ANNUAL REVIEWS

Annual Review of Linguistics

Cognacy Databases and Phylogenetic Research on Indo-European

Paul Heggarty

Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany; email: Paul.Heggarty@gmail.com

Annu. Rev. Linguist. 2021. 7:371–94

The *Annual Review of Linguistics* is online at linguistics.annualreviews.org

https://doi.org/10.1146/annurev-linguistics-011619-030507

Keywords

Indo-European, databases, cognacy, phylogenetics, Bayesian, chronology

Abstract

Repeatedly in recent years, phylogenetic analyses of linguistic data have reached the world's leading scientific journals, but in ways hugely controversial within linguistics itself. Phylogenetic analysis methods, taken from the biological sciences, have been applied to date and track how major language families dispersed through prehistory, with implications also for archaeology and genetics. As this approach is extended to ever more language families worldwide, this review offers methodological perspectives and cautionary tales from the most high-profile and hotly disputed case of all: Indo-European. This article surveys the checkered history of these phylogenetic methods and of the cognacy databases they have relied on for their linguistic input data. It clears up cross-disciplinary misconceptions about this new methodology, identifies major flaws in the current state of the art (hence its highly inconsistent results), diagnoses the causes, and outlines new solutions that might bring the field closer to living up to its potential.

1. INTRODUCTION

This article reviews an emerging, highly controversial field that has shaken and stirred historical linguistics over the last few decades. Highly complex mathematical methods for phylogenetic ("tree-drawing") analysis, developed principally for applications in the biological sciences, have been co-opted for the study of descent with modification in our languages. They are claimed to offer tools that not only can recover the structures of language family trees but, from those, can infer the time scales and early geographical dispersals of given families as they diversified through prehistory. The results have appeared in high-profile journals outside linguistics, and the authors of these studies claim to set language families into contexts that extend also to archaeology and to the fast-changing new field of ancient DNA. But in historical linguistics, these methods have faced enduring criticism and skepticism (e.g., Campbell 2013, section 17).

The impact and controversy have been felt nowhere more than on the oldest puzzle of them all in linguistics: the origins of the Indo-European language family, now the linguistic lineage of almost half the world's population. A great deal is already known about Indo-European, of course, and especially about some of its major branches. It thus also serves as a pathfinder and test case for these new techniques. Methodological lessons from this review should thus hold also for other language families and their prehistories, in other parts of the world. Similar studies elsewhere have raised much less *Sturm und Drang* than the Indo-European case, however, which has become something of a touchstone of a palpable tension between qualitative and phylogenetic camps (see Heggarty 2014, p. 567). The stand-off is almost ideological—a microcosm of the wider tension between supposedly incompatible qualitative versus quantitative ways of doing scholarship or science, even between the humanities and natural sciences.

In this context, an important aim of this review is to help defuse these tensions by clearing up some cross-disciplinary misconceptions and suspicions. This review is certainly no hard sell of either cognacy databases or the phylogenetic methods applied to them. On the contrary, it frankly exposes many concerns with both (as shown below in **Figures 1–3**). But it also rejects any knee-jerk fatalism that the concerns are qualitatively irreparable. Rather, it aspires to diagnose the causes and to outline potential methodological solutions to take the field forward.

In this article, I review the methodology for both main components in a process that might simplistically be paraphrased as proceeding from words to characters to numbers to trees to dates. The first stages are encapsulated in the databases of cognacy (shared word origins) that have provided the main source data (see the overview in **Table 1**). Mathematical, phylogenetic analyses require as input true data sets, in formats that are computationally tractable. So comparative language data need to be "encoded"—that is, reduced to expression as abstract states or numbers. Most phylogenetic methods ultimately require this encoding to be in simple binary, yes/no form: presence versus absence (usually expressed mathematically as 1 versus 0). The contrast could hardly seem starker with the intricate, detail-rich qualitative reference works most often associated with Indo-European linguistics, so the controversy (along with some misconceptions) extends to these databases too.

The later stages (numbers to trees to dates) are the task of the phylogenetic analyses themselves. The methodology for these analyses has been far from static over the last few decades; it has changed radically and diversified (see **Table 2**). The different techniques have their weaknesses and their strengths, as explored particularly in section 2 of the **Supplemental Text**. Certainly, notwithstanding much confusion and suspicion, methodology has come a very long way from the old, widely discredited techniques of lexicostatistics and glottochronology. The new breed of methods bears no real relationship to those at all (see sections 3.2 and 3.3 of the **Supplemental Text**). Crucial also is that the two components—cognacy databases and phylogenetic

Supplemental Material >

Table 1 Comparison of main cognate data sets of Indo-European languages used in major published quantitative and phylogenetic analyses

In Supplemental Text, see these sections	Publication	Data set name	Relationship to other databases: Minor changes: <10% of data changed/added; major changes: >10% of data changed/added	Languages: Total	Languages: Present-day	Languages: Past: usable as date calibrations	Data types, scale, and format: Phonological	Morphological	(Cognacy in) Lexical meanings	Cognate sets as binary characters (present/absent)	Lexemes	Download URL
3.1	Dyen et al. 1992	DKB	Ab initio (originally 95 languages, but only 84 used)	84	84	—			200	c. 2,400	c. 19,800	**https://thevore.com/comparative-indoeuropean-database-collected-by-isidore-dyen/**
3.3	Gray & Atkinson 2003	DKB +3	Minor expansion of DKB 1992	87	84	3			200	2,449	>19,800	
3.3	Ringe et al. 2002	CPHL	Ab initio	24	4	20	22	15	333	Multistate	?	**https://www.cs.rice.edu/~nakhleh/CPHL**
3.4	Nakhleh et al. 2005	CPHL "screened"	CPHL minus characters deemed to show parallel changes	24	4	20	22	13	259	Multistate	?	
3.5	Bouckaert et al. 2012	IELex (DKB+CPHL)	Major alteration of DKB 1992, part merger with CPHL	103	83	20			207	6,279	22,549	**http://hdl.handle.net/hdl:1839/00-0000-0000-0016-AD42-A**
3.5	Bouckaert et al. 2013	IELex (corrected)	IELex with technical correction to remove empty columns	103	83	20			207	5,996	22,549?	
3.6	Chang et al. 2015 (table 5)	IELex full	IELex corrected, plus minor revisions	103	83	20			207	5,992	22,353	Within supplementary material at **http://doi.org/10.1353/lan.2015.0005**
		IELex subset "broad"	IELex corrected, minus 9 languages and minus 20 meanings	94	77	17			197	5,694	20,802	
		IELex subset "medium"	IELex corrected, minus 21 languages and minus 64 meanings	82	68	14			143	3,279	13,388	
		IELex subset "narrow"	IELex corrected, minus 51 languages and minus 64 meanings	52	39	13			143 (92)	2,350	8,615	

Question marks indicate that the exact number of lexemes is not stated explicitly with the data set.

Supplemental Material >

Table 2 Comparison of major published quantitative and phylogenetic analyses of cognate data sets of Indo-European languages

←In Supplemental Text, see these sections	Publication	Data set used (see Table 1)	Data format and scale			Phylogenetic analysis type		Bayesian approach to:		Bayesian priors		Constraints/ calibrations				Results: inferred Indo-European chronology	
			Distance matrix or characters	Characters: multistate or binarized	Number of such characters	"Perfect" or Bayesian	If Bayesian, which software used?	Phylochronology	Phylogeography	Main analysis: tree prior type	Main analysis: trait model	Clade constraints	Date calibrations (time constraints): Non-modern languages	Date calibrations (time constraints): Assumed clade splits	Ancestry constraints	Main analysis: root date (95% range estimate)	Main analysis: root median date estimate
	In Supplemental Text, see these sections →		3.2	4.1		3.2.3		3.3.4	3.5	3.7	3.7	3.7				3.5, 3.6	
3.1	Dyen et al. 1992	DKB	Distance matrix														
3.3	Ringe et al. 2002	CPHL	Characters	Multistate	322	"Perfect phylogeny"											
3.4	Nakhleh et al. 2005	CPHL screened	Characters	Multistate	294	"Perfect Phylogenetic networks"											
3.3	Gray & Atkinson 2003	DKB	Characters	Binarized	2,449	Bayesian phylogenetics	MrBayes v. 2.01	Yes		(flat)	Restriction site	5	1	13	—	9200–7100 BP	8700 BP
3.5	Bouckaert et al. 2012	IELex (DKB+CPHL)	Characters	Binarized	6,279	Bayesian phylogenetics	BEAST v. 1.7	Yes	Yes	Coalescent	Stochastic Dollo	— (16)	20	14	—	10410–7116 BP	c. 8400 BP (figure S1)
3.5	Bouckaert et al. 2013 (correction to Bouckaert et al. 2012)	IELex corrected	Characters	Binarized	5,996	Bayesian phylogenetics	BEAST v. 1.7	Yes	Yes	Coalescent	Covarion	— (16)	20	14	—	9351–5972 BP	7579 BP
3.6	Chang et al. 2015	IELex corrected	Characters	Binarized	5,992	Bayesian phylogenetics	BEAST v. 1.7	Yes		Coalescent	Restriction site	35	Up to 17	8 (—)	8	7580–5060 BP (A3 corrected)	6240 BP (A3 corrected)

Several publications test a range of different data subsets, models, priors, and constraints. This table shows only the main analysis for each publication, plus some alternative settings in brackets.

Supplemental Material >

analyses—need to be applied not independently but specifically with regard to each other. A key lesson to emerge from this review is how critical it is for cognacy databases to structure, analyze, and encode their language data in ways appropriate to how the phylogenetic methods will make use of them. Likewise, those phylogenetic methods must beware of assuming properties of the input data that might not in fact hold for *language*.

This review seeks to plot a middle road through this tense methodological debate. It starts with no illusions. The dangers of false analogies between disciplines are real. But there are also valid analogies on abstract, conceptual levels. Historical linguistics itself sees appealing logic in tree models and linguistic descent with modification. The actual practice of combining cognacy databases with Bayesian phylogenetic models, however, has left much to be desired and much in dispute. The route through this dispute first has to avoid certain misconceptions on both sides. When these are cleared up, much of the controversy dissolves, and it becomes easier to grasp not just the potential in theory but also any shortcomings in practice.

This review seeks a middle road on another level, too. Much of the tension particular to the Indo-European case boils down not so much to the methodological issues reviewed here, but to the individual *results* of published Bayesian phylogenetic analyses, depending on which of the preexisting hypotheses of Indo-European origins they support. So to follow this wider context, a brief sketch of that debate is required. The Steppe hypothesis sees all Indo-European as having first spread out of the Pontic–Caspian Steppe from ca. 6,000 years ago, mostly with horse-based pastoralism. This scenario was set out and most prominently advocated by archaeologists (e.g., Gimbutas 1970, Mallory 1989, Anthony 2007) and came to be supported also by a large and vociferous group in Indo-European linguistics. It is most often set against the rival hypothesis by Renfrew (1987) and Bellwood (2005), both also archaeologists, that Indo-European first spread with early farming out of Anatolia, some 3,000 years earlier than the Steppe hypothesis would have it. A third proposal sees the Steppe as a staging post for a secondary expansion of the Indo-European languages of most of Europe, but not Greek and Albanian, nor most main branches outside Europe, which are traced back to an original homeland south of the Caucasus. This hypothesis was set out by the linguists Gamkrelidze & Ivanov (1984, 1995), so there is in fact no consensus within linguistics on the answer to the Indo-European question, or even on which forms of linguistic evidence and methods are probative to this end (see Heggarty 2018b). Ancient DNA is now adding a revolutionary new external perspective, which supports some elements of some hypotheses, but there seems to be no straightforward fit with all elements of any single hypothesis (see Heggarty 2018a). Alternative scenarios seem possible that would combine components of different hypotheses.

This review does not take a position on the "answer" to the Indo-European question. To do so would undermine the task of providing a balanced review of database and phylogenetic methodology. And from that perspective it is also natural not to take a position, because there is no agreed, consensus phylogenetic answer—far from it. Phylogenetics has sought to contribute above all through "phylochronology," using a Bayesian inference approach (see Section 3) to estimate when (Proto-)Indo-European first began to diverge. But the major papers report contradictory results. Central date estimates have varied widely, from ca. 8700 BP (Gray & Atkinson 2003) to ca. 6000 BP (Chang et al. 2015), in line with either the farming hypothesis or the Steppe hypothesis, or falling in between the two (Bouckaert et al. 2013). Irrespective of what the "right" answer may be, it has to be a concern for methodology that on the same basic question the answers could come out so differently. They cannot all be right. At least some of the methodological approaches must have been somehow inappropriate—and on closer inspection (Section 5.1), all main published results seem to harbor methodological artifacts, which this review traces back to their source.

The need for a methodological stocktaking seems clear. Reviewing the story of the field helps explain how it came to its current (still imperfect) state of the art. More positively, it also identifies where and how to make progress—and not just on the Indo-European issue, but in general methodology for both phylogenetic methods and for the cognacy databases they rely on.

The space limitations on this review necessitate a tight focus on the current state of the art: the most recent state of the field from 2012 onward, and future prospects. Nonetheless, to set the scene, the next two sections provide overviews of the nature of cognacy databases (Section 2) and of the history of the field up to 2012 (Section 3). Those are necessarily brief outlines, however. For much fuller exploration and explanations, readers are directed to sections 2 and 3 of the **Supplemental Text**. Those longer sections expand on all of the issues touched upon in Sections 2 and 3 below and provide additional illustrative figures and more extensive references to earlier work, before 2012.

Supplemental Material >

2. COGNACY DATABASES: WHAT AND WHY?

Cognacy databases are already familiar enough to historical linguistics, so here I assume some basic awareness of them. They are known also as the object of much suspicion and criticism, with their value for historical linguistics disputed, not least as input data for phylogenetic analyses of Indo-European. Here, I identify and summarize the main vexed issues, outlining the misconceptions and clarifications that can offer at least some reassurance that cognacy data should at least in principle be broadly viable, even if database practice has not yet lived up to the methodological ideals. For a fuller justification of this position on each of the main issues of dispute, and for a reference overview of how such databases (should) work, readers may consult section 2 of the **Supplemental Text**.

The most basic concern and suspicion surrounds why *cognacy* databases have become the data type of choice for phylogenetic analyses, especially when qualitative historical linguistics favors and "trusts" data in phonology and morphology instead. This is one case, however, where much boils down to some common misconceptions, and clarifying them can help take the heat out of disputes over which data types are most valid for phylogeny or chronology. Most crucial is to distinguish the respective roles of lexemes versus cognate sets in these databases. Indeed, while suspicious critics prefer to dub these databases "lexical," that is in fact something of a misrepresentation; it is much more accurate to describe them as cognacy databases. For such a database is by no means just a set of lexical lists, but far more than that. It is not lexemes that go forward into phylogenetic analyses, but cognacy states. And to establish those cognacy states properly and correctly entails a very full role for phonological and morphological data in any case. It goes without saying that a cognacy database cannot aspire to accurate cognacy coding without the full input of qualitative historical linguistics. In the Indo-European case, a cognate state is defined by a common proto-form that those cognates all go back to, as established by qualitative scholarship in Indo-European studies, and as ought to be explicit in the data coding. This also helps assuage another major concern. It is clear that "lexical" data are highly exposed to borrowing—that is, horizontal transfer that confounds the true, vertically inherited phylogenetic signal. Taken strictly, however, cognacy by definition already excludes loanwords.

One must also be realistic about the limitations of *all* data types. In practice, many phonological and morphological data characters that can bear on the earliest branching structure of Indo-European turn out to be open to interpretation and dispute, and thus have never brought a consensus qualitative answer on that deep phylogeny anyway. Furthermore, it turns out that there are too few clear-cut characters in phonology and morphology for the most powerful phylogenetic approaches to work with effectively, so cognacy characters are needed to make up the

bulk of the data in any case (Ringe et al. 2002, SI section 2). Finally, the traditional phonological and morphological characters are of such a nature that they cannot serve as data for the chronological inference technique used in phylochronology. To that end, cognacy data based on "universal" meaning slots that are constantly valid though time actually offer some positive advantages (and without going anywhere near the patently flawed assumptions of glottochronology; see section 3.3.3 of the **Supplemental Text**).

Supplemental Material >

To sum up, then, popular criticisms of "lexical" data turn out to be less valid than they appear, at least for a properly constructed cognacy database. Many other considerations remain—for instance, regarding the output data formats from cognacy databases, either taking each meaning as a multistate character or each cognate set as a binary (present/absent) character (see **Supplemental Figures 1** and **2**). Furthermore, serious dangers lurk in little-noticed inconsistencies in cognacy databases. The main weaknesses of the cognacy databases widely used in phylogenetic analyses of Indo-European are reviewed in Section 5.1, along with the new methodological approach needed to solve them (Section 5.2). Again, though, the problems are less of principle than of implementation in practice in existing cognacy databases for Indo-European published to date. First signs from applying a new database methodology (Section 6) are that many of the artifacts in results based on past cognacy databases now disappear, and outcomes are much closer to phylogenies expected from traditional phonological and morphological data.

On the basis of these clarifications, this review proceeds for now on the assumption that cognacy databases might, at least in principle, be useful for phylogenetic research on Indo-European. For fuller exploration and explanation of the nature of cognacy databases and all of the issues touched upon in this section, readers are directed to section 2 of the **Supplemental Text**.

Supplemental Material >

3. COGNACY DATABASES AND PHYLOGENETIC METHODS: A VERY BRIEF HISTORY

The focus of this article is to review the state of the art in the latest and most prominent phylogenetic studies into Indo-European origins, in both main components: cognacy databases and phylogenetic analysis techniques. The two latest major papers—one by Bouckaert et al. (2012) in *Science*, the other by Chang et al. (2015) in *Language*—both in fact used essentially the same cognacy database, IELex, and the same general analytical approach, Bayesian phylogenetics. Other papers have since explored variant analyses but have also used the same database and the same Bayesian framework. It is this latest phase, focused on that single database and single approach to phylogenetics, that I review from Section 4 onward.

Nonetheless, this latest research by no means came out of nowhere. It represents just the current phase that has emerged out of a much more diverse tradition of quantitative and phylogenetic approaches to Indo-European, over a long process of methodological debate and development. An extensive history of the field can be found in section 3 of the **Supplemental Text**. That section also provides additional explanation of the Bayesian phylogenetic approach that dominates the current phase, by explicit comparison and contrast with the (very) different methods that preceded it.

Supplemental Material >

Even for readers generally familiar with Bayesian inference and how it contrasts with the better-known frequentist statistics, it is worth clarifying what a "Bayesian" approach means for the purposes of this review—that is, as applied specifically to the phylogenetics of a language family, and to dating its time-depth. Central to the Bayesian approach is to conceive of and express things in terms of probabilities, and indeed this vision can be applied in different ways and on different levels. First, on the very broad and basic question of Indo-European origins, and assuming that the family had a single true history, probabilistic methods can evaluate the relative plausibility of different hypotheses on what that single history was—for instance, the respective probabilities

(given the language data and phylogenetic model used) of the Steppe hypothesis, farming hypothesis, or any other hypothesis. Second, a reality need be not simplex, but can be a combination of different components, and probabilistic methods can evaluate the most plausible relative strengths of those different contributions. In a language family, different data characters need not all show the same divergence histories. The overall picture can thus be a complex combination of different subparts, tantamount even to different, cross-cutting tree histories, and probabilities can represent the respective contributions of each.

Key characteristics that define the Bayesian approach and distinguish it from frequentist statistics are the explicit concepts and roles of *prior* and *posterior*, and the expression of both of these in terms of probabilities [respectively, *P(A)* and *P(B)* within the formulation of Bayes's Theorem]. In language phylogenetics the prior involves a set of explicit prior starting assumptions, including, for instance, a formal model of how (language) lineages diversify. The actual language data (cognate relationships) are then analyzed against this prior, to end up in a posterior result: a distribution of the respective probabilities of many different tree configurations and with that, the probabilities also of their corresponding time-depth estimates. Section 4.3 returns to this in more depth, and section 3.3.3 of the **Supplemental Text** explores this more fully still. For further, more general background on Bayesian inference, readers are referred to Yang (2006, section 5) and Beaumont (2010); for more on Bayesian phylogenetics in particular, readers are referred to Yang (2006, section 5.6).

Supplemental Material >

To put this latest, Bayesian phase into the wider methodological context that led to it, **Table 1** summarizes the main milestones in cognacy databases, while **Table 2** shows the main steps in quantitative and phylogenetic methods over the history of the field. Developments in methodology can be broken down grossly into four main phases (before the current fifth phase). Modern quantitative approaches to Indo-European origins began during the (1960s) heyday of lexicostatistics. That method produced simple counts of overlap/difference in cognacy between pairs of languages, and these gross measurements ("distances") formed the input to the early types of computational tree-drawing methods used in this first phase. But despite much early enthusiasm, serious objections were raised from the start, and the second phase was one in which lexicostatistics progressively became largely discredited. Meanwhile, methodological developments in phylogenetic analysis (and ever-increasing computational power) were escaping from the weaknesses of blunt, overall distance measures, as in lexicostatistics, and switching to alternative, more powerful approaches that retained all the detail in every data character. This new departure came to linguistics around the turn of the millennium, and soon the first major applications to Indo-European had appeared: a study by Ringe et al. (2002) and a hugely controversial one by Gray & Atkinson (2003). These character-based approaches in this third phase came in very different types, however, with a particular contrast between the search for a (mathematically) "perfect" phylogeny and a less all-or-nothing, probabilistic, Bayesian approach. Indeed, the following, fourth phase can be seen as one of methodological comparisons between these two competing approaches (and variants of each), exploring their respective strengths, weaknesses, and general validity. That exploration then led into the current (fifth) phase, in which—at least to judge from publications since 2012—the Bayesian approach seems to be winning out as the phylogenetic method of choice. As explored in Section 4, however, the broad Bayesian label still subsumes much scope for alternative models and parameters, and indeed starkly different results on the time-depth of the Indo-European family. For full details on each of the preceding four phases, see the corresponding sections 3.1–3.4 in the **Supplemental Text**.

Supplemental Material >

Progress in cognacy databases was much less radical. A series of very diverse phylogenetic analyses, over many years, continued to use essentially the same database originally collated and cognate-coded in the 1960s by Isidore Dyen (although best known from the much later publication

of Dyen et al. 1992). This formed the basis of the data on all modern languages even in the IELex database still used in the most recent major papers. The main alternative was the CPHL database of (mostly) ancient Indo-European languages, by Ringe et al. (2002). Alongside its phonological and morphological data, it too had a significant set of cognacy data, which would then be combined with Dyen's data for modern languages to create the "modern plus ancient" IELex. In other words, as quantitative and phylogenetic analysis methods changed fundamentally over the decades, the methodology for the cognacy databases that fed data into them did not keep up. Data collected and encoded for the old methods turn out not to be necessarily well adapted for what the latest methods do with them, and this mismatch helped create one of the major weaknesses that undermines even the latest results, as revealed in Section 5.1.

With this brief summary of the history of the field in mind, we are now forearmed to survey the current state of the art. For much fuller exploration and explanation of the entire history of the field and all of the issues touched upon in this section, readers are directed to section 3 of the **Supplemental Text**.

Supplemental Material >

4. WHERE WE ARE: LATEST, CONTRADICTORY FINDINGS

4.1. From Phylochronology to Phylogeography: Support for the Anatolian Hypothesis?

After the furor surrounding Gray & Atkinson's (2003) phylochronology of Indo-European, there ensued a decade or so of methodological exploration. Many objections (and some misunderstandings) were raised, and attempts made to respond and to test or even combine variant approaches. This stand-off phase burst back into open controversy, however, with the next major methodological innovation, when Bouckaert et al. (2012) introduced a major update and addition to the phylochronology of Gray & Atkinson (2003). IELex provided a data set significantly enlarged with Ringe's ancient languages, and thereby also many more time calibration points. The Bayesian analysis software used was likewise new: not MrBayes [as used by Gray & Atkinson (2003)], but the BEAST package [Bayesian Evolutionary Analysis by Sampling Trees (Drummond et al. 2012)]. And now the flexibility of Bayesian inference was extended to a new dimension: phylogeography. The phylo(geny) part remained as before: a probability distribution of tree structures through which the patterns in the cognate data most likely arose, according to a model of descent with modification (see section 3.3.2 of the **Supplemental Text**). The dates of all languages, modern and ancient, were again used to provide end points in time from which to infer chronology (see section 3.3.3 of the **Supplemental Text**). But now, the languages' locations were taken to provide end points in geographical space, too. The phylogeny could thus stand as a guide by which to estimate these languages' divergence not just through time but also through space (i.e., the dispersal process by which they came into their geographical end points). Or so, at least, was the principle. Historical linguists have themselves long assessed hypotheses for Indo-European dispersal in terms of compatibility with an assumed structure of the family tree—but not formally and computationally. Turning that principle into an actual component of a Bayesian phylogenetic analysis requires many formal assumptions, within an additional spatial diffusion model of how geographical dispersal proceeds. Bouckaert et al. (2012) used a "relaxed random walk"—akin to Brownian motion but relaxed in allowing wide variation in rates of movement. They also tested variants that adjusted probabilities for movement either by land or over large bodies of water. All variants supported the same result: an Indo-European homeland in Central Anatolia, in line with the farming hypothesis. In "Bayes factors," this result appeared overwhelmingly supported over the Steppe hypothesis.

Supplemental Material >

Implementing spatial diffusion into the Bayesian phylogenetic framework appears an impressive methodological achievement, of undoubted mathematical complexity (see Bouckaert et al. 2012, SI section 4). Nonetheless, the spatial diffusion models still struck commentators as highly idealized as well as potentially biased toward the gradual "demic diffusion" of the farming hypothesis, and against long-range migrations of highly mobile, horse-borne Steppe pastoralists. Also, the geographical coverage of IELex lacked any of the historical Iranic languages of Central Asia and the Steppe (the IELex language sample simply continued that of its source databases, which had been devised for very different, non-geographical purposes). As for the earliest attested Indo-European languages—those closest to the root—their locations reflect also the origins and spread of writing, not just of languages themselves. Moreover, technical errors in the input file necessitated a correction, and while the phylogeography result remained in the corrected version of Bouckaert et al.'s (2013) article, the root date estimate moved many centuries more recent, into a range centered on 7579 BP—all but ambivalent between the farming and Steppe hypotheses (for more detailed assessment, see Heggarty 2014).

In Indo-European circles, much reaction was caustic once more. Pereltsvaig & Lewis (2015) extensively criticized the entire process—from cognacy databases to phylochronology and phylogeography—but also revealed how deep the cross-disciplinary misconceptions still run. Their book's back cover already jumps to mischaracterizations. It describes "phylogenetic and phylogeographical analysis" as "treating cognates like genes and conceptualizing the spread of languages in terms of the diffusion of viruses." This seems a woeful misunderstanding of what is a basic abstraction of historical linguistics too. Indo-European languages descended and diverged out of their common ancestor, by changing through time as they also spread across geographical space. This abstracts away from any particular *modes* of descent or dispersal. Other things descend with modification through time and space, too, including species and virus strains, so phylogeography has been applied to research them too—but again, abstracting away from case-specific details of modes of transmission, descent or dispersal. Nothing in the study by Bouckaert et al. (2012) was conceptualized in terms of viruses.

Not all the objections were actually well founded, then. And stepping back to a more general level of methodology, at least, there was an echo of the occasional positive reactions from a few historical linguists to Gray & Atkinson's (2003) phylochronology a decade earlier. As April McMahon had put it then (cited in Whitfield 2003), "This kind of study is exactly what linguistics needs," and to some this new phylogeography again appeared intriguing, as an illustration of methodological potential on an entirely new level. Many questions remained over its particular implementation, however, leaving its results with no real acceptance among linguists, and there has been no other major attempt at Indo-European phylogeography since. Even those positively disposed in principle tend to see the value of Bouckaert et al. (2012, 2013) as a proof of concept, and far from a slam-dunk conclusion to the Indo-European debate.

4.2. Riposte: Back to a Steppe Chronology

Nonetheless, by now other disciplines had seen a drip-drip of results, which, even if hugely controversial with Indo-Europeanist linguists, seemed repeatedly to undermine the Steppe hypothesis. The onus seemed to be switching to its proponents to show some counterevidence that the new methods did not automatically favor the rival farming hypothesis (Heggarty 2014). The response was not long in coming. Chang et al. (2015, p. 199), "using the same model and data set as Bouckaert and colleagues (2012, 2013), but with incremental changes to both," reported a set of Bayesian phylochronological results that were indeed consistent with the short chronology of the Steppe hypothesis instead. The article by Chang et al. (2015) was even fêted as paper of the year in

Language, although that could also be read as a sign of the times: that Bayesian phylogenetics was finally gaining methodological acceptance in mainstream linguistics. Had its first proponents perhaps lost the Indo-European battle, but nonetheless won the methodological war for the Bayesian approach? (See section 3.3.3 of the **Supplemental Text**.)

Supplemental Material >

On the other hand, the welcome for Chang et al. (2015) in Indo-European linguistics may owe more to their (Steppe) result than to the methodology itself. Their paper set out some welcome technical developments of Bayesian methods for language diversification. But only one step proved crucial to swinging the result into the Steppe hypothesis time frame: enforcing a set of eight "ancestry constraints," that is, forcing, for example, Vedic to be directly ancestral to modern spoken Indic languages. While at first glance the idea may seem reasonable, on closer inspection and in stricter methodological terms, the most crucial constraints actually go against much linguistic orthodoxy. High-status ancient written languages are almost by definition *not* the direct sources of modern *spoken* languages (particularly not in lexical register). As Clackson (2016, p. 12) puts it of the Romance languages, for example, "The origins of the Romance languages lie in the (irrecoverable) spoken language. . .[and] there will always be a mismatch between the Latin sources and the parent of the Romance languages." Chang et al. (2015) nonetheless also forced Classical Latin to be directly ancestral to all Romance. Similarly, specialists are clear that West Saxon (the "Old English" of IELex) was not the dialect directly ancestral to modern English (e.g., Finegan 2009, p. 65) and that Vedic is not itself the direct source of modern Indic languages (Lazzeroni 1998, p. 102). The works that Chang et al. (2015, p. 206) cite as if in support do not in fact justify enforcing *direct* ancestry, but more the contrary. What Masica (1991, p. 51) actually states about Vedic, for example, is that it was an already distinct "far-western dialect," not compatible with seeing it as the direct ancestor of all modern Indic languages.

A number of other methodological decisions seem dubious, too. Most of Chang et al.'s (2015) range of analyses (all those coded l2) override the identification of loanwords in the IELex data set, treating loanwords instead simply as if they were true cognates with their loan source lexeme. Most of their analyses also discard data, reducing the IELex database to a series of increasingly smaller subsets of fewer languages and fewer meanings (Chang et al. 2015, p. 199), progressively limiting the data from which phylogenies, rates of change, and chronologies can be estimated (Section 2). Above all, their primary innovation of ancestry constraints did not so much solve the problems in Bouckaert et al.'s (2012, 2013) trees, as simply push them elsewhere. After the Classical Latin and Vedic constraints, for instance, their results show no divergence at all in Romance for over a millennium, until a sudden, frenetic burst of splitting only from AD 1000 (see **Figure 1**). Indic, similarly, has no internal diversity for 1,500 years after Vedic.

These results seem contradicted by the known and plausible histories for these clades. The authors explain this away with appeals to vague concepts of ongoing linguistic unity and mutual intelligibility. This is not actually valid, however, for a method in which a lineage split represents something very specific and different, and which long predates the end of mutual intelligibility. In these trees, even a single difference in cognacy, in a single meaning in the IELex 207 set, entails separate lineages, already split. (Differences in occasional meanings in this set are typically found even between regional varieties of English, such as in words like *dull/blunt*, as used of a knife.) Results need to be interpreted strictly in the terms of the method they are based on, and in those terms, Chang et al.'s (2015) results effectively deny even the most minimal divergence in the Swadesh lists of all Romance language lineages, even as late as AD 1000. They underestimate, then, the (very beginnings of the) divergence of the Romance languages, by up to 50%. Divergence dates are too late by a similar proportion for Indic, too, and for smaller clades such as West Scandinavian, High German, and Goidelic. This leaves little confidence in the shallow date estimates for the divergence of Indo-European as a whole.

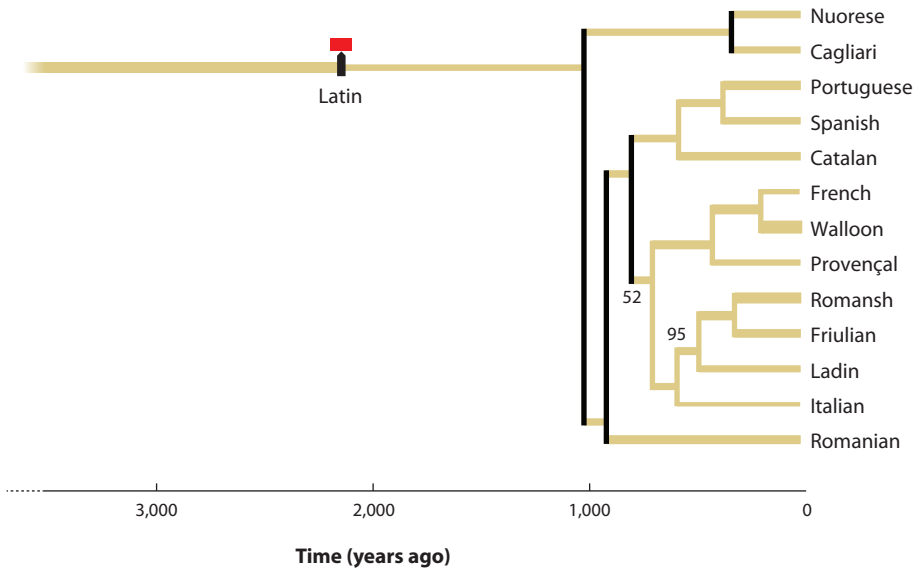

Figure 1

The implausibly late diversification of Romance in Chang et al. (2015, figure 1). The thickness of branch lines reflects relative rates of change along each branch segment (thicker line = faster change). Vertical black bars show clade constraints enforced. Numbers indicate percentage posterior probability values on nodes, where less than 98%. The red bar indicates the span of the time calibration constraint placed on Latin. Figure adapted with permission from Chang et al. (2015, figure 1).

Relief with a dating that finally supports the Steppe hypothesis should not have us turn a blind eye to such deep problems. For those seeking to assess this methodology as a potential tool for language history, here were two major papers on "the same model and data set" (Chang et al. 2015, p. 199) but which flatly contradicted each other's results on the same question, essentially because of a single changed methodological assumption. For methodological skeptics, this triggers some bemusement. All are Bayesians now, perhaps, but can one just pick one's assumption to get one's desired answer? With this as the state of the art, the task we now face is to get to the bottom of how such contradictions could have come about, to work out how to progress beyond them.

4.3. Exploring Differences Among Bayesian Phylogenetic Models

So far, I have sought to spare readers the intricate and technical details of differences in methodology within the broad field of Bayesian phylogenetics. But such differences lie at the heart of the latest phase in methodology, given that they seem responsible for the ongoing disparity in results. For while Chang et al. (2015, p. 199) used "the same model and data set" as Bouckaert et al. (2013), that by no means exhausts the set of methodological variables. It remains beyond the scope of this review to enter into any great detail here, although I hope to clarify at least the broadest issues, and direct readers to fuller treatments.

In fact, both Bouckaert et al. (2012, esp. SI) and Chang et al. (2015) themselves already discuss and test a range of methodological variants. Two more recent papers even switch the focus away from answering the Indo-European dating question itself, and make the explicit comparison of alternative methods their main purpose: one by Rama (2018) and, particularly recommended for clarity, one by Ritchie & Ho (2019). Both compare multiple phylogenetic methods on the same

Indo-European data sets, to explore which differences in methods can lead to which differences in results, particularly datings. They explore differences on two main levels.

First, the Bayesian approach needs a prior. This is "prior" in being set in advance of considering the actual data set or running any analysis on it. The prior is a form of broad initial framework, and the tree prior can be seen as embodying broad starting assumptions about how diversification into a tree structure proceeds. It defines only a general process and pattern of splitting (and extinction) of lineages, usually determined by one or other existing "evolutionary" model of how lineages diversify (where "evolution" effectively means nothing more than "descent with modification"; see section 3.3.2 of the **Supplemental Text**). It is *not* a presumed specific phylogeny for a family. Working out which specific tree structures are the best-supported for a given family is a separate analytical task (see below on the trait model, and also on clade constraints). The tree prior is used, however, to define a prior distribution of probabilities of tree structures—a form of broad, default, neutral assumption against which to test whether the real data and phylogenetic model actually support a more specific, narrower answer.

Supplemental Material >

There are alternative types of tree prior, and Ritchie & Ho (2019) provide a clear explanation of the difference between the two types most commonly used: "coalescent" and "birth-death" (and variants of each). Their tests find the birth-death type generally preferable for language families, and specifically for Indo-European—although both Bouckaert et al. (2012, 2013) and Chang et al. (2015) actually used a coalescent-type prior for their main analyses. The birth-death model does, however, require an additional sampling parameter, to specify the scale of coverage of language taxa in the data set used, as a proportion of all taxa in the family—which itself can be problematic, given the dialect/language issue. (Note also that "birth" and "death" rates here refer to lineages splitting or going extinct, not to changes arising in data characters, which are part of the trait model instead; see below.)

The tree prior also sets a very broad date range deemed plausible and worth testing at all. In the Indo-European case, for example, the prior could already exclude dates equal to or younger than the first attestations of distinct Indo-European languages, and dates much older than envisaged in any serious hypothesis. This prior date range would normally be set to a distribution broad enough for all main hypotheses to fit into, and fairly neutral between those to be tested, without presuming or excluding any one of them.

Second, to get to the actual results, the phylogenetic analysis on the data set needs a model of how those data characters ("traits") change through time along a tree structure—in this case, how cognacy states switch between being present or absent in a given lineage. This is variously termed the trait model, cognate substitution model, cognate replacement model, cognate evolution model, and so on. It is this model, applied to the cognate data set, that assigns the likelihood score to any tree structure (see section 3.3.3 of the **Supplemental Text**). Starting out from a random tree structure, the usual Markov Chain Monte Carlo (MCMC) process is then guided by these scores. The initial tree structure is tweaked in some respect, and if the new, tweaked tree has a higher likelihood score, then it goes through to become the basis for the next tweaking iteration. This process becomes a progressive search, sampling through the vast multidimensional tree space of theoretically possible structures, for trees with ever higher likelihood scores (for more detail, see Greenhill & Gray 2009, Greenhill et al. 2020). Relative to the prior distribution, this likelihood-guided search should progressively narrow down to a more highly supported subdistribution. Indeed, it may well shift the probability balance significantly away from that of the prior distribution. Once any such shifting stabilizes on a new probability distribution, the search can be ended, to leave this as the posterior distribution (i.e., the result).

Supplemental Material >

Of the various types of trait model available, Bouckaert et al. (2012) found the "Stochastic Dollo" to be the best-fitting model, and report results from that, whereas with their corrected

data set (Bouckaert et al. 2013), the "covarion" model was the best-fitting. The main analyses reported by Chang et al. (2015, p. 219), meanwhile, were based on a third type, a "restriction site" model. Nonetheless, both teams actually tested and compared all three models. Both also duly report that all three models returned broadly the same main chronological result—even though that differs significantly between the papers.

Other variables here lie in how chronology is estimated—flexibly, without any strict glottochronological constant. Rates of change can vary on two levels. First, a "relaxed clock model" accommodates rates of change that differ from one language lineage to the next ("among-branch" variation). Second, different (sets of) data characters—cognate sets or meanings—can also change at different rates. This "among-trait variation" can be accommodated by various "site models," as tested, for example, by Heggarty et al. (2021, SI sections 4.1 and 4.4).

Separately, and in addition to the general tree prior as described above, it is also possible to constrain the distribution of tree structures in much narrower ways, specific to a given language family. Three main types of constraint have been employed in Indo-European phylogenetics. A first type is a clade constraint, which means that only tree structures that include an individual, preassumed clade are considered. Assessing the balance of support for and against all clades is, of course, what the phylogenetic model and data set are intended to do in the first place, though. All approaches typically return the uncontroversial main 10 clades in Indo-European, along with most uncontroversial narrower ones, without needing any clade constraints. Indeed, forcing them can in principle risk certain circular presumptions in testing competing hypotheses. Nonetheless, Chang et al. (2015, table 9) constrain a wide set of 35 clades. These include "nuclear" and "inner" Indo-European, which in effect force an early branching order of Anatolian, then Tocharian, and are not entirely uncontroversial. A second type of constraint is effectively essential for chronological estimation: time constraints that are likewise far more specific than the very broad, neutral range of the tree prior. These constrain the date ranges (best also expressed as probability distributions) of individual non-modern languages or, much more questionably, assumed dates of particular lineage splits in the tree. The more ancient languages there are in the data set, the more of these date calibration points are available to contribute to the model's estimates of the distribution of rates of change, and thus of chronologies. Bouckaert et al. (2012) and Chang et al. (2015) both use date constraints for all their non-modern languages, as well as a number of assumed lineage splits, although they do not agree on the exact specifications. Finally, most crucial for their result is that Chang et al. (2015, table 2) also apply a set of eight constraints of a third type: ancestry constraints (see Section 4.2).

What lessons emerge, then, from the extensive, explicit comparisons of competing tree priors and trait evolution models? Some findings are partly reassuring: Ritchie & Ho (2019) report that varying some aspects of the prior, for example, does not affect dates very significantly. Other findings, however, raise more concerns than they dispel. When Ritchie & Ho (2019) tested different trait models on cognacy data sets for three other language families too, they found that the "best-performing" model varied by family. Also, this validation is based on measures of goodness of fit between competing models, particularly marginal log likelihoods. These are principled mathematical approaches (for discussion and examples, see Rama 2018, pp. 195–96; Bouckaert et al. 2013, SI table S3). It is not clear, however, that they necessarily correlate with how linguists might evaluate the plausibility or accuracy of the phylogenetic results. And for Indo-European, the same measures rated different models as best-performing on very similar data sets. Ritchie & Ho (2019) also report that these different trait models do lead to significant differences in date estimates.

These concerns are echoed on a much broader level in that the methodological variants have generally been developed primarily for applications in the biological sciences (i.e., outside

linguistics). This helps guarantee a steady flow of ongoing methodological development, of new and more flexible approaches. But variant approaches coexist within the biological sciences in any case—for instance, birth-death type tree priors generally for the species level, coalescent type priors for the population level. It thus remains open which (if any) of these might be most highly realistic and appropriate for how languages actually diverge, depending also on the different real-world contexts and processes affecting their speaker populations through time (Heggarty 2015, pp. 600–4). Ritchie & Ho (2019) repeatedly and rightly caution on just this point. Assessing this question calls for detailed and difficult cross-disciplinary thinking-through, and empirical exploration. There remains some ideological tension between the disciplines, too. Modeling is founded on known idealizations ("lies that can lead us to the truth"), and the Bayesian analyses are already hugely complex and computationally intensive, so there is a trade-off to be had. A model can soon become overparameterized, as the computational perspective would have it. That is, adding more parameters can heavily undermine computational efficiency, and leave so much scope for uncertainties as to compromise the model's ability to return clear-cut results at all. Many linguists, on the other hand, distrust the models as oversimplified, and would aspire to add further complex parameters to tailor them more closely to language, in order to have confidence in their results in the first place.

In sum, the recent exploratory papers do not seem to offer big, convincing findings to close down the large methodological wriggle room that has allowed different papers to produce different dating results. Among skeptics, methodological bemusement continues, then. There is also another sense in which the papers in this latest phase remain limited as contributions to the field. They all take existing data sets and existing phylogenetic approaches, and recombine them in multiple ways. Considerable expertise is needed to run these analyses, certainly, but assuming that, then multiple analysis runs still come relatively cheap compared to the demands of creating a major database of many thousands of rigorously researched cognate judgments. Ever since Bouckaert et al. (2012) introduced it, and aside from only minimal amendments by Chang et al. (2015), all major papers have essentially just recycled (or shrunk) the same Indo-European data set: IELex, itself still majoritarily the work of Isidore Dyen. Some studies even seem to rather naively assume that any data set (or subset) goes, as if there could be no valid methodological (linguistic) criteria on which one data set might be more valid and appropriate than any other, as input to these analyses. The focus on phylogenetic methodology seems to have been largely blind to the possibility that a major source of the inconsistencies and problems in results, in all analyses, might in fact lie in a common, flawed data set.

As the next section will now reveal, bad data turn out to have been far more of an issue than has been realized, and part of what has prevented these direct model comparisons from bringing us any real step forward in working out the sources of conflicting results. Indeed, if changing just one assumption could cause such a change in dating results, from Bouckaert et al. (2013) to Chang et al. (2015), one interpretation is that the data set could be oversensitive to particular assumptions. To conclude this review, and in a much more optimistic vein, I aspire to make a start by proposing a radical fix to the data set issue. This seems long overdue as the next phase that the field now needs to move into.

5. METHODOLOGICAL ISSUES FOR COGNACY DATABASES—AS INPUT TO PHYLOGENETIC ANALYSES

We have seen in Section 4.2 that with the IELex data set, forcing particular languages to be direct ancestors to others resulted in those descendant languages diversifying far too late (e.g., Romance in **Figure 1**). But there is a flip side to this, visible most clearly in **Figure 2**. Why did ancestry constraints seem necessary in the first place?

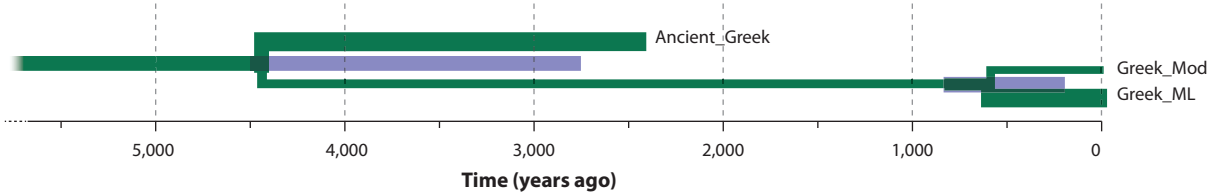

Figure 2

The implausibly early diversification of Greek lineages in Bouckaert et al. (2013, SI figure S1). The thickness of the green branch lines reflects relative rates of change along each branch segment (thicker line = faster change). The violet bar around each split between branches shows the 95% credible [HPD (highest posterior density)] interval of estimates for the date of that split. Abbreviations: Greek_Mod = Greek: Modern; Greek_ML = Greek: Modern Lesbos (Dyen et al. 1992, p. 100). Figure adapted with permission from Bouckaert et al. (2013).

Strictly, ancient written languages like Classical Latin were not themselves the exact variety and register ancestral to modern spoken languages like those of the Romance family. But in many cases they were at least close. From this type of data set and phylogenetic analysis, the expectation is that such near-ancestor languages should split from the lineage ancestral to the spoken languages, yes, but on only a short branch. (Differences in register or regiolect effectively end up expressed in terms of branch length—the only dimension on which language taxa can vary in these idealized models.) In the Greek branch, register contrasts and the *koine* phase entail that modern spoken Greek does not derive absolutely directly from written Ancient (Attic) Greek. But the differences would be few and slight, and they cannot remotely explain the scale of the gulf between their lineages in **Figure 2**.

Bouckaert et al.'s (2013, SI figure S1) "maximum clade credibility" (MCC) summary tree has the lineages to Modern and to Ancient Greek diverging from each other already by ca. 4400 BP, almost two millennia too early to be plausible. Even granting the wide "error bar," the bulk of the chronological distribution (violet bars in **Figure 2**) lies very far off the mark. In other branches, such as Indic and Romance, effects are less extreme, but branches to near ancestors likewise split too deeply to be historically plausible.

Results like those shown in **Figures 1** and **2** seem so awry that skeptics might seize on them as if evidence that these phylogenetic methods just do not work for languages. It is sure that something is deeply amiss in both cases. On closer inspection, though, the phylogenetic methods themselves are rather less at fault than it might appear. Under normal conditions, these methods do not produce results so off the mark as these. As we shall now see, in **Figures 1** and **2**, the methods actually did what they were devised to do, with the particular data (and constraints) input to them. Rather than giving up, or downplaying the weird results and attempting to explain them away, it is more productive to try to diagnose how they could have arisen.

Figures 1 and **2** are in fact two sides of the same coin. In both cases, the analysis was clearly producing excessive branch lengths. In Bouckaert et al.'s (2012) paper, this excess came *before* each near-ancestor language (shared out also on the parallel branch leading to its near-direct descendants). Chang et al. (2015) essentially reused the same data and very similar analyses, however, and thus did not address the source of the excess branch lengths. So when they simply added ancestry constraints, this did not remove the excess branch lengths but just forced them to flip *after* each near-ancestor language instead. (It also had the side effect of altering the distributions of rates of change over time, in a way that shifted the root date estimate shallower.)

Forcing direct ancestry was technically unnecessary anyway. As per the expectation explained above, these analysis methods are free to set (near-ancestor) languages on very short, near-zero branch lengths, which in effect equates to (very near) direct ancestry. But the data and model

came nowhere near to supporting that, and insisted on overlong branch lengths. This problem remained, its source undiagnosed and untreated.

5.1. Diagnosing the Data Problem

There has long been criticism of Dyen's original database (e.g., Embleton 1995, McMahon & McMahon 2005, Holm 2011) and of its IELex update (Pereltsvaig & Lewis 2015). Objections have focused on questionable choices of lexemes in a given language, and in particular on some small sets of undetected loanwords, wrongly entered as true cognates (an example is provided below for Danish). But individual erroneous data points of these types were still not at such a scale as could produce the artifacts in **Figures 1** and **2**. Effects of that magnitude go back to a far greater problem: a general flaw of database policy, part of a wider, urgent need for a whole new methodology for cognacy databases. This systemic flaw can best be appreciated from **Figure 3**.

A golden rule of a data set is that it should be consistent. **Figure 3**, however, reveals serious inconsistency within the data files input to the analyses of Bouckaert et al. (2013) and Chang et al. (2015). The outlier block near the value 1.8 on the righthand side of **Figure 3*b*** represents Ancient Greek. For its set of 207 reference meanings, the IELex file had 363 cognate sets present (binary value 1) in Ancient Greek. For Modern Greek, it had only 229. One expects a share of differences within the overlap of 229 anyway, but over and above that, this inconsistency in itself necessarily entails 134 extra differences between these languages: 1 in Ancient Greek versus 0 in

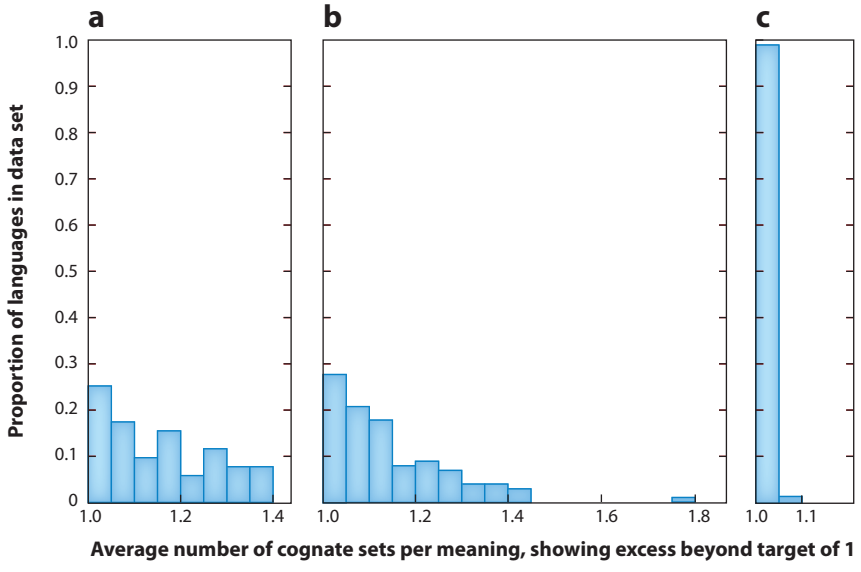

Figure 3

Data set (in)consistency in three Indo-European cognacy databases. (*a*) Data set d0-c0-g1 used in the main analysis A1 in Chang et al. (2015) and their figure 1. (*b*) IELex data set of Bouckaert et al. (2013), reproduced also as data set d2-c4-g3 in Chang et al. (2015). (*c*) New IE-CoR data set of Anderson et al. (**https://iecor.clld.org**, publications in preparation). The *x* axis is divided up into bins. The first bin includes languages with an average ≥1.00 and <1.05 cognate sets per meaning, the second bin includes languages with ≥1.05 and <1.10 cognate sets per meaning, and so on. The histogram bars show how many languages, as a proportion of the total covered in that data set, fall into each bin. (In an ancient language, data may not be known in some meanings, which therefore do not count toward average excess synonymy in that language.) For further details and discussion of this figure, see Heggarty et al. (2021, SI section 1).

Modern Greek. In the model, these differences correspond to changes: gains of cognate sets on the branch to Ancient Greek and/or losses on the branch to Modern Greek. The direct consequence of this data inconsistency is hugely to inflate the branch lengths between these languages with the additional 134 changes, and correspondingly to overestimate how long ago their lineages split from each other. (Without these excess cognate sets and branch lengths, the time-depth of this split in the Greek lineages is returned much more accurately; see Section 6 below.)

Such inconsistency is found across the IELex data files and Chang et al.'s (2015) various subsets of them. For example, their d0-g2 data set (Chang et al. 2015, SI file /sup/runs-post/a1-c2-d0-g2-l2-s1-t1-z3/ieo.xml) uses a slightly smaller set of 197 reference meanings, across which Catalan has 282 cognate sets present, Portuguese 267, but Latin only 202—to give just a few examples. In Indic, Vedic Sanskrit has 248, Punjabi 205, and Marathi 260. This scale of inconsistency is enough to cause the big distortions to branch lengths and divergence time estimates within Romance, Indic, and other branches.

How could a language come to have many more cognate sets than the number of reference meanings in the database? All languages are required to have a lexeme specified for each meaning (if attested). Each lexeme has its cognate set identified (a single one, except in rare cases of compound lexemes). So the baseline is that each language has one cognate set present (1) per meaning. Dyen, however, explicitly tolerated the entry of near synonyms, too, such as both *schlagen* and *treffen* for the meaning HIT in German (see Dyen et al. 1992, IE-DATA.txt section 3a). These near synonyms, and others from the CPHL data set (e.g., for Ancient Greek), were carried into IELex. Any synonym adds a cognate set beyond the baseline of one per meaning, and directly results in extra branch length (for detailed explanation, see Heggarty et al. 2021, SI). Compounding the problem, different contributors of lexical lists provided widely varying numbers of synonyms for different languages. There seems to have been no database policy applied to ensure consistency in this.

How could this inconsistency—and its consequences—have been missed? One possible reason is that it lay hidden in raw data files that are far from human-friendly and demand painstaking poring-over. (This suggests that future publications should provide more transparent versions of their input data files, to facilitate precise data-checking and peer review.) In fact, some efforts were made within IELex to thin down the number of synonyms, but still left the very high proportions shown in **Figure 3*a***. (The input file behind **Figure 3*b*** seems to have suffered also from an unnoticed bug in the script that created it, such that most "excluded" synonyms and loanwords were in fact left in.)

More generally, the oversight seems a legacy of the history of the field. While phylogenetic methodology changed radically, the methodology behind cognacy databases failed to keep up (Section 3). Dyen's database was devised for lexicostatistics. He tolerated synonyms, perhaps even encouraged them, in order to address one particular objection to that method. A language may have more than one candidate lexeme for a given meaning, as in Dyen's example of both *schlagen* and *treffen* for HIT in German. Choosing only one of them seemed to risk overstating the difference with certain languages (particularly closely related ones) that used a cognate of the other: Dyen pointed to Afrikaans *slaan* and Danish *traeffe*. (In fact, the latter is an example of a miscoding in Dyen and IELex: *traeffe* is a loan, not a cognate, of German *treffen*.) In such cases, Dyen allowed any one of the synonym lexemes to suffice to declare a cognacy match, a full similarity of 1 for that language pair in that meaning, to go forward into the total score across all meanings. So for HIT, Dyen scored German as cognate both to Afrikaans and to Danish, even though they scored as 0 similarity to each other (see Dyen et al. 1992, IE-DATA.txt section 3a).

Supplemental Material >

Later, when character-based Bayesian phylogenetic methods came on the scene (see section 3.3 of the **Supplemental Text**), their proponents seized upon Dyen's database as the main, readily

available one. But Dyen's generous inclusion of synonyms was within a distance-based method (see section 3.1 of the **Supplemental Text**) that could take an "any one" approach to them, with the net effect of limiting differences between closely related languages. The new, character-based approach, by contrast, took *all* synonyms, and each of them entailed a new, additional cognate set character present (1) in that language. The net effect was the opposite: This could not limit differences, but only add new ones (and branch lengths) vis-à-vis any languages coded with fewer synonyms. Another major change in how language data were handled was binarization (see **Supplemental Figure 2**), which in effect magnified the impact of synonyms. Other Bayesian phylogenetic methods that take multistate input data might potentially attenuate the impact of synonyms, although the inconsistency will remain: Some languages will have many more characters with multiple concurrent states than other languages.

Supplemental Material >

In short, the switch from lexicostatistics to (binary) character-state phylogenetic methods was a huge change in approach, but there was no corresponding change in cognacy databases. Dyen had made a particular design decision (on synonyms) for his lexicostatistical purpose. When co-opted for a totally new type of analysis method, Dyen's database still had its language data handled appropriately for his different, original purpose. Unforeseen by those who applied the new phylogenetic analyses, Dyen's synonyms led to the artifacts in **Figures 1** and **2**.

5.2. Curing the Data Problem: Toward a New Methodology for Cognacy Databases

In fact, the mismatches go beyond just the handling of synonyms. Cognacy databases face a long series of other potential challenges, several of which have indeed proved damaging in practice. So numerous and wide-ranging are they that the field needs far more than just a small-scale correction to IELex. Rather, what is called for is nothing short of a whole new methodology for cognacy databases, because the problem is a systemic one. Cognacy databases have structured, sampled and encoded language data in ways that failed fully to realize what the new, character-based phylogenetic methods would make of them. Or from the other perspective, those methods have been applied without fully realizing the consequences of the policies used in creating the language data sets they used. There still seem to be some crossed lines, then, between historical linguistics and the complex mathematics of evolutionary modeling. And while new phylogenetic methods powered ahead in high-profile journals, database methodology seemed typecast as the poorer, less glamorous relation. Paying too little regard to it, however, turned out to be the Achilles' heel that led to results such as those shown in **Figures 1** and **2**, and helped undercut confidence in the methods too.

This review can only sketch out here, on each issue, the basic methodological rethinking that can justify some confidence that the task is feasible: that alternative database structures and policies can be found to solve or at least significantly mitigate the challenges. Much more detailed, comprehensive treatments are forthcoming from Heggarty et al. (2021, SI) and Anderson et al. (**https://iecor.clld.org**, publications in preparation).

On the most critical issue of inconsistency in numbers of cognate sets per meaning and per language, the only viable target is as close to 1 as possible: that is, minimal tolerance for "synonyms." A database of core lexicon can only ever aspire to represent a sample—a small fraction—of any language's vocabulary in any case. The appropriate objective is not to extract, from any one reference meaning, as broad a coverage as possible of the full range and lexical wealth of a language in that semantic field (that is a dictionary's job). Rather, for a cognacy database, the imperative is *consistency* in the representative sample. This requires precise (re)definitions of all reference meanings: not as loose, broad semantic fields, but focused on a specific target sense.

Indeed, they should be explicitly defined to clarify which other related senses are not the target, and to avoid the ambiguities of the English lexemes used to label them. German *schlagen* and *treffen* are not true synonyms, and they appear so only if the reference meaning is left too loosely defined, as if by the polysemy of English *hit*. There is in fact a wide range of principled linguistic tests and criteria—register, neutrality, (non)markedness, etc.—that can very effectively reduce the scope for inconsistency and inaccuracy in lexeme determinations. In the vast majority of instances, such redefinitions and criteria can make possible a crosslinguistically consistent identification of a single lexeme, even in cases that may first appear the hardest to call, like English *small* versus *little*, or Spanish *pájaro* versus *ave* in the meaning BIRD.

More generally, there are also principled (and empirically supported) ways to target a more appropriate set of reference meanings in the first place. Swadesh's choices are not sacrosanct, and many criteria can be applied to optimize the set. It should avoid, for instance, "meanings" that are too dependent on the idiosyncrasies of particular lexemes in English, and grammatical functions that English happens to express by lexical means. The meaning set can also be optimized in ways that are language-family-specific, to steer clear of meanings that pose particular problems in Indo-European.

Supplemental Material >

The synonyms problem aside, the single commonest concern in the literature has instead been with the loanword problem. Although section 2 of the **Supplemental Text** clarifies the principle that cognacy by definition excludes loanwords, this issue seems never to have been handled satisfactorily in practice. Dyen adopted the right basic approach in theory, sequestering loanwords out of the cognate set to which their source form belonged. He was let down in practice, however, by missing just how many of his "cognates" were actually loans, even in English (see McMahon & McMahon 2005, p. 118). In IELex, meanwhile, many loanwords were just tagged as such, and kept within the cognate set of their source lexeme. Bouckaert et al. (2012) report that they excluded loans, but most seem to have slipped through into the input file nonetheless. Chang et al. (2015, p. 205), in most of their analyses (their l0 type), made the highly questionable choice to deliberately override the identification of loanwords and explicitly include them as if truly cognate with their source lexeme. In short, no previous treatment is advisable. Some more fundamental thinking is required regarding the various different processes and stages that all fall under the loose umbrella term "loanwords," and how each actually relates to cognacy and phylogeny.

Loans should certainly, by definition, not be included within the cognate set of their source (although they can usefully be cross-referenced to it, to allow for other analyses). Also, simply excluding loans from the input to a phylogenetic analysis creates problems of its own. It leads to missing data, in nonrandom patterns, and can actually discard valuable phylogenetic information. For once a loan has been integrated in the borrower language's lexicon, from that point on it can form the basis of a new, true cognate set, inherited into its own descendants. Loans of Latin *piscis* 'fish' into contemporary stages of the Albanian and Brythonic Celtic lineages have since continued, inherited, into modern Albanian and Brythonic varieties. The solution is to establish separate cognate sets defined by these individual *loan events* as their origin (e.g., separate sets for Albanian and Brythonic, notwithstanding their common Latin source). Other loanwords from a common source, but borrowed into related languages *after* they had already begun to diverge, must be treated as a separate type: "parallel loans."

Supplemental Material >

Another challenge is parallel semantic shift (see Heggarty 2018b, pp. 99–101). As noted in section 3.3.3 of the **Supplemental Text**, assessing whether an innovation is shared or parallel can run risks of subjectivity and circularity. Nonetheless, at least in a good proportion of cases, objective linguistic criteria can provide a clear, consensus determination that particular developments began only long after branches had already diverged. A database can implement specific structures

to handle such parallel shifts, to ensure that they are not input to phylogenetic analyses as if they were cases of true vertical descent.

Returning to the need for consistency, this applies not just in determining the lexemes in each language but also in determining cognacy status between languages. Cognacy status is not always clear-cut, but can be partial in various ways (see McMahon et al. 2005). In the meaning HEART, for example, most Romance languages share a common root, as in Italian *cuore* and French *cœur*, but Spanish *corazón* adds an extra suffix, *-azón*. Again, a consistent policy is possible, particularly if based on root cognacy. This can be made even more explicit if defined in terms specific to the proto-language of the family in question. In the Indo-European case, cognacy determination can call and rely upon compendious standard reference works, such as the *Lexikon der Indogermanischen Verben* (*LIV*^{2}, Rix et al. 2001) and *Nomina im indogermanischen Lexikon* (*NIL*, Wodtko et al. 2008). Preferably, all cognacy decisions should be fully justified and referenced by citation to the corresponding entries in those works and others in (qualitative) Indo-European linguistics. Indeed, even if optimized for applications in phylogenetic analysis, a cognacy database of Indo-European should meet the same exacting linguistic standards needed for it to serve as a new comparative resource for qualitative research too.

More generally, it stands to reason that all cognacy determinations at the Indo-European level should be the work of specialists in Indo-European linguistics. Likewise, other branch-specific cognate sets require expertise within each major branch of Indo-European. Lexeme determinations, meanwhile, call for specialists in individual languages, best placed also to identify loanwords. In short, data accuracy requires a large, diverse team of linguist contributors, for whom it is all the more important to have strict, explicit policies for consistency.

Within a consortium-based approach, having more specialists also makes it possible to cover more languages. Indeed, language coverage should also be principled, to be as comprehensive and balanced as possible on each of three main levels: phylogeny, geography, and time-depth. Such an approach should also avoid misdirecting resources by oversampling language varieties so closely related that they bring very little to the overall phylogenetic results. Conversely, for phylogenetic and chronological inference, and to provide known historical test cases for assessing the validity of results, ancient or historical stages of language lineages are especially valuable. They also enrich the set of date constraints (Section 4.3) to better calibrate analyses on the time dimension. Furthermore, covering such languages does away with any need for circular, subjective and speculative date constraints on lineage splits. The time-depths of these splits can only be guesstimated, either on hypothetical correlations claimed with archaeological datings, or relative to a historical written language presumed to be close to the proto-language stage. Such guesstimation is unnecessary if that historical written language is covered directly itself, with its own, reliable date calibration.

In sum, once appropriate methodological policies are in place for all of these issues, problems that seemed serious in principle can be so mitigated that in practice they affect a much smaller proportion of the full data set than feared.

6. WHERE NEXT? REMAINING TASKS AND FUTURE PROSPECTS

The last section above makes for an intimidating checklist. Nonetheless, a major new attempt has been made to implement all of this new methodology for cognacy databases in practice, to create a new cognacy database of Indo-European: IE-CoR, by Anderson et al. (**https://iecor.clld.org/**, publications in preparation). The experience of creating this database over the last few years seems reassuring that even if the task is arduous, it can be achieved. In preliminary results from its entirely new data set, many of the artifacts of previous studies have vanished, not least the excess branch lengths after Latin in **Figure 1** and before Ancient Greek in **Figure 2**, and without any need

for ancestry constraints. As these artifacts that clouded past results dissolve, one should be able to see through more clearly to aspects of the results that can still seem unexpected. This in turn helps diagnose their causes, to identify which next methodological steps may be able to address those too. And as database methodology catches back up, these lessons may extend also to the phylogenetic models employed, to suggest how they too might be tailored more closely to how descent with modification proceeds specifically in language.

This is not to say that no concerns or obstacles remain. Section 5.2 above outlines how the analysis of loanwords can and must be much improved, as in IE-CoR. But even when handled entirely appropriately, loanwords still mask phylogenetic signal, and not randomly but in patterns that still potentially disrupt results. With poorly attested ancient languages, meanwhile, large amounts of missing data remain a concern, as does recognition bias: Words are identifiably "attested" especially if they remain "good Indo-European cognates." Further explorations are needed to fully investigate these potential confounds.

As for parallel semantic shifts, Bayesian phylogenetic models are themselves intended to help tease apart which patterns in character data more likely arose through shared innovation or parallel changes. Most methods still seem configured rather too narrowly in favor of parsimony, however, to judge from results for languages in close contact, especially in dialect continua. If phylogenetic analyses could allow particular language taxa to be identified as being in such common contexts with each other, then the model could correspondingly adjust the respective likelihoods of parallel versus shared innovations. Similarly, certain ancient written languages could be specified as "fossilizing," so that changes would be taken as more likely to be happening instead on (nonfossilizing) sister lineages. This might attenuate the remnant of the tendency toward excessive branch lengths leading up to near-ancestor written languages. Also on the chronological side, empirical distributions of rates of change might usefully contribute to the rating of tree likelihoods, downrating those tree structures that necessitate unrealistically high or low rates of change.

What all of these suggestions have in common is that they propose adding what linguists would consider material information to understanding language histories, but which is not currently taken into account by phylogenetic methods (although see Section 4.3 for a reason why). On the other hand, linguists ought not to demand unrealistic magic from phylogenetic methods. Yes, at the level of fine-scale relationships between closely related languages, especially within dialect continua, results can still seem poor in some cases. But the analysis may be coming up against the limits to the resolution not of the methods, but of the data set. Cognacy databases of Indo-European have deliberately targeted those reference meanings where cognacy relationships tend to survive most, to provide the key signal on the deepest relationships between the very disparate branches of an old and broad language family. The inevitable flip side is much less resolution on fine-scale divergence, between language varieties that vary in cognacy in only a handful of those same, slowly changing meanings. A simple (albeit very labor-intensive) solution is to extend the set of reference meanings to include many more in which cognacy states typically change much faster. Alternatively, these could form a dedicated separate database for finer-resolution analyses, while the existing "high-stability" databases remain used for deep Indo-European research, without unrealistic expectations as to their resolution for fine-scale and dialectal divergence.

7. ENVOI

The history of the field surveyed here has been a long and tortuous one, with many road bumps along the way. Its results still command no great acceptance in Indo-European prehistory. This review hopes at least to have clarified many misunderstandings, so that some easing of the qualitative/quantitative stand-off may now be at hand. Realistically, a consensus result from Bayesian phylogenetic analyses, applied to cognacy databases of Indo-European languages, still

seems a long way off. But with past flaws now diagnosed, and potential methodological cures identified, the next steps are now clearer. A radical new database methodology, once implemented, should allow a new cognacy database to command respect as both a qualitative and a quantitative resource for Indo-European linguistics. That should in turn support further explorations to clarify which approaches to phylogenetic analysis are most appropriately configured for language. This new combination of data and methods might even aspire to earn rather more confidence in whatever revised perspectives it may offer on Indo-European origins.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author thanks Mark Liberman, Cormac Anderson, and Adam Powell for their reviews and many valuable comments and suggestions, and the Annual Reviews team for their attentive work on the figures and copyediting.

LITERATURE CITED

Anthony DW. 2007. *The Horse, the Wheel, and Language: How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World*. Princeton, NJ: Princeton Univ. Press

Beaumont MA. 2010. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41:379–406

Bellwood P. 2005. *First Farmers: The Origins of Agricultural Societies*. Oxford, UK: Blackwell

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, et al. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337(6097):957–60. **http://doi.org/10.1126/science.1219669**

Bouckaert R, Lemey P, Dunn M, Greenhill SJ, Alekseyenko AV, et al. 2013. Correction to: Mapping the origins and expansion of the Indo-European language family. *Science* 342(6165):1446. **http://doi.org/10.1126/science.342.6165.1446-a**

Campbell L. 2013. *Historical Linguistics: An Introduction*. Cambridge, MA: MIT Press

Chang W, Cathcart C, Hall D, Garrett A. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1):194–244. **http://doi.org/10.1353/lan.2015.0005**

Clackson J. 2016. Latin as a source for the Romance languages. In *The Oxford Guide to the Romance Languages*, ed. A Ledgeway, M Maiden, pp. 3–13. Oxford, UK: Oxford Univ. Press

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* 29(8):1969–73. **http://doi.org/10.1093/molbev/mss075**

Dyen I, Kruskal JB, Black P. 1992. An Indoeuropean classification: a lexicostatistical experiment. *Trans. Am. Philos. Soc.* 82(5):1–132. **http://doi.org/10.2307/1006517** (data set available at **https://thevore.com/comparative-indoeuropean-database-collected-by-isidore-dyen/**)

Embleton SM. 1995. Review of *An Indoeuropean Classification: A Lexicostatistical Experiment*, by Isidore Dyen, Joseph B. Kruskal, and Paul Black. *Diachronica* 12(2):263–68. **http://doi.org/10.1075/dia.12.2.10emb**

Finegan E. 2009. English. In *The World's Major Languages*, ed. B Comrie, pp. 59–85. London: Routledge

Gamkrelidze TV, Ivanov VV. 1984. *Indoevropjskij jazyk i indoevropejcy: rekonstrukcija i istoriko-tipologieskij analiz prajazyka i protokultury [The Indo-European Language and the Indo-Europeans: A Reconstruction and Historical-Typological Analysis of a Proto-Language and a Proto-Culture]*. Tblisi, Ga.: Tbilisi Univ. Press

Gamkrelidze TV, Ivanov VV. 1995. *Indo-European and the Indo-Europeans: A Reconstruction and Historical Analysis of a Proto-Language and a Proto-Culture*. Berlin: Mouton de Gruyter

Gimbutas M. 1970. Proto-Indo-European culture: the Kurgan culture during the 5th to the 3rd millennia B.C. In *Indo-European and Indo-Europeans*, ed. G Cardona, HM Koenigswald, A Senn, pp. 155–98. Philadelphia: Univ. Pa. Press

Gray RD, Atkinson QD. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426:435–39. **http://doi.org/10.1038/nature02029**

Greenhill SJ, Gray RD. 2009. Austronesian language phylogenies: myths and misconceptions about Bayesian computational methods. In *Austronesian Historical Linguistics and Culture History: A Festschrift for Robert Blust*, ed. KA Adelaar, A Pawley, pp. 375–97. Canberra, Aust.: Pac. Linguist.

Greenhill SJ, Heggarty P, Gray RD. 2020. Bayesian phylolinguistics. In *The Handbook of Historical Linguistics*, ed. RD Janda, BD Joseph, BS Vance, pp. 226–53. Hoboken, NJ: Wiley-Blackwell

Heggarty P. 2014. Prehistory by Bayesian phylogenetics? The state of the art on Indo-European origins. *Antiquity* 88(340):566–77. **http://doi.org/10.1017/S0003598X00101188**

Heggarty P. 2015. Prehistory through language and archaeology. In *The Routledge Handbook of Historical Linguistics*, ed. C Bowern, B Evans, pp. 598–626. London: Routledge. **https://www.routledgehandbooks.com/doi/10.4324/9781315794013.ch28**

Heggarty P. 2018a. Indo-European and the ancient DNA revolution. In *Talking Neolithic: Proceedings of the Workshop on Indo-European Origins Held at the Max Planck Institute for Evolutionary Anthropology, Leipzig, December 2–3, 2013*, ed. G Kroonen, JP Mallory, B Comrie, pp. 120–73. Washington, DC: Inst. Study Man

Heggarty P. 2018b. Why Indo-European? Clarifying cross-disciplinary misconceptions on farming versus pastoralism. In *Talking Neolithic: Proceedings of the Workshop on Indo-European Origins Held at the Max Planck Institute for Evolutionary Anthropology, Leipzig, December 2–3, 2013*, ed. G Kroonen, JP Mallory, B Comrie, pp. 69–119. Washington, DC: Inst. Study Man

Heggarty P, Anderson C, Scarborough M, ... Greenhill SJ, Kühnert D, Gray RD. 2021. *Language trees with sampled ancestors support an early origin of the Indo-European language family*. Work. Pap., Max Planck Inst. Evol. Anthropol., Leipzig, Ger.

Holm HJ. 2011. "Swadesh lists" of Albanian revisited and consequences for its position in the Indo-European languages. *J. Indo-Eur. Stud.* 39(1–2):45–99

Lazzeroni R. 1998. Sanskrit. In *The Indo-European Languages*, ed. AG Ramat, P Ramat, pp. 98–124. London: Routledge

Mallory JP. 1989. *In Search of the Indo-Europeans*. London: Thames & Hudson

Masica CP. 1991. *The Indo-Aryan Languages*. Cambridge, UK: Cambridge Univ. Press

McMahon AMS, Heggarty P, McMahon R, Slaska N. 2005. Swadesh sublists and the benefits of borrowing: an Andean case study. *Trans. Philol. Soc.* 103(2):147–70. **http://doi.org/10.1111/j.1467-968X.2005.00148.x**

McMahon AMS, McMahon R. 2005. *Language Classification by Numbers*. Oxford, UK: Oxford Univ. Press

Nakhleh L, Ringe D, Warnow T. 2005. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language* 81(2):382–420. **www.jstor.org/stable/4489897**

Pereltsvaig A, Lewis MW. 2015. *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*. Cambridge, UK: Cambridge Univ. Press

Rama T. 2018. Three tree priors and five datasets: a study of Indo-European phylogenetics. *Lang. Dyn. Change* 8(2):182–218. **http://doi.org/10.1163/22105832-00802005**

Renfrew C. 1987. *Archaeology and Language: The Puzzle of Indo-European Origins*. London: Jonathan Cape

Ringe DA, Warnow T, Taylor A. 2002. Indo-European and computational cladistics. *Trans. Philol. Soc.* 100(1):59–129. **http://doi.org/10.1111/1467-968X.00091**

Ritchie AM, Ho SYW. 2019. Influence of the tree prior and sampling scale on Bayesian phylogenetic estimates of the origin times of language families. *J. Lang. Evol.* 4(2):108–23. **https://doi.org/10.1093/jole/lzz005**

Rix H, Kümmel MJ, Zehnder T, Lipp R, Schirmer B. 2001. *Lexikon der Indogermanischen Verben (LIV²)*. Wiesbaden, Ger.: Reichert. 2nd ed.

Whitfield J. 2003. Language tree rooted in Turkey. *Nature*. **https://doi.org/10.1038/news031124-6**

Wodtko DS, Irslinger B, Schneider C. 2008. *Nomina im indogermanischen Lexikon (NIL)*. Heidelberg, Ger.: Winter

Yang Z. 2006. *Computational Molecular Evolution*. Oxford, UK: Oxford Univ. Press