Nilo-Saharan, Sino-Tibetan, Tai, Trans New Guinea, and Turkic. The new BiSSE and QuaSSE methods provide the means to test these prominent and long-standing hypotheses about the factors that have shaped our modern-day language diversity.

In biological evolution, diversification rates are relatively constant over time after a burst of diversification. It has been suggested that the burst occurs as the species diversifies into new niches. After this initial burst these niches become filled and therefore constrain further diversification (Etienne et al. 2012). Evidence for this "density dependence" comes from many molecular studies showing a slowing down of diversification rates in many species. For example, a meta-analysis of bird families showed significant decreases in diversification in 23 out of 45 families with bigger decreases in larger families consistent with density-dependent constraints on diversification (Phillimore and Price 2008). To date, this idea has not been applied to cultural evolution. This omission is striking as there are strong hints that density dependence operates on cultural diversity. For example, a study of 264 islands in the Pacific found that 195 (74%) had only one language (Gavin and Sibanda 2012). This suggests that once a language or culture fills a niche, it heavily restricts the birth of new languages or cultures. Thus one possible explanation for the immense diversity of the Austronesian language family might be that the invention of better canoe technology combined with a shift to agriculture opened a range of new niches in the Pacific that facilitated the diversification of these cultures. In contrast, the substantially less diverse Mayan family had to compete for niches with hunter-gatherer groups and other agriculturalist populations belonging to the Mixe-Zoquean, Oto-Manguean, and Uto-Aztecan language families.

What Drives Linguistic Disparity (and What Constrains It)?

In his influential book, *Wonderful Life*, Stephen Jay Gould (1989) distinguished between diversity and disparity. He argued that the number of species was a poor measure of the overall amount of phenotypic variation. Diversity in overall body plan does not necessarily correlate well with the number of species in a clade. Whereas there might be millions of species of beetles, they are all still beetles. A similar distinction could be made in linguistics. With over 100 languages spoken across its islands, Vanuatu has one of the highest densities of languages in the world (Lewis 2009). However, all these languages belong to

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

just two subgroups (North/Central Vanuatu and South Vanuatu) of the Oceanic group, which is itself a subgroup of Austronesian. How, then, should we measure language disparity?

Languages differ not only in their lexicon but also on numerous structural levels, including the organization of the sound system (phonology), systems for the combination of meaningful elements into words (morphology) and phrases (syntax), as well as systems for indicating spatial and temporal relationships, speaker attitude, and epistemological status (see Evans, this volume). It is not possible to combine these variables into a global measure of linguistic disparity, just as it is not possible to come up with a global measure of biological disparity (see MacLaurin and Sterelny 2008). We are thus skeptical whether it is possible to conceptualize the "absolute design space" for all possible languages. It is, however, possible to develop measures of disparity relative to particular traits and for specific questions, just as David Raup did in his famous diagram of possible and actual ammonoid shell morphologies (Raup 1967).

These local representations of morphospace have provided theoretical morphologists and evolutionary biologists with powerful tools for analyzing both the drivers and constraints on morphological evolution. Phylogenies can be used to trace phenotypic evolution through these spaces and infer factors that accelerate or constrain the evolution of disparity. A similar approach could be adopted in studies of linguistic and cultural evolution (see Hauser 2009; Levinson 2012b). Just as Kemp and Regier (2012) constructed a design space of possible kinship systems, linguists could construct phonological and typological spaces. For example, we could classify the world's languages based on primary word order (i.e., the order of the Subject, Object, or Verb in a sentence). There are six possible ways of structuring this information, however, not all combinations are as likely (Dryer 1992, 2011):

- 41% order the elements as SOV, while 35% use SVO.
- 13% use no dominant order.
- Less frequent are VSO (7%), VOS (2%), and OVS (0.8%).
- The least common is OSV, with only 0.2% of the world's languages choosing this ordering.

This difference in the frequency of word orders requires explanation. Whereas linguists often claim that these patterns reflect cognitive and functional constraints, the role of historical contingencies needs to be evaluated as well (see Levinson and Gray 2012).

Let us extend this idea of word-order space to many aspects of language typology. If we take the World Atlas of Language Structures (WALS) as an example, then there are 140 different traits that characterize language. Each of these traits has on average 4.6 states. If we trace all possible combinations of these traits, then there are 2.5×10^{89} possible ways of constructing a language. However, in this "WALS space" of possible languages, not all regions will be equally likely. Phylogenetic methods could be used to map the movement of

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

language lineages through this space and thus evaluate the roles of cognition, function, and history in explaining the patterns of disparity that we see among the languages of the world today.

If the rates of both language cladogenesis and anagenesis are constant, then measures of language diversity and disparity will be congruent. Language diversification is unlikely to be constant (see above), and the rate of change in lineages is known to vary markedly (Blust 2000). Phylogenetic methods can be used both to estimate rates of change and to test hypotheses about the factors that influence them. Thus, rather than rate variation being a nuisance, it can become an object of study (just as it is in evolutionary biology). There are numerous hypotheses about the factors that might affect rates of linguistic change. Trudgill (2011), for example, lists five major factors: group size, density of social networks, amount of shared information, social stability, and levels of contact with other speech communities. There is, however, no consensus on which factors most influence rates of change, and little has been done to quantify the relative roles and interaction between these factors. Bayesian phylogenetic model comparison offers a way forward. Rather than fitting a model with a single rate, multiple rates can be estimated for different branches on the tree. Where there is a prior hypothesis about a factor that might affect the rate of linguistic change, the posterior probability of a single rate model can be compared with one that fits different rates for branches with different values of that parameter. For example, if the hypothesis suggested that hunter-gatherer languages had higher or lower rates of lexical replacement than agricultural ones, the hypothesis could be tested by constructing a language phylogeny from lexical data and comparing the posterior probability of a single rate model with one that allowed different rates for hunter-gatherer versus agricultural languages. Alternatively, where there are no prior hypotheses, the analysis could be done in an exploratory fashion using the local random clock approach proposed by Drummond and Suchard (2010), where a Bayes factor is estimated for the probability of a multiple local rates versus a single rate model.

How might social processes affect historical patterns of language change? An important insight from sociolinguistics is that language functions as a mechanism for marking social boundaries (Labov 1963). Human groups under pressure often exaggerate the language differences to make ethnic barriers—a process Bateson (1935) dubbed *schismogenesis* and Thurston (1987) labeled *esoterogeny*. The effect of this process is likely to be particularly marked when speech communities split. If speech communities exaggerate differences at the time when they are drifting apart, then lineages that have been through more splitting events will undergo more change (see Figure 15.5). Atkinson et al. (2008) used phylogenetic methods to quantify the impact of this effect. They used basic vocabulary data to construct phylogenies for the Austronesian, Bantu, and Indo-European language families. Their results revealed that between 10–33% of the vocabulary differences in these families arose during

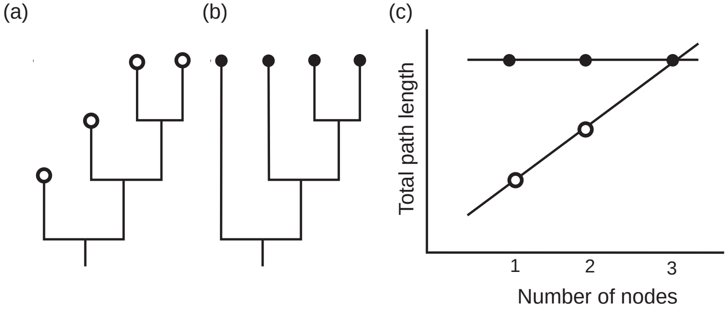

Figure 15.5 The phylogenetic tree in (a) shows the pattern produced by an increase in the rate of linguistic change at the splitting of speech communities; the branch lengths are longer in lineages that have been through more splitting events. (b) In contrast, if the rates of change are not affected by the number of splitting events, then all of the tips of all the branches will be equal irrespective of the number of splitting events the lineage has been through. (c) The size of any schismogenic/esoterogenic effect can be quantified by plotting the path length from the root of the tree to each of its tips against the number of nodes (splitting events) through which the path goes. The slope of the resulting graph estimates the magnitude of the effect.

rapid bursts of change associated with language-splitting events. One interesting extension of this approach would be to see if it holds equally for all aspects of language. To the extent that closely related speech communities differ more in accent than they do in vocabulary, and more in vocabulary than in language structure, it might be predicted that the schismogenesis effect would be most pronounced in phonetics and least in structural features of language.

Can We Infer Cultural and Linguistic Homelands? More Generally, How Do Language Expansions Unfold across a Landscape?

Questions about the origins of human groups and the languages they speak have an enduring fascination. The early European explorers in the Pacific speculated on the origins of the Polynesians after noticing that many words were shared across remote Oceania (Andrews 1836), and for over two hundred years scholars have debated the origins of the Indo-European languages (Jones 1786/2013). Diamond and Bellwood (2003) dub Indo-European the “most recalcitrant problem in historical linguistics.” Linguists typically attempt to make inferences about possible homelands by using arguments based on either linguistic palaeontology or area-of-maximum-diversity. The diversity argument postulates that the most likely point of origin of a language family is the area of greatest diversity (Sapir 1916/1949). Linguistic paleontology arguments rely on reconstructions of words tied to specific locations, such as animal and plant names, to locate the homeland. Both arguments are far from

From “Cultural Evolution: Society, Technology, Language, and Religion,” edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

infallible. How language family diversity is measured in linguistics is often subjective, and the apparent center of diversity can move as language families expand (Nichols 1997). Reconstruction of the form of ancestral words is a rigorous process based on inferences about sound change. However, the reconstruction of the semantics of these forms is much more speculative; for example, does the proto-Indo-European reconstruction for horse (PIE *éḱwos) actually refer to domesticated horses, wild horses, or some more generic four-legged mammal (see Heggarty and Renfrew 2013)?

Linguists are hardly alone when it comes to rather loose inferences about geography. In biology, studies of phylogeography often consist of a rigorously derived phylogenetic tree and a geographical just-so story. The recent advent of stochastic models have, however, enabled more rigorous phylogeographic inferences (Lemey et al. 2009, 2010). These models have proved particularly adept at tracing the spread of human viruses such as the H1N1 outbreak (Lemey et al. 2009) and the yellow fever virus (Auguste et al. 2010). Virus evolution is perhaps a closer analog to language evolution than is vertebrate evolution (Gray et al. 2007). The obvious question that arises is: Could these phylogeographic methods be adapted to make inferences about linguistic geography?" Walker and Ribeiro (2011) used a relaxed random walk (RRW) model in the Bayesian phylogenetic program BEAST (Drummond and Rambaut 2007) to make inferences about the expansion of the Arawak language family. The RRW model is essentially a Brownian diffusion model in which the rate of diffusion can vary along branches of a tree. Rather than assuming a constant rate of diffusion, rate heterogeneity among branches is accommodated via a single additional rate distribution parameter, $P(r)$, allowing support for rate variation and the degree of rate variation (or "relaxation") to be estimated from the data itself. This approach treats language location as a continuous vector (longitude and latitude) which evolves through time along the branches of a tree. It seeks to infer ancestral locations at internal nodes on the tree, simultaneously accounting for uncertainty in the tree. Thus, the phylogeny and the geographic diffusion are co-estimated. Although there was considerable spread in the posterior distribution of ancestral root locations, Walker and Ribeiro found that the most likely origin of Arawak was in Western Amazonia, with subsequent expansion into the Caribbean and across the lowlands. Interestingly, although Northwest Amazonia has the largest number of Arawak languages, the phylogeographic models did not support the region as a potential homeland.

Could the same approach be used to shed light on the "recalcitrant problem" of the Indo-European homeland? We think so. As part of a large team of mathematical biologists and linguists we have recently assembled a large data set of cognate-coded basic vocabulary for 103 ancient and contemporary Indo-European languages (Bouckaert et al. 2012). To increase the realism of the spatial diffusion modeling, we extended the RRW process in two novel ways. First, to reduce potential bias associated with assigning point locations

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

to sampled languages, we used geographic ranges of the languages to specify uncertainty in the location assignments. Second, to account for geographic heterogeneity we accommodated spatial prior distributions on the root and internal node locations. By assigning zero probability to node locations over water, we incorporated prior information about the shape of the Eurasian landmass into the analysis. Although we do not allow for different rates of movement across specified land types, this approach could, in principle, be extended to incorporate other geographic features such as mountains, rivers, or deserts.

Although there are numerous hypotheses about the origins of the Indo-Europeans, most of the current debate revolves around two theories. The "Steppe hypothesis" proposes an Indo-European origin in the Pontic steppe region north of the Caspian Sea, perhaps linked to an expansion into Europe and the Near East by "Kurgan" seminomadic pastoralists, beginning 5–6 KYA. Evidence from "linguistic palaeontology" and putative early borrowings between Indo-European and the Uralic language family of northern Eurasia (Koivulehto 2001) are argued to support a steppe homeland (Anthony 2007). However, the reliability of inferences derived from linguistic palaeontology and claimed borrowings remain controversial (Heggarty and Renfrew 2013). The "Anatolian hypothesis" holds that Indo-European languages spread out of Anatolia (in present-day Turkey) with the expansion of agriculture, beginning 8–9.5 KYA. Our results unambiguously support an Anatolian origin (see Figure 15.6). To quantify the strength of support for an Anatolian origin, we calculated the Bayes factors comparing the posterior to prior odds ratio of a root location within the hypothesized Anatolian homeland (yellow polygon, Figure 15.6) with two versions of the Steppe hypothesis (blue polygons). The Anatolian homeland was over 150 times more likely in both these analyses. Note that the relaxed diffusion model supports substantial variation in rates of diffusion through time and fits the data significantly better than a model which assumes a constant rate of diffusion, even accounting for the extra rate variation parameter. Nevertheless, there is enough regularity in the inferred rates to allow substantial support to emerge for one hypothesis over another. Additionally, it is not simply the case that these methods return the geographic midpoint of the language distributions. The geographic centroid of the languages we analyzed falls within the broader Steppe hypothesis (green star, Figure 15.6); this indicates that our model is not simply returning the center of mass of the sampled locations, as would be predicted under a simple diffusion process that ignores phylogenetic information and geographic barriers.

The RRW approach avoids internal node assignments over water but assumes the same underlying migration rate across water as land. To investigate the robustness of our results to heterogeneity in rates of spatial diffusion, we developed a second inference procedure that allows migration rates to vary over land and water. We examined the effect of varying relative rate parameters

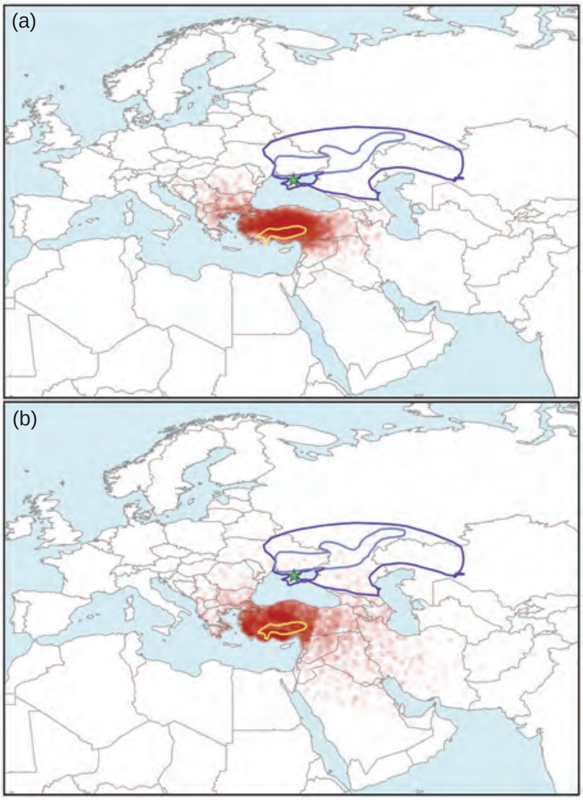

Figure 15.6 (a) Map showing the estimated posterior distribution for the location of the root of the Indo-European language tree. Each point sampled in the posterior is plotted in translucent red such that darker areas correspond to increased probability mass. (b) The same distribution under a landscape-based analysis in which movement into water is 100 times less likely than movement into land. The blue polygons delineate the proposed origin area under the Steppe hypothesis: dark blue shows the initial suggested homeland whereas light blue shows a later version of the Steppe hypothesis. The yellow polygon delineates the proposed origin under the Anatolian hypothesis. A green star in the steppe region shows the location of the centroid of the sampled languages. Reprinted with permission from Bouckaert et al. (2012).

to represent a range of different migration patterns. Figure 15.6b shows the inferred Indo-European homeland under a model in which migration from land into water is 100 times less likely than from land to land. Once again the Anatolian origin is overwhelmingly more likely.

Thus, phylogeographic modeling not only enables us to make probabilistic inferences about ancestral homelands, it also enables us to investigate the robustness of these inferences to a range of assumptions about the spread of languages. Figure 15.7 shows how these phylogeographic models can even be used to plot the spread of an entire language family in space and in time. This

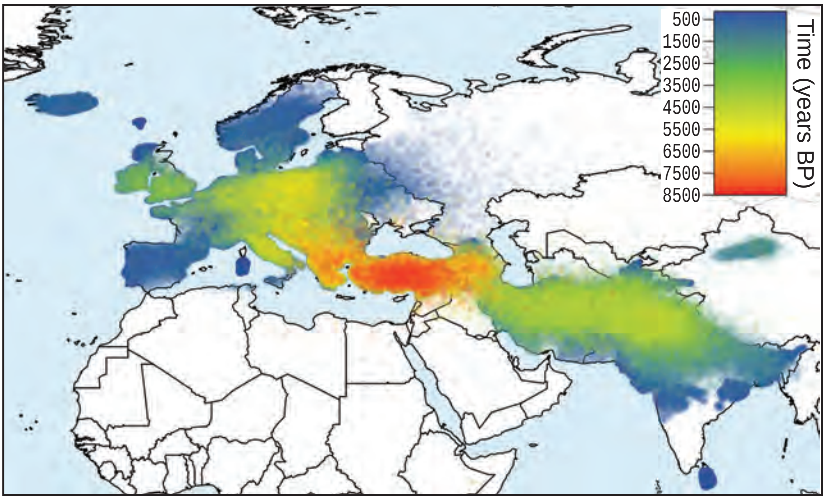

Figure 15.7 Spatial and temporal reconstruction of the expansion of Indo-European languages. The posterior distribution of node location estimates through time is plotted as opaque points with a color that indicates their corresponding age estimate. Older nodes are shown on the foreground to depict clearly the temporal diffusion pattern. Reprinted with permission from the supplementary material of Bouckaert et al. (2012).

figure needs to be interpreted with the caveat that we can only represent nodes corresponding to divergence events between languages that are in our sample. Nodes that are associated with branches not represented in our sample will not be reflected in this figure. For example, the lack of Continental Celtic variants in our sample means we miss the Celtic incursion into Iberia, and instead infer a late arrival into the Iberian Peninsula associated with the Romance languages. The chronology represented here, therefore, offers a minimum age for expansion into an area. Expanding and enhancing these methods to accommodate other aspects of geographic heterogeneity and other language expansions will allow us to test increasingly detailed hypotheses about human prehistory and the processes that drive language diversity and disparity in space and time.

It may even be possible to infer population migration events on a global scale. Atkinson (2011) highlights a global trend of decreasing phoneme diversity with distance from Africa, which is consistent with a serial founder effect in phoneme diversity following the human expansion from Africa. The observed relationship fits with theoretical models of cultural and linguistic transmission (De Boer 2001; Henrich 2004b) and holds after controlling for modern population size, density, and language relatedness. While the finding is, of course, only correlational and remains controversial (e.g., Wang et al. 2012), there are clear geographic trends in language variation across the globe that require explanation.

From "Cultural Evolution: Society, Technology, Language, and Religion," edited by Peter J. Richerson
and Morten H. Christiansen. 2013. Strüngmann Forum Reports, vol. 12, J. Lupp, series editor.
Cambridge, MA: MIT Press. ISBN 978-0-262-01975-0.

Conclusion

The combination of large databases and computational methods has revolutionized inferences in evolutionary biology. While we should not ignore the numerous subtle differences between biological and cultural evolution, the three questions we have framed in phylogenetic terms show that there is much to be gained from the nuanced application of this approach to questions about the evolution of languages across the globe. Such an approach would provide a powerful way of resolving questions about human prehistory by integrating genetic, linguistic, and cultural data in a common analytical framework. This ambitious undertaking is not without obstacles, such as the rigorous inference of cognate vocabulary and the detection of borrowing, but already computational approaches are rising to these challenges (Bouchard-Côté et al. 2013; Nelson-Sathi et al. 2011).

Acknowledgments

Some of the ideas in this manuscript were developed at the NESCent-funded workshop on "Modeling the diversification of languages." We thank Claire Bowern, Michael Dunn, and Michael Gavin for useful discussion and the editors and an anonymous referee for comments.