New Perspectives on Indo-European Phylogeny and Chronology[1]

ANDREW GARRETT

Professor of Linguistics
Nadine M. Tang and Bruce L. Smith Professor of Cross-Cultural Social Sciences
University of California, Berkeley

O ne of the oldest questions in linguistics concerns the origins of Indo-European languages—a large family of several hundred languages, including English and the other Germanic languages, Spanish and the other Romance languages, and their ancestors (including Old English and Latin) and relatives spoken two thousand years ago across much of Europe and central and south Asia. The earliest Indo-European languages are known from historical documents beginning in the first half of the second millennium BCE; today they are spoken around the world. But just as the Romance languages descended from Latin, so all the Indo-European languages must have a single prehistoric ancestor. When and where was that language spoken? In answering that question here, I explore the relationship between traditional methods of linguistic reconstruction and newer tools that have begun to transform the field.

Language Diversity and Diversification

The development of comparative and historical linguistics was a by-product in the late 18th and early 19th centuries of European colonial adventurism. This was especially important in three sites. One was the Pacific, where colonial and scientific explorations as early as Captain Cook's voyages exposed Europeans to Austronesian languages and their relationships.[2] A second was India, where mercantile incursions and military conquests brought Europeans to encounter Sanskrit—the earliest attested language of the Indo-Iranian branch of Indo-European, and the ancestor of modern Indo-Aryan languages—and

---

1   Read 29 April 2017 as part of the *Indo-Europeanization of Europe* symposium.

2   See Patrick Vinton Kirch, *On the Road of the Winds: An Archaeological History of the Pacific Islands before European Contact*, revised and expanded edition (Oakland: University of California Press, 2017), 188–89.
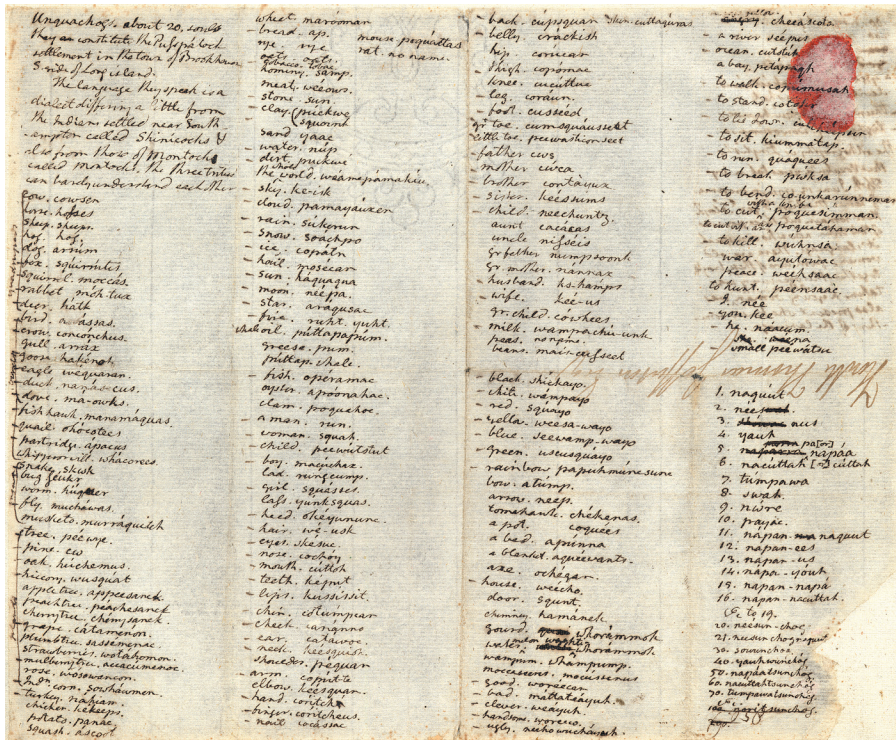
FIGURE 1. Language documentation notes from Thomas Jefferson's "Vocabulary of the Unquachog Indians," 1791. American Philosophical Society.

to start working out the relationships between Sanskrit and languages of India and Iran and their European relatives.

The third site was America, where the European invasions led to encounters with a large number of very different languages. I illustrate with material famously held at the American Philosophical Society. Figure 1 shows Thomas Jefferson's Unquachog vocabulary, recorded in Brookhaven, Long Island; this gives us lexical information about an Algonquian language that has not been used since the 18th century. In his description of his own fieldwork, Jefferson clearly indicates his interest: "The language they speak is a dialect differing a little from the Indians settled near Southhampton . . . also from those of [Montauk]. . . . The three tribes can barely understand each other." Whether or not this last claim was true, Jefferson's interest was in understanding the relationships among the languages he encountered.[3] He put the matter more generally as follows:

3   Quotations that follow are from his *Notes on the State of Virginia*. See Thomas Jefferson, *The Writings of Thomas Jefferson*, ed. H. A. Washington (Cambridge: Cambridge University Press, 2011), vol. 8, 344–45.

Algonquin	orni mis	ajack	pen ainse
Tawa	mimi	hiwauki	papasé
Shawanee			
Nanticoke	wectah-tomps	ur capitz	hitty cawndipqua
Mohiccon			pah-pah-cogh
Unquachog	má-ewks	ápacus	ohócotees
Oneida	oleedagó	ohquás	
Cayuga			qua-sen
Onondaga	tschiéochara	nonniagaree	
Miami			
Cherokee	visaki memiah	bokkiah	
	co, táquos; cone, heh coo, la	tan, tis, te che, is, qua	cuqueh. ho, que
Chickasaw	pat, che, u, so, buh	co, fis, to	co, fe
Choctaw	patche, u, so, buh	co, fit, cho, to	co, fe
Creek	pat, che ho, we	o, ke, te, neat, cuh	co, wag, ge
Tuscarora			oh-qua-sen
Chetimacha		heihankape	pupe conhimamche
Atacapa		enkestt	nelkitson

FIGURE 2. Words for different birds in selected Native American languages, from Thomas Jefferson's "Comparative Vocabularies of Several Indian Languages, 1802–1808." American Philosophical Society.

> The knowledge of the languages of North America would be the most certain evidence of [Native American people's] derivation which ever can be produced. In fact, it is the best proof of the affinity of nations which ever can be referred to. How many ages have elapsed since the English, the Dutch, the Germans, the Swiss, the Norwegians, Danes and Swedes have separated from their common stock? Yet how many more must elapse before the proofs of their common origin, which exist in their several languages, will disappear?

The two questions asked here highlight his concern, already at this time, with chronology and with methods for determining out the chronology of undocumented earlier languages.

From this concern came a challenge to generations of linguists: "Were vocabularies formed of all the languages spoken in North and South America . . . it would furnish opportunities . . . to construct the best evidence of the derivation of this part of the human race." Jefferson himself contributed to this project through comparative vocabularies like the one in Figure 2, showing vocabulary for birds in a variety of

languages. Many linguists have taken this project seriously; if we can learn enough about many languages, we may be able to draw historical inferences about the people who spoke them.

Based on modern versions of comparative data like the vocabulary in Figure 2, linguists ask several kinds of historical questions, including questions about phylogeny (the relationships of languages) and, once relationships are identified among languages, the location in space and time of their ancestors. We can identify families of languages, like the Austronesian languages of the Pacific, the Algic languages of North America, and the Indo-European languages of Eurasia. When and where do they come from?

This question seemed particularly acute to European observers of North America, which is home to many languages and far more language families than Jefferson might have expected. In North America as a whole, before the European invasion there were some 300 distinct languages, including over two dozen isolates (languages with no known relatives) in addition to languages belonging to over two dozen unrelated language families.[4] By comparison, in present-day Europe, most languages belong to a single-language family, Indo-European; there is only one isolate, Basque; and the other languages belong to only six language families in all (Northwest and Northeast Caucasian, Mongolic, Semitic, Turkic, and Uralic). The discrepancy between the linguistic diversity of the Old and New Worlds was highlighted as follows by Jefferson:

> [T]here will be found probably twenty [language families] in America, for one in [East] Asia. . . . 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. . . . A greater number of those radical changes of language having taken place among the [people] of America, proves them of greater antiquity than those of Asia.

We need not accept the statistics today, but the underlying question remains: If we have identified a language family, can we figure out when its ancestor was spoken and how long its descendants have taken to diversify?

---

4   See Ives Goddard, ed., *Languages*, Handbook of North American Indians, vol. 17 (Washington: Smithsonian Institution, 1996), and Marianne Mithun, *The Languages of Native North America* (Cambridge: Cambridge University Press, 1999).

Indo-European Phylogenetics

The ancestral chronology of the Indo-European language family has been of interest since the 19th century, given the historical significance of languages belonging to this family (from Greek, Latin, and Sanskrit to the present day); see Brian D. Joseph's contribution in this issue. The map in Joseph's Figure 3 highlights the broad distribution of Indo-European languages. Already by 2000 BCE, the language family had spread over a large area, from Ireland and Spain in the west to India and western China in the east. The earliest languages of the family (belonging to the Anatolian, Greek, and Indo-Iranian branches) were already very different when first documented in the second millennium BCE, so their last common ancestor must predate them by a significant time span.

There are two main models for interpreting the origin and early diversification of the Indo-European languages. The *steppe model*, favored by most linguists (and many archaeologists), suggests that Indo-European originated in the steppe north of the Black and Caspian Seas around 4500 BCE (6500 BP) and was spread by pastoralists who had learned to breed horses and build wheeled vehicles. According to the *Anatolian model*, developed by the archaeologist Colin Renfrew in the 1980s, Proto-Indo-European was spoken 3,000 years earlier, south of the Black Sea in Anatolia, and spread into Europe with the initial diffusion of agriculture.[5] The two models thus posit both a significant chronological difference and distinct sociocultural matrices for language transmission.

The question of chronology is thus crucial. Given what we know about Indo-European languages beginning in the second millennium BCE, if linguistic data can tell us when their ancestor was spoken, we may be able to decide between the two models of diversification. A variety of arguments from linguistic data have been used to reconstruct chronology, each with strengths and weaknesses. Among the best-known arguments in the Indo-European case is the argument from what is sometimes called "linguistic palaeontology." For example,

---

5   For the steppe model, see J. P. Mallory, *In Search of the Indo-Europeans: Language, Archaeology and Myth* (London: Thames & Hudson, 1989); David W. Anthony and Don Ringe, "The Indo-European Homeland from Linguistic and Archaeological Perspectives," *Annual Review of Linguistics* 1 (2015): 199–219; and earlier literature they cite. For the Anatolian model, see Colin Renfrew, *Archaeology and Language: The Puzzle of Indo-European Origins* (London: Cape, 1987) and many subsequent papers; a recent defense of the Anatolian model is by Paul Heggarty, "Prehistory through Language and Archaeology," in *The Routledge Handbook of Historical Linguistics*, eds. Claire Bowern and Bethwyn Evans (London: Routledge, 2015), 598–626.

comparison of Indo-European languages has led some linguists to conclude that their common ancestor had a word *$k^wék^wlos$ ("wheel") and other terms associated with wheeled transport.[6] If this is correct, then Proto-Indo-European must have been spoken in a time and place with wheeled vehicles; this is known to postdate the spread of agriculture. But the number of crucial vocabulary items is relatively small, and each one is disputed by skeptics; so, skeptics claim, perhaps the word for "wheel" actually referred to some other object. Arguments over these details have not resolved the matter.

Ancestry Relationships in Phylogenetics

Another argument about ancestral chronology relies on the extent of differentiation among descendant languages. This was traditionally an impressionistic exercise. If the languages in a newly identified language family seem about as diverse as the Romance languages do today, then, since we know the date of the Romance ancestor (Latin, spoken around 2,000 years ago), we may guess a similar date for the ancestor of the new family. This kind of inference has two key ingredients: a way of measuring similarity; and some assumption about rates of linguistic change, based on the intuition that similar levels of differentiation between pairs of languages should mean they have been diverging for similar time periods. To avoid the problem that similarity impressions were unquantified, some linguists in the 1950s began using an approach called *lexicostatistics*, but this had two flaws: it was what we would now call a distance-based method—it quantified linguistic differences rather than reconstructing an evolutionary tree—and it relied on the assumption of a universally constant rate of language change. The latter assumption in particular is known to be false.[7]

More recently, Bayesian phylogenetic methods have been adapted for linguistic research from biological systematics, with impressive

6   See Bill Darden, "On the Question of the Anatolian Origin of Indo-Hittite," in *Greater Anatolia and the Indo-Hittite Language Family*, ed. Robert Drews (Washington, DC: Institute for the Study of Man, 2001), 184–228; Asko Parpola, "Proto-Indo-European Speakers of the Late Tripolye Culture as the Inventors of Wheeled Vehicles: Linguistic and Archaeological Considerations of the PIE Homeland Problem," in *Proceedings of the Nineteenth Annual UCLA Indo-European Conference*, eds. Karlene Jones-Bley, Martin E. Huld, Angela Della Volpe, and Miriam Dexter Robbins (Washington, DC: Institute for the Study of Man, 2008), 1–59; and Anthony and Ringe, "Indo-European Homeland."

7   For classic lexicostatistical analysis, see Morris Swadesh, "Salish Internal Relationships," *International Journal of American Linguistics* 16 (1950): 157–67; Robert B. Lees, "The Basis of Glottochronology," *Language* 29 (1953): 113–27; and Swadesh, "Towards Greater Accuracy in Lexicostatistic Dating," *International Journal of American Linguistics* 21 (1955): 121–37. For a classic critique, see Knud Bergsland and Hans Vogt, "On the Validity of Glottochronology," *Current Anthropology* 3 (1962): 115–53.

results in a variety of language families.[8] A simple example shows generally how such work proceeds. This involves vocabulary traits, which are relatively easy to assemble (from dictionaries) and are therefore often used. Consider the following words in four Polynesian languages and their reconstructed ancestor, Proto-Polynesian:

	Hawaiian	Māori	Samoan	Tongan		Proto-Polynesian
A	niu	niu	niu	niu	"coconut"	*niu
	pua	pua	pua	pua	"flower"	*pua
	kapu	tapu	tapu	tapu	"taboo"	*tapu
	maka	mata	mata	mata	"eye"	*mata
	1	1	1	1		
B	makika	namu	namu	namu	"mosquito"	*namu
	pulelehua	pepe	pepe	pepe	"butterfly"	*pepe
	1	2	2	2		

The words in each row of A are "cognates": all four languages have the same word, with observed differences due to known or inferable changes (like the shift of *t* to *k* in most Hawaiian dialects); as shown in the last column, the ancestor word was probably the same or very similar. In B, Hawaiian has a different word from the other three languages; this can be encoded numerically by indicating that Hawaiian has character state 1 and the other languages have character state 2; some language or languages changed state. Data set B is then an informative data set, unlike data set A (in which all four languages show the same state). Given many sets like B, an analysis is possible.

A phylogenetic analysis has several key ingredients. The first is a set of linguistic traits, often from basic vocabulary, ideally including hundreds or even thousands of informative sets. These are reduced to numerical characters. A second is a model of trait evolution, expressing assumptions about how linguistic traits change over time; a third is a clock model, expressing assumptions about how rates of change vary across the whole language tree and from trait to trait. The clock model is crucial if a goal is chronology: to hypothesize when the ancestral language was spoken, assumptions about rates of change are needed.[9] Finally, there are hard constraints, for example based on historical evidence—that Old Irish was spoken from 700 to 900 CE, or that Clas-

---

8   For overviews of this field, see Michael Dunn, "Language Phylogenies," in *The Routledge Handbook of Historical Linguistics*, eds. Claire Bowern and Bethwyn Evans (London: Routledge, 2015), 190–211; and Claire Bowern, "Computational Phylogenetics," *Annual Review of Linguistics* 4 (2018): 281–96.

9   This is different from assuming a universally constant rate of change, as in traditional lexicostatistics. Here, it is feasible to assume that some languages or some linguistic traits change faster than others, as long as the overall variation in rates of change can be modeled statistically.

sical Latin was spoken in the first century BCE. Given such ingredients, computer programs implement Bayesian phylogenetic methods to evaluate millions of possible language trees for likelihood (of the data, given the tree and model parameters). A set of probable language trees can then be identified.

Work along these lines was introduced to linguists in a 2000 paper on the Austronesian language family (of which Polynesian is a small part).[10] Using a novel methodology, that work yielded results bolstering traditional linguistic analyses that had not involved massive analysis of large lexical data sets. When some of the same authors and other collaborators turned to the Indo-European language family, they produced work that was very widely publicized because its results—unlike in the Austronesian case—contradicted the consensus of linguists.[11] Most importantly in a 2012 paper in *Science*, a team of scientists then based in New Zealand claimed that Bayesian phylogenetic methods point to a Proto-Indo-European date of about 6000 BCE (8,000 years ago). This would support the Anatolian (first agriculturalists) model of Indo-European origins, and would contradict the steppe (pastoralists) model.

The informal reaction to this work in some linguistic circles has been outright dismissal. Some linguists who accept the steppe model seem to have concluded that the newer methods are flawed in fundamental ways.[12] Others have simply been puzzled; the methods require a broader toolkit than many of us have. In collaboration with other linguists and computer scientists at Berkeley, my own approach has been to try to understand these methods and work out why they produce the results they do.[13]

10   See Russell D. Gray and F. M. Jordan, "Language Trees Support the Express-train Sequence of Austronesian Expansion," *Nature* 405 (2000): 1052–55. See also R. D. Gray, A. J. Drummond, and S. J. Greenhill, "Language Phylogenies Reveal Expansion Pulses and Pauses in Pacific Settlement," *Science* 323 (2009): 479–83.

11   See Russell D. Gray and Quentin D. Atkinson, "Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin," *Nature* 426 (2003): 435–39, and Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin Atkinson, "Mapping the Origins and Expansion of the Indo-European Language Family," *Science* 337, no. 6097 (2012): 957–60.

12   An example of the more dismissive approach is found in a book by Asya Pereltsvaig and Martin W. Lewis, *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics* (Cambridge: Cambridge University Press, 2015); see the review article by Claire Bowern, "The Indo-European Controversy and Bayesian Phylogenetic Methods," *Diachronica* 34 (2017): 421–36.

13   The work described here was published as Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett, "Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis," *Language* 91, no. 1 (2015): 194–244.
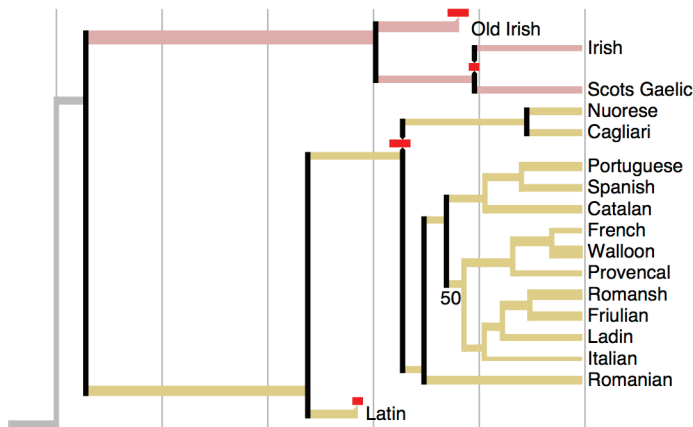
FIGURE 3. Detail of Analysis C3 summary tree (based on Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin Atkinson, "Mapping the Origins and Expansion of the Indo-European Language Family," *Science* 337, no. 6097 [2012]: 957–60). Reproduced by permission from Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett, "Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis," *Language* 91, no. 1 (2015): 194–244.

In our work, we first replicated the results of the New Zealand team using the same data set and methodologies. Like them, we found an ancestral Proto-Indo-European date of nearly 6000 BCE; this is consistent with the Anatolian model, not the steppe model. Have linguists simply been wrong about wheeled vehicle vocabulary and diversification in Indo-European?

On inspection, it turns out that the analyses of the New Zealand team yield a very surprising pattern. This is shown in Figure 3, which presents a part of the summary tree of our replication of their work. Shown here is the relationship between Old Irish and Modern Irish and Scots Gaelic, and between Latin and the Romance languages. In linguistics, it is conventional to assume that Latin is the ancestor of the Romance languages, and that Old Irish is the ancestor of Modern Irish and Scots Gaelic. But in the inferred tree in Figure 3, produced by phylogenetic methods, Old Irish is actually an aunt or uncle of Modern Irish and Scots Gaelic, and Latin is an aunt or uncle of the Romance languages. The same effect—replacing putative ancestry relationships with avuncular relationships—is present throughout the New Zealand team's tree, with Modern Greek not descended from Ancient Greek, the modern Indo-Aryan languages not descended from Sanskrit, and so on.

Moreover, the avuncular relationships posited in these analyses are not even very close. The vertical lines in Figure 3 demarcate millennia, so the branches from Old Irish, up to its ancestor, and down to its putative sibling (the ancestor of Old Irish and Scots Gaelic) add up to nearly 2,000 years of linguistic evolution. Likewise there are over 1,000 years of linguistic change between Latin and the posited ancestor of Romance. In short, the analyses of the New Zealand team posit that the attested, commonly assumed ancestors of later Indo-European languages are in fact rather distant from the actual ancestors of the latter.

This distance can be estimated. Shown below, based on our replication of the New Zealand team's analysis, are estimated basic-vocabulary divergences between each attested ancient language and the otherwise undocumented sibling that the analysis requires as the ancestor of modern languages. For example, comparing Latin and its putative sibling, just over 10 percent of the basic vocabulary would differ.

Old Norse 1.4%	Classical Armenian 8.6%
Old English 3.0%	Latin 10.2%
Old High German 3.1%	Vedic Sanskrit 13.1%
Ancient Greek 7.1%	Old Irish 17.1%

This is a substantial divergence in basic vocabulary traits. To put these figures in context, there is a basic-vocabulary divergence of 11 percent between Portuguese and Spanish, or French and Italian. In other words, the difference posited between Latin and the putative Romance ancestor is like that between French and Italian, two mutually incomprehensible languages. Based on what we know about Latin, this is simply false. There is now a lot of evidence about colloquial Latin; the scholarly consensus is that sociolinguistic variation in that language was like variation in familiar modern languages. Latin varieties were not as different as Portuguese and Spanish, or French and Italian.[14]

We argued that the effect in Figure 3 arises throughout the inferred tree (in the New Zealand team's work and our replication) because of

---

14   At a time when less was known about colloquial Latin speech (e.g., from papyri and other ephemeral written sources), it was more common to talk about "Vulgar Latin" as a distinct variety or even language. But to quote James Clackson and Geoffrey Horrocks: "there was so much geographical and social mobility" in Latin of the first three centuries CE that "local differences in speech tended to become levelled, and long-term divisions in the language were kept to a minimum." *The Blackwell History of the Latin Language* (Hoboken, NJ: Wiley-Blackwell, 2011), 236. For further discussion, see Chang et al., "Ancestry-constrained Phylogenetic Analysis," 206–208.

innovations that occur independently in related languages but were not present in their common ancestor. This happens often when languages share the same ingredients (related words with similar uses). For example, the related Latin and Old Irish words for "man, male person" (Latin *vir*, Old Irish *fer*) were replaced in Romance languages and in Modern Irish and Scots Gaelic, respectively, by cognate words that originally meant "human being" (Latin *homō*, Old Irish *duine*). When this happens, the phylogenetic tools tend to assume that parallel innovations were in fact ancestral. So, for example, undocumented siblings of Latin and Old Irish are posited in which *homō* and *duine* had already come to mean "male person," even though these semantic changes occurred later. The same kind of change and inaccurate inference are found throughout the inferred tree.[15]

The whole family tree is like an elastic structure. If avuncular relationships are inaccurately introduced in every sub-part, the added time-depth seen in Figure 3 stretches the overall tree like an accordion. As a global effect, therefore, falsely imputing parallel innovations to the ancestor language effectiveness forces an earlier ancestral date.

To address this problem, we introduced ancestry relationships as hard constraints. Just as the attested dates of ancient languages are stipulated in a phylogenetic analysis, known ancestry relationships can be stipulated. We hard-coded known relationships, constraining the space of possible trees to include only those where Latin is the ancestor of Romance languages, Ancient Greek is the ancestor of Modern Greek, etc.[16] In effect, this is to use what is known (language histories over observed time) to make inferences about what is in dispute (language histories in prehistoric times). From documented relationships, we were able to infer overall rates of change, from which we could infer rates of change over the entire tree. This results in not overstretching the prehistoric branches.

---

15   A few similar examples discussed by Chang et al., "Ancestry-constrained Phylogenetic Analysis": the semantic shift from "foot" to "leg" is common across languages, and is independently observed in the history of Greek, Indo-Aryan languages, and Iranian languages; when the Proto-Indo-European word for "snake" fell out of use, it was independently replaced by derivatives of the verb "creep, slither" in Albanian, Latin, and Indo-Aryan languages; and when the Proto-Indo-European word for "kill" fell out of use, it was independently replaced by derived verbs meaning "cause to die" in Indo-Aryan languages, Iranian languages, Irish, and Romance languages. In technical terms, the trait models used in current phylogenetic work do not account for homoplasy due to lexical drift.

16   We did this with eight languages (Ancient Greek, Classical Armenian, Vedic Sanskrit, Latin, Old Irish, Old English, Old Norse, Old High German) and for different versions of the analysis used three subsets of the same data: a narrow set with 8,615 words, 143 meanings, and 52 languages (2,350 traits); a medium set with 13,388 words, 143 meanings, and 69 languages (3,279 traits); and a broad set with 20,802 words, 197 meanings, and 94 languages (5,694 traits).
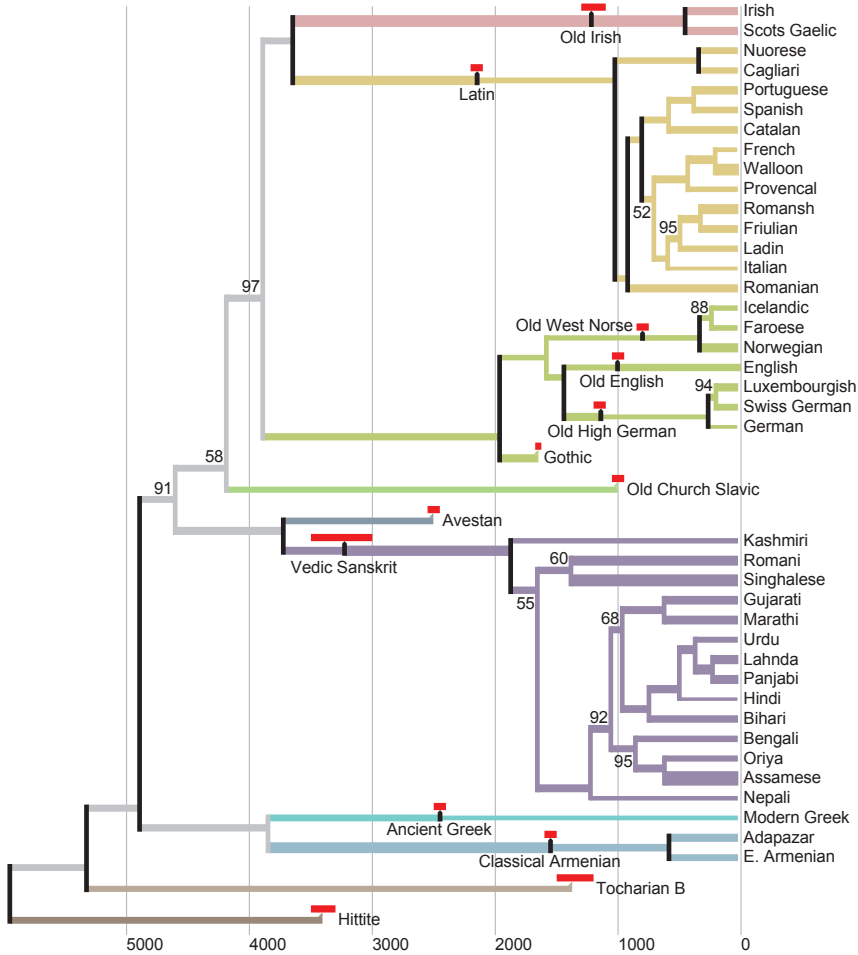
FIGURE 4. Analysis A1 summary tree (based on the narrow dataset). Reproduced by permission from Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett, "Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis," *Language* 91, no. 1 (2015): 194–244.

Our main result is seen in Figure 4, based on an analysis that used only modern languages that have attested ancestors. The New Zealand team's overstretching is absent, and the inferred date of Proto-Indo-European is about 4000 BCE. Figure 5 shows a result based on an analysis with modern languages from all Indo-European subfamilies, yielding an inferred date that is a few hundred years earlier, around 4500 BCE. The results in Figures 4–5 are consistent only with the steppe model, not the Anatolian model.
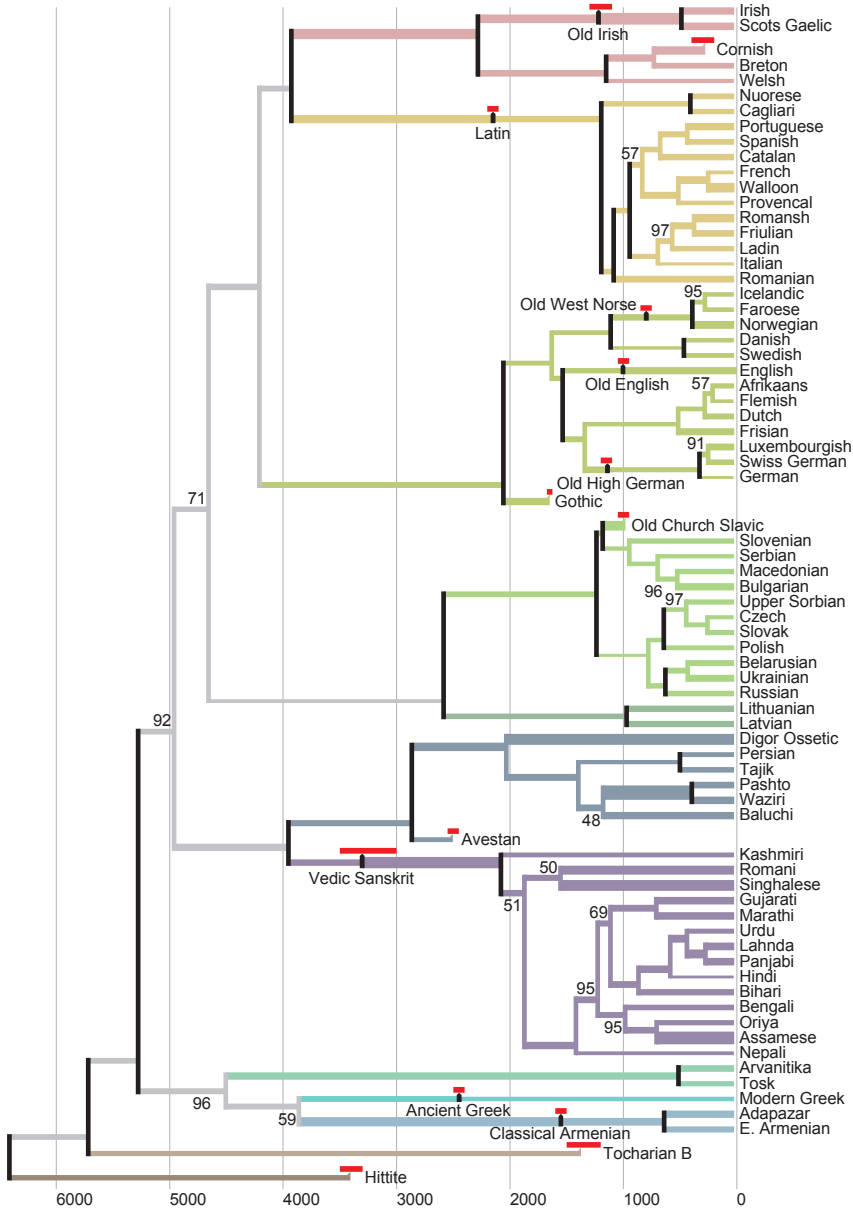
FIGURE 5. Analysis A2 summary tree (based on the medium dataset). Reproduced by permission from Will Chang, Chundra Cathcart, David Hall, and Andrew Garrett, "Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis," *Language* 91, no. 1 (2015): 194–244.

In summary, if what is known about relationships between ancestor and descendant languages is allowed to inform phylogenetic analysis, inferred rates of change become more realistic and the overall inferred chronology of Indo-European supports the prior consensus of linguistic work—work that was based on entirely different lines of argument.[17]

Conclusion

I end by highlighting two kinds of results from the work I've described. The first relates to Indo-European chronology and diversification. Where and when did this language family begin its long-term expansion? The steppe model is favored by a consensus now emerging from linguistics, archaeology, and genetics (see David Reich's contribution to this issue). Because the Anatolian model had the virtue of being tied to a well-understood mechanism of language spread—the diffusion of agriculture—the emerging consensus for Indo-European means that models of global language spread must explain more clearly how pastoralists expand their territories.

A second result is for linguistics. Comparative and historical linguistic research is often qualitative in nature, but we should not be afraid of computational tools in phylogenetics; we need not be more skeptical than is healthy in all scholarly discourse. For two large language families in very different settings, Austronesian and Indo-European, the new methods of computational phylogenetics have confirmed the broad lines of previous linguistic research. We should therefore also be engaged with applications to more difficult cases in the future.[18] As a limiting case where much is known about today's languages and there is not much documented time depth, this will include the Native languages of North America. If so, in time, it may be possible to use these new methods to pursue Jefferson's goal—to understand prehistoric relationships in North America and elsewhere on the globe.

---

17   The same results are also independently supported by analysis of molecular genetic data; see Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt et al., "Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe," *Nature* 522 (2015): 207–11. See also David Reich, *Who We Are and How We Got Here: Ancient DNA and the New Science of the Human Past* (New York: Pantheon, 2018).

18   See, for example, the recent application to another widely dispersed language family by Remco R. Bouckaert, Claire Bowern, and Quentin D. Atkinson, "The Origin and Expansion of Pama–Nyungan Languages across Australia," *Nature Ecology and Evolution* 2 (2018): 741–49.