There's no escaping phylogenetics*

David Goldstein

Abstract

The comparative method depends crucially on the phylogenetic tree of the languages under comparison, but in many linguistic families, including Indo-European, the true tree is unknown. To circumvent this issue, frequency heuristics have been devised to enable comparative reconstruction over consensus trees. These heuristics come in different forms, but they are all based on the same methodological principle: if the number of homologous elements (e.g., lexical cognates) in the daughter languages meets a minimum threshold (canonically three), their ancestor is reconstructed to the root of the tree. In this paper, I demonstrate that frequency heuristics are not only unreliable but fundamentally misguided. As an alternative, I present a Bayesian method for inferring ancestral states that accounts for phylogenetic uncertainty by estimating the probability of a character state over a set of resolved phylogenetic trees.

1 Introduction

Inferences about Proto-Indo-European are drawn primarily on the basis of the comparative method (Hale 2015, Weiss 2015). It is well known that this method depends crucially on the phylogeny of the languages under comparison (e.g., Hale 2007:225, Olander 2018). Several aspects of the phylogenetic tree of Indo-European are poorly understood, however. Research over the past few decades has elucidated the position of Anatolian, but the order in which other early clades formed remains elusive (Garrett 1999:147, Widmer 2018:374). As a result, Indo-Europeanists commonly subscribe to the topology in Figure 1 (Olander 2018:184, with references to earlier literature), according to which Anatolian and Proto-Nuclear-Indo-European are sisters and the remainder of the tree is unresolved. Some scholars contend that the Tocharian clade was the second to form (e.g., Ringe 2017:6, Weiss 2018), but this view has yet to become the consensus (Malzahn 2017).

*It is with the deepest appreciation that I offer this contribution to Mark. There are few scholars who have moulded my approach to syntax, language change, and linguistic theory as much as he has. His work ushered in a new era in the investigation of archaic Indo-European syntax and paved the way for my own research. I remain indebted to him for years of scholarly guidance and fruitful intellectual debate. I also want to take this opportunity to thank Hans Hock, Stephanie Jamison, Thomas Jügel, Martin Kümmel, Benjamin Slade, and Michael Weiss for answering a variety of questions. Fault for all remaining shortcomings lies solely with me.

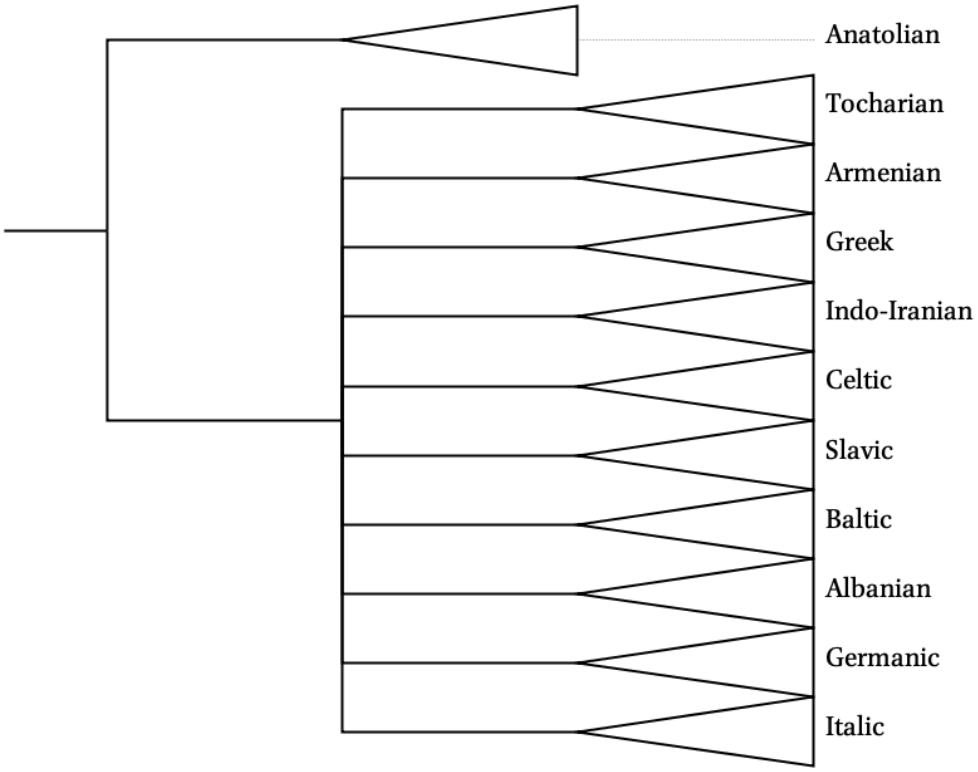

Figure 1: Nuclear Indo-European star phylogeny

The unresolved sub-tree in Figure 1, Nuclear Indo-European, is ambiguous. It can denote the more or less simultaneous diversification of Indo-European, a scenario that biologists refer to as a HARD POLYTOMY (Baum et al. 2013:58). Alternatively, it can represent uncertainty. According to this interpretation, the unresolved sub-tree represents all the possible topologies that could exist under the Nuclear Indo-European node (cf. Hale 2007:242). Biologists use the term SOFT POLYTOMY to describe unresolved trees of this sort (Baum et al. 2013:58–59). Few if any scholars believe that the formation of the Nuclear Indo-European clades began more or less simultaneously, so the star phylogeny in Figure 1 is standardly interpreted as a consensus tree.

If the comparative method depends on the phylogeny and the true tree is unknown, the question arises as to whether linguistic reconstruction is possible at all. In response to this challenge, Indo-Europeanists have devised FREQUENCY HEURISTICS to draw inferences about Proto-Indo-European on the basis of consensus trees such as that in Figure 1. Although these heuristics vary in their details, they are all based on the same methodological principle: if the number of homologous elements (e.g., lexical cognates) in the daughter languages reaches some minimum threshold their ancestor is reconstructed to Proto-Indo-European.

In this paper, I show that frequency heuristics are not only unreliable, but fundamentally misguided. These heuristics are designed to liberate the comparative method from its dependence on tree topology by licensing reconstruction over consensus trees, but the comparative method simply cannot be divorced from phylogenetics (cf. Paradis 2014:4). As an alternative, I present a Bayesian method developed within

evolutionary biology that estimates the probabilities of linguistic states at interior nodes over a set of fully resolved trees. This method thus confronts phylogenetic uncertainty directly, in contrast to frequency heuristics, which circumvent tree topology altogether.

The remainder of this paper is structured as follows. Section 2 critically evaluates two frequency heuristics commonly used for linguistic reconstruction. Section 3 presents a Bayesian method that estimates the probability of ancestral states given uncertainty in the phylogeny. I illustrate this method with two case studies: the prehistory of the conjunction */ǵkʷe/ and the augment. Section 4 brings the paper to a close with brief concluding remarks.

2 Frequency heuristics

In this section, I introduce and evaluate two frequency heuristics, the lower-bound heuristic and the majority-rule heuristic. These heuristics are sometimes coupled with the Anatolian criterion, which is also discussed. To avoid the extraneous complications that often accompany real data, I rely primarily on schematic examples.

2.1 The lower-bound heuristic

The lower-bound heuristic dictates a minimum number of cognates—canonically three—that must be attested for their common ancestor to be assigned to Proto-Indo-European. Gaitzsch et al. (2017:86) write that "An item is considered to be of Proto-Indo-European provenience if it has descendants in at least three derivative languages with nearly the same meaning." Zimmer (2017:78) likewise subscribes to the three-language heuristic, but also introduces considerations of geography and borrowing:

> A venerable rule of thumb, reasonably valued in IE studies at all times, holds that a word may only be ascribed to the common mother tongue if it is attested in at least three languages, preferably in non-contiguous ones, and if no suspicion of loan relations may be raised.

The venerable status of this rule of thumb may be due to its presence in Meillet's *Introduction* (Meillet 1937:380): "La coïncidence de trois langues non contiguës suffit donc pratiquement à garantir le caractère «indo-européen» d'un mot, au sens indiqué ci-dessus." Despite the pedigree of the lower-bound heuristic, it suffers from at least three critical faults.

First, the lower-bound heuristic as presented in the quotations above can only assign the ancestor of homologous elements to the root of a tree and ignores the possibility that their shared ancestor arose only later. For example, in the unresolved tree in the left panel of Figure 2, the three circles share a common ancestor that the lower-bound heuristic assigns to the root of the unresolved tree. If we interpret the unresolved tree as a consensus tree, one of the trees it represents is in the right panel. Given this topology and distribution of circles, their common ancestor would be assigned to a much later interior node. Without knowing how probable the resolved tree is, there is no reason to trust the reconstruction from the consensus tree in the left panel. This simple example reveals that the lower-bound heuristic can overestimate the antiquity of ancestral forms—a problem that is intrinsic to the method, since frequency heuristics ignore topology by design.

When there are multiple cognate sets with at least three members, the lower-bound heuristic demands that the ancestor of each be reconstructed to the root. In Figure 3, for instance, both the circle and the

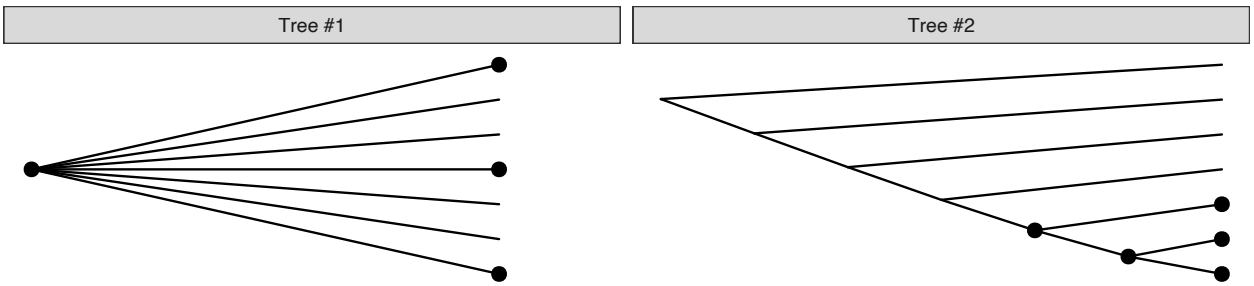

Figure 2: Reconstruction with an unresolved and a resolved tree

triangle meet the lower-bound threshold, so the ancestor of each would be assigned to the root. But it is entirely possible that only one of the character states was present at the root and that the other was a later innovation. In fact, both character states could be later innovations. The lower-bound heuristic cannot countenance either of these possibilities, however, and again is liable to overestimate the antiquity of ancestral forms.

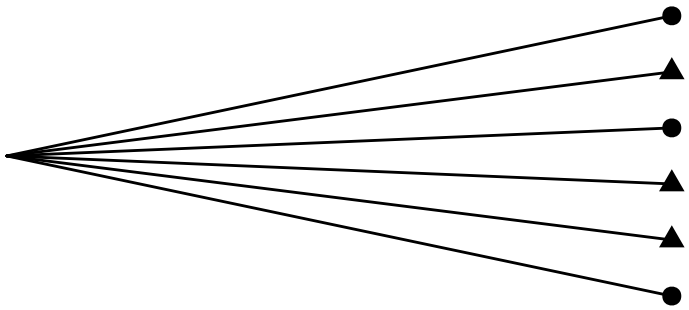

Figure 3: Competing cognate sets

Second, the lower-bound heuristic assumes that unique states are not archaisms. In figure 4, none of the shapes at the tips of the tree are cognate. According to the lower-bound heuristic, these must all be innovations, since no shape meets the critical threshold for being reconstructed to the root. But it is of course entirely possible that the ancestor of one (or more) of the shapes actually did exist at the root. Again the lower-bound heuristic fails to countenance the full range of empirical possibilities.

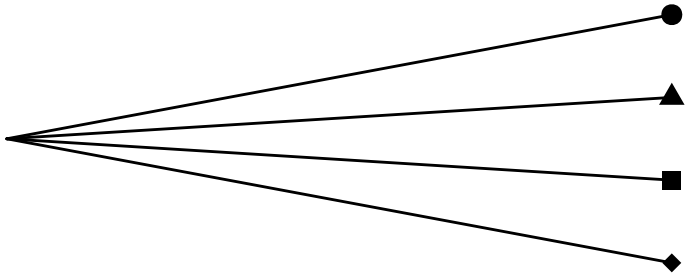

Figure 4: Unique states

Finally, the lower-bound heuristic is unmotivated. Why is three the magic number? Neither Gaitzsch et al. (2017:86) nor Zimmer (2017:78) offers any motivation as to why this particular number should be the

lower bound. I presume that three attained its venerable status on the basis of some sort of probabilistic reasoning, but to the best of my knowledge it has never been divulged.

2.2 The majority-rule heuristic

In contrast to the lower-bound heuristic, which relies on an absolute lower bound, the majority-rule heuristic is relative: if a homologous linguistic element is attested in the majority of the languages being compared, it has the greatest claim to being the ancestral state at the root of the tree (e.g., Fox 1995:84). Prima facie this heuristic seems more reliable than the lower-bound heuristic, since in larger datasets it imposes a higher threshold for reconstruction to the root.

The underpinning of the majority-rule heuristic is the maximum parsimony optimality criterion: linguistic reconstructions requiring fewer changes to account for the observable data are superior. If a linguistic element is attested in a majority of the tips of a star phylogeny, assigning that element to the root of the tree will always be the most parsimonious reconstruction. Consider the unresolved tree in Figure 5, which has five circles and two triangles at the tips. Since circles preponderate, the majority-rule heuristic assigns a circle to the root of the tree.

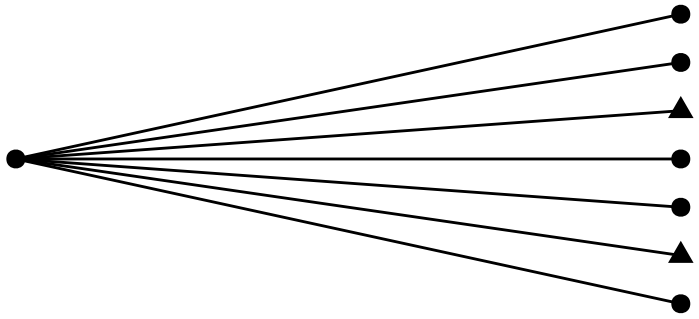

Figure 5: The majority-rule heuristic with an unresolved tree

There are 10,395 possible fully resolved rooted trees for seven languages. Here I focus on the two presented in Figure 6, which are annotated with the most parsimonious ancestral state assignments on their interior nodes. In tree one, the triangle languages form first. In tree two, they form last. One change has taken place on each tree: from triangle to circle on tree one and from circle to triangle on tree two. The state of the root in tree two (circle) agrees with the root state of the unresolved tree in Figure 5, but the state of the root (triangle) in tree one does not.

The two trees in Figure 6 illustrate why the majority-rule heuristic is otiose. If we happen to know that one of these trees is the true tree, the majority-rule heuristic is of no use, since ancestral states will be inferred on the basis of an optimality criterion (e.g., posit as few changes as possible). If we do not know the true tree, the majority-rule heuristic is of no help, since the reconstruction of a triangle or a circle depends on which of the two trees in Figure 6 is more probable. In other words, what matters for reconstruction are the phylogeny and the optimality criterion. This is true not only for the majority-rule heuristic, but also for the lower-bound heuristic.

Our honorand comes to a similar conclusion in his own discussion of the majority-rule heuristic:

> If...a "flat" descent model is a "shorthand" for a set of equally possible subgroupings, then

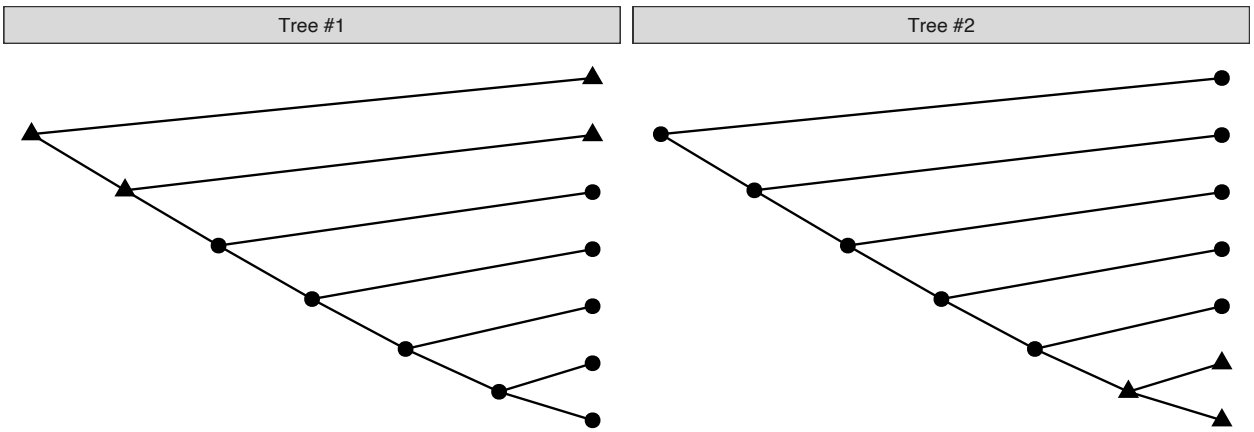

Figure 6: Two resolved topologies with maximum-parsimony reconstructions

> clearly the number of terminal daughters which show a given feature cannot be relevant to the issue of what should be reconstructed for the protolanguage. (Hale 2007:242)

While I certainly agree with Hale's conclusion concerning the majority-rule heuristic, the number of tips that attest a given feature can be relevant to reconstruction. In likelihood-based approaches, for instance, certain character-transition models incorporate the frequency of character states at the tips, which plays a role in estimating the probability of character states at interior nodes (Baum et al. 2013:225–231 provide an accessible discussion of such models).

2.3 The Anatolian criterion

Since Anatolian is now widely agreed to have been the first clade to form, some scholars have argued that Proto-Indo-European ancestry can only be established for cognate sets that include an Anatolian language. Although this criterion is not accepted by everyone, its introduction of a topological constraint is certainly a step in the right direction. The problem is what one is supposed to reconstruct when homologous traits are found exclusively in Anatolian or exclusively in Nuclear Indo-European. Olander (2018:193–194) mentions the case of */kʷékʷlo-/ 'wheel', descendants of which are attested in several branches of Nuclear Indo-European, but not in Anatolian. On account of this absence, some question whether */kʷékʷlo-/ existed as early as Proto-Indo-European and suggest that it arose only after the formation of Anatolian (e.g., Anthony 2007:63–76). There are a number of examples of this type (e.g., the feminine, the dual), which of course have provided the grist for the *Schwundhypothese* debate (see, e.g., Melchert 2018). The Anatolian criterion seems to suggest that homologous traits attested only in Anatolian or only in Nuclear Indo-European should not be reconstructed to Proto-Indo-European. Such a view is clearly unviable, because even in such scenarios it is possible for the ancestor of such traits to have been present in Proto-Indo-European.

2.4 The deeper problem

Intrinsic to the frequency heuristics discussed above is the assumption that reliable linguistic reconstruction via the comparative method is possible with consensus trees. This assumption is untenable. Given that consensus trees represent a range of possible topologies and that the differences among these

topologies can have a crucial impact on the inferences that we draw about ancestral states it is pointless to use them for reconstruction. For all of the uncertainty surrounding the phylogeny of Indo-European, one thing we know for sure is that the consensus sub-tree in Figure 1 is not the true tree, since consensus trees are necessarily incorrect (Yang 2014:129). Why should we expect to draw true inferences from false trees?

3 Reconstruction in the face of phylogenetic uncertainty

We now face a conundrum. The true Indo-European tree is unknown, but reconstruction over consensus trees with frequency heuristics is unviable. How then are we supposed to go about reconstruction? Olander (2018) has recently answered this question by reconstructing aspects of Proto-Indo-European and other early interior nodes on the basis of a specific phylogenetic tree (cf. Adams et al. 1997:555, Ringe 1998, Winter 1998, and Gąsiorowski 1999). His approach is certainly a welcome development, but the accuracy of his inferences depend on the accuracy of the tree that he assumes. In this section, I introduce Bayesian ancestral state estimation, which offers a method for carrying out linguistic reconstruction in the face of phylogenetic uncertainty.[1] This method is illustrated with two case studies in sections 3.1 (the prehistory of the conjunction */ɸkʷe/) and 3.2 (the prehistory of the augment).

Bayesian ancestral state estimation was originally developed within evolutionary biology (e.g., Pagel et al. 2004). At the heart of this method is Bayes' Theorem,[2] which is used to calculate the probability of unattested linguistic states on the basis of the observable data, a set of phylogenetic trees, and a model of linguistic change. This value is referred to as the POSTERIOR PROBABILITY. The analyses below take phylogenetic uncertainty into account by estimating the posterior probability of ancestral states over a random sample of one hundred phylogenetic trees generated from the A3 dataset and model of Chang et al. (2015a).[3] This dataset contains ninety-four Indo-European languages, of which seventy-eight are contemporary and sixteen are ancient. The tree sample is plotted in Figure 7, which reveals variation in topology, branch length, and root age. The maximum a posteriori (MAP) tree from the A3 dataset and model is presented in Figure 8, which is also used to display the results of the case studies in Figures 9 and 10 below.

Linguistic change in Bayesian phylogenetics is standardly modeled as a continuous-time Markov chain (CTMC).[4] CTMCs model language change as a stochastic phenomenon with rate parameters that govern the amount of time between transition events. It is worth highlighting the assumptions that these models bring with them. First, character states at the nodes of a tree are assumed to depend only on the state of their immediate ancestors and the length of the branch along which they evolved (Cathcart 2018:4). Second, the probability of a transition depends only on the current state of a language. Its previous history is irrelevant. This is known as the MARKOV PROPERTY. Finally, rates of gain and loss are assumed not to vary across the tree.

[1] The description of Bayesian inference in this section has been simplified to keep the discussion as accessible and brief as possible. I have accordingly omitted any discussion of prior probability distributions. For more detail on Bayesian ancestral state information, see Pagel et al. 2004 and Ronquist 2004.

[2] For accessible introductions to Bayes' Theorem, see McGrayne 2012 and Stone 2013.

[3] This is configuration file `a1-c2-d0-g2-l2-s1-t1-z3`, the BEAST `.xml` file for which is available in Chang et al. 2015b. To run the `.xml` file, one needs the customized version of BEAST available from `https://github.com/whdc/ieo-beast`. See Chang et al. 2015b:6 for further details.

[4] Cathcart (2018) provides an introduction to the linguistic use of CTMCs.

In most real-world applications of Bayes' Theorem, it is not possible to calculate the posterior probability analytically. The standard practice is instead to use Markov Chain Monte Carlo (MCMC) to sample from the posterior distribution. In the analyses presented below, six independent MCMC chains were run for one hundred thousand generations each, with samples being taken every one hundredth generation.[5] The first twenty-five percent of these samples were then discarded as burn-in for the calculation of the posterior probabilities of ancestral states. Visual inspection confirmed convergence of the six chains.

Although Bayesian ancestral state estimation is not yet common in historical linguistics,[6] these methods offer a number of important advantages. For one, the posterior probabilities of ancestral states need not be based on any specific tree. So linguistic reconstruction can be carried out even when the true tree is unknown. Furthermore, these methods provide a way to quantify the uncertainty of ancestral state inferences. Given that linguistic reconstruction is fundamentally a probabilistic endeavor, measures of uncertainty are absolutely essential. Despite these advantages, Bayesian methods are not a replacement for traditional methods. Indeed, there are things that these methods cannot (or at least cannot yet) achieve. Fine-grained segmental and prosodic reconstruction is not a possibility, for instance.

[5] The analyses in sections 3.1 and 3.2 were performed with `RevBayes` version 1.0.13 (Höhna et al. 2016).

[6] For recent examples, see Haynie et al. 2016, Dunn et al. 2017, Cathcart, Carling, et al. 2018, and Cathcart, Hölzl, et al. 2020.

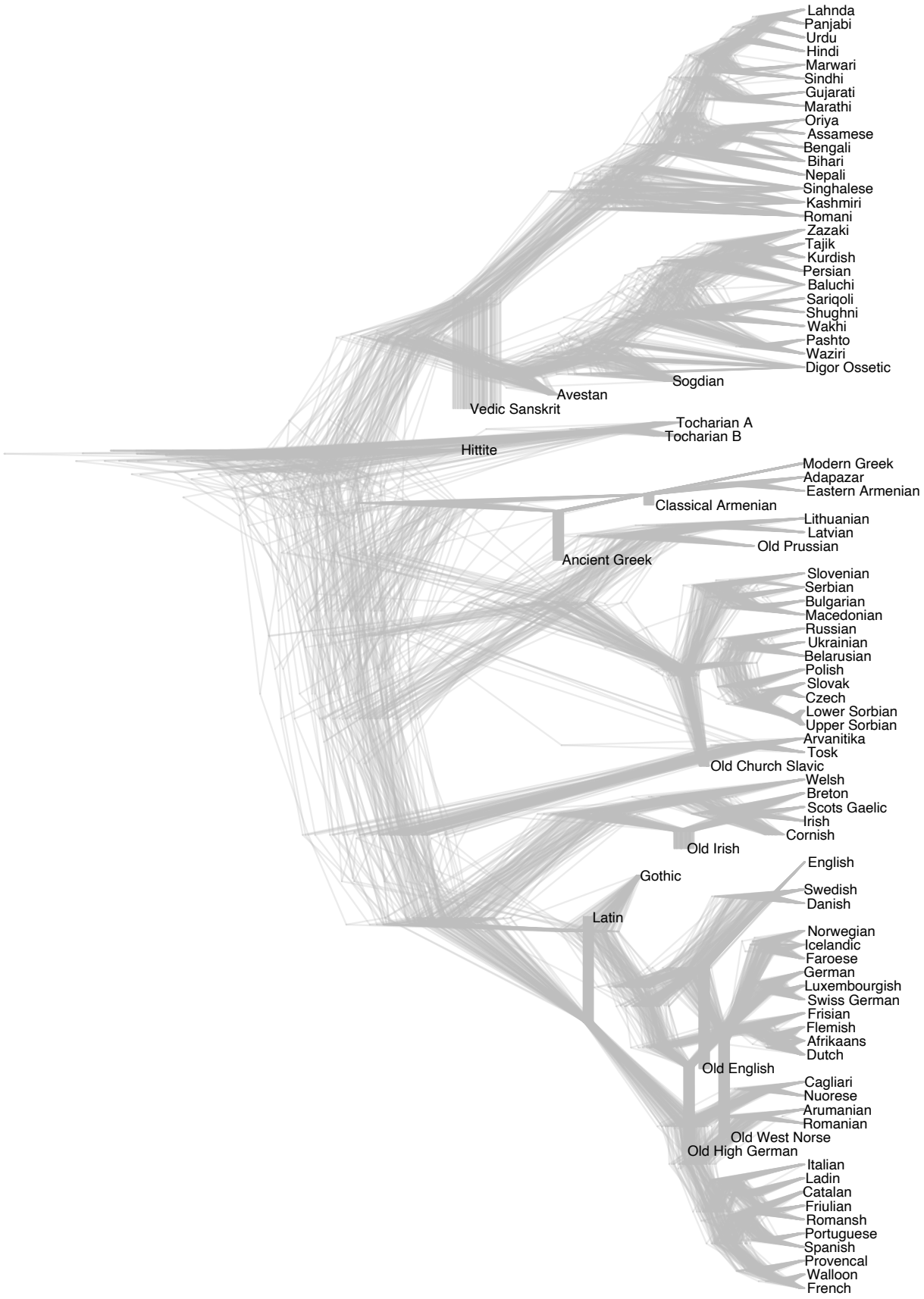

Figure 7: One hundred phylogenetic trees from the A3 dataset and model of Chang et al. 2015

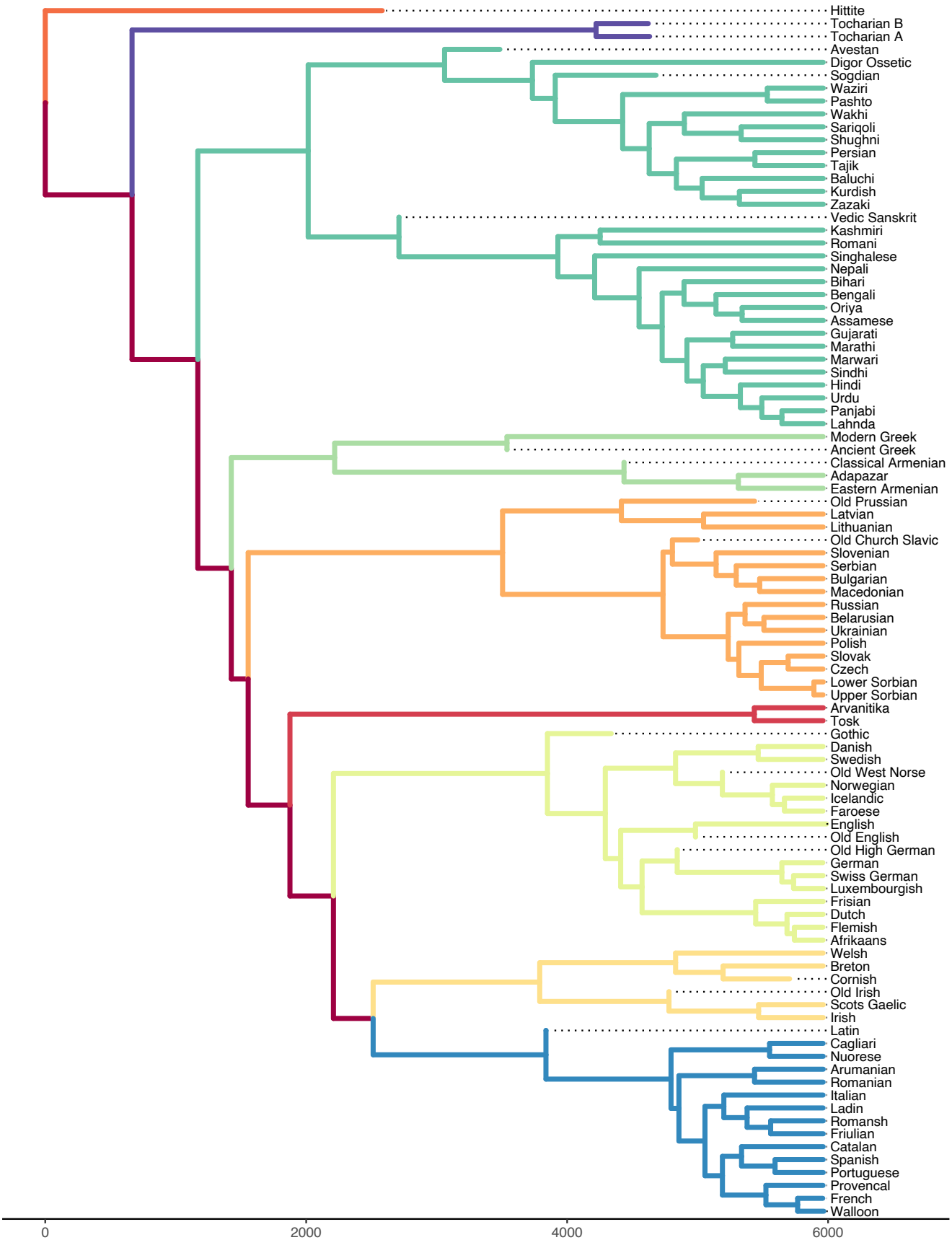

Figure 8: The maximum a posteriori tree for the A3 model and dataset of Chang et al. (2015a)

3.1 The prehistory of the conjunction *$/\!\!=\!\!k^{w}e/$

My first case study investigates the prehistory of *$/\!\!=\!\!k^{w}e/$ 'and', from which a number of conjunction exponents in the earliest archaic Indo-European languages descend:

(1) *Descendants of* *$/\!\!=\!\!k^{w}e/$ 'and'

a. *Hittite* (Watkins 1985:495–496)
[n=aš] ēšzi=pát natta=**ku**w[=aš=apa ar]āi.
'She remains seated **and** she does not get up.'
KBo 19.163 ii.33′–34′ (NH)

b. *Vedic*
devásya mártyasya=**ca**
'of the divine **and** mortal'
RV 2.7.2b

c. *Mycenaean Greek* (*DMG*²:345–346, Miller 2014:299–301)
ta-ra-nu a-ja-me-no / e-re-pa-te-jo / a-to-ro-qo i-qo-**qe** po-ru-po-de-**qe** po-ni-ke-**qe**
'Footstool inlaid with an ivory man **and** a horse **and** an octopus **and** a palm-tree'
PY Ta 722.1

d. *Latin*
de domino bono colono bono=**que** aedificatore melius emetur.
'It is better to buy (a farm) from a good farmer **and** a good builder.'
Cato *Agr.* 1.4

Beyond these four languages, *$/\!\!=\!\!k^{w}e/$ is also continued in Avestan, Oscan, and Celtiberian, among other languages (see, e.g., the collection of evidence in Dunkel 2014:690–692). Given this robust presence among the early daughter languages, a conjunction *$/\!\!=\!\!k^{w}e/$ is standardly reconstructed to Proto-Indo-European (e.g., Mallory et al. 2006:62, 421–422, Fortson 2010:149, Goldstein 2019).[7]

Table 1 presents the languages in my dataset that possess a conjunction descending from *$/\!\!=\!\!k^{w}e/$. To carry out ancestral state estimation, a single binary variable registering the presence or absence of a conjunction descending from *$/\!\!=\!\!k^{w}e/$ was used. The languages in Table 1 were all assigned a value of one; all other languages in my sample were assigned a value of zero.

[7]Descendants of *$/\!\!=\!\!k^{w}e/$ are of course not exclusively conjunctions. Consideration of the full profile of PIE *$/\!\!=\!\!k^{w}e/$ lies beyond the scope of the present discussion, however. For a useful collection of data, see Dunkel 2014:442–446, 689–786.

CLADE	LANGUAGE	CONJUNCTION	REFERENCE
Anatolian	Hittite	*=(k)ku*	*HED*:204–205, *EDHIL*:483–484
Celtic	Old Irish	*-ch*	Thurneysen 1921:299
Italic	Latin	*=que*	*LEW*:401–402, *DELL*:555
Indic	Vedic	*=ca*	*KEWA*:365
Indic	Singhalese	*=t*/*=d*	*CDIAL*:246, Slade 2011:158 n. 11, Masica 1991:398
Iranian	Avestan	*=čā*	Rastorgueva et al. 2003:195
Greek	Greek	=τε	*EDG*:1457
Germanic	Gothic	*=h*	Lehmann 1986:374

Table 1: Conjunctions that descend from */=kʷe/

The results of my analysis are presented in Figure 9. The circles at the tips of the tree are either solidly black (which denotes the absence of a conjunction descending from */=kʷe/) or solidly gray (which denotes the presence of such a conjunction) because these states can be observed and are therefore known with certainty. The interior nodes of the tree are annotated with pie graphs displaying the posterior probability of */=kʷe/. The posterior probability of the conjunction at the root of the tree and along its spine is above ninety-five percent. It is only within the later histories of the major clades that the absence of */=kʷe/ begins to predominate. In short, we can be confident that */=kʷe/ existed in Proto-Indo-European. The results of my analysis thus agree with the traditional reconstruction.

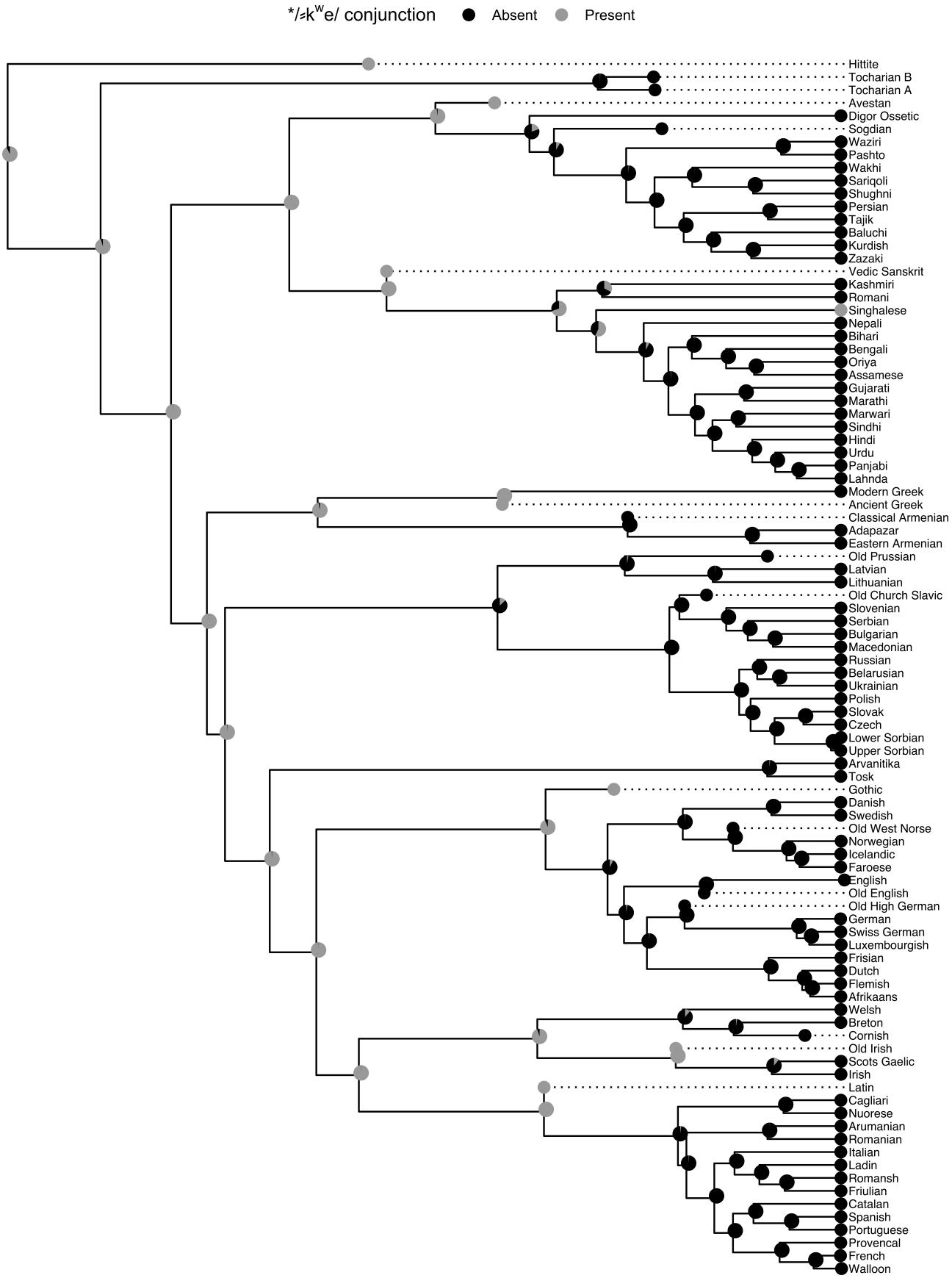

Figure 9: Posterior probabilities of the conjunction */ɟkʷe/ and its descendants in Indo-European

3.2 The prehistory of the augment

My second case study takes up a far more uncertain case of reconstruction, that of the augment. Among the archaic Indo-European languages, the augment is attested in the following languages:[8]

(2) *The archaic augment languages*

a. *Sanskrit*
a-dāt
AUG-give.3SG.IMPF.ACT.IND
'He has given, he gave'

b. *Avestan*
a-məhmaidī
AUG-think.1PL.AOR.MED.IND
'We have thought'

c. *Old Persian*
a-bara
AUG-bear.3SG.IMPF.ACT.IND
'He bore, carried'

d. *Greek*
ἔ-δωκε(ν)
AUG-give.3SG.AOR.ACT.IND
'He gave'

e. *Phrygian*
e-daes
AUG-set.up.3SG.AOR.ACT.IND
'He set up'

f. *Classical Armenian*
e-tes
AUG-see.3SG.AOR.ACT.IND
'He saw'

The distribution of the augment is not uniform among these languages. Within Greek, it appears at most twice in Mycenaean (PY Fr 1184.1 and PY An 724). In Homer, its appearance with finite past indicative verbs varies, but the conditioning factors have yet to be worked out (e.g., Mumm 2004, Bakker 2005, Willi 2018:358–392). By the time of classical Greek, the augment is obligatory with such verb forms. In Classical Armenian, the augment is preserved in aorist forms that would have been monosyllabic without it (e.g., Klein 2007:1074). In Vedic, the appearance of the augment varies, but in Classical Sanskrit it has become obligatory. In Avestan, the augment is rare, but in Old Persian it is obligatory (Skjærvø 2007:865–866). The augment in Indo-Iranian is remarkable for being prefixed to the future stem in the Sanskrit conditional (Cardona 2007:789–790) and to optative verb forms in Old Persian and Young Avestan (Skjærvø 2009:87, 90, 213). Such patterns are not found, for instance, in Greek. In later Indic, the augment survives in Pali and evidently in the modern languages Khawar and Kalasha (Masica 1991:289). In later Iranian, it survives in Middle Persian, Chorasmian, Sogdian, Tumshuqese, and modern Yaghnobi (Skjaervø 2006:22). It is also present in Modern Greek.

A number of scholars argue for the PIE antiquity of the augment (e.g., Brugmann 1916:13, Hoffmann 1970:530, Rix [1976] 1992:§246, Meiser 2002:§34.7, Mallory et al. 2006:65, Tichy 2009:54, 125–126, Meier-Brügger 2010:315, Beekes 2011:252), but it is probably the case that more consider it a post-PIE innovation (e.g., Meillet 1908:97–101, Porzig 1954:87, Lehmann 1993:244, Sihler 1995:484–485, Szemerényi 1996:297, Bartolotta 2009:511, Fortson 2010:§5.44, Hajnal 2009, Drinka 2013:385, Zahn 2014:119, Bartolotta 2017, Matasović 2017:22, Ringe 2017:30, Lundquist et al. 2018:2141). Weiss (2011:384 n. 28) deems the evidence too uncertain to allow for a conclusion either way. Scholars who favor a post-PIE innovation rarely identify

[8]The meanings of augmented forms in Homeric Greek and Vedic Sanskrit have been the source of dispute. The glosses in example (2) are provided only to give the reader a basic sense of the meaning of the verbal forms and should not be interpreted as an attempt to weigh in on any of the debates.

the point at which the augment emerged. Zahn (2014:119) is exceptional in this regard: he maintains that the augment is an innovation of Greco-Indo-Iranian.

The reconstruction of the augment raises a number of issues. Tense, aspect, and mood are predominantly realized by suffixation in archaic Indo-European, so the augment stands out for being a prefix. This development is presumably unusual enough that the augment was either inherited from PIE or emerged once. If it only arose once, it could have developed in the immediate common ancestor of all the languages that possessed the augment, i.e., the first node containing Greek, Phrygian, Armenian, and Indo-Iranian. The problem is that it is anything but clear that these languages formed a clade. An additional complication is the possibility that the augment once existed in more clades than those for which we currently have evidence. Indeed, attempts have been made to ferret out traces of the augment in Hittite, Tocharian, Germanic, Baltic, Slavic, Italic, Celtic, and Albanian (see Szemerényi 1996:297, Olander 2018:190, Willi 2018:357 n. 1), but they have met with little to no acceptance. Alternatively, the augment could have arisen once and then spread to neighboring languages. Porzig (1954:87) and Drinka (2013:385, 401) propose accounts of this type.

Reconstructing the augment to Proto-Indo-European circumvents the problems that arise under an innovation analysis, but such a move runs into other issues, the most obvious of which is the suspicious pattern of survival. Greek, Phrygian, Armenian, and Indo-Iranian not only form a geographic band, but also share a number of linguistic characteristics (e.g., Euler 1979, Clackson 1994). One might reasonably expect the survival of the augment to be more evenly distributed among the major clades. The second problem is that under an inheritance analysis it is difficult to account for the variable appearance of the augment in Mycenaean, Homeric Greek, Vedic, and Avestan. One would have to assume that the situation in PIE was similar to what we find in these languages (so Brugmann 1916:13, Meier-Brügger 2010:315), but such an analysis is difficult to implement given that the distribution of the augment in these languages seems to differ. Positing a non-obligatory augment (whatever exactly that would mean) in PIE makes it easier to account for its loss in most of the family, but it would be puzzling that it suddenly becomes obligatory in classical Greek and Sanskrit after what would have been millennia of non-obligatory behavior. A third issue is the absence of relics. If the augment were inherited from PIE, one could reasonably expect to find lexicalized relics of augmented forms in languages that have otherwise abandoned the morpheme. As noted above, a number of scholars have claimed to have discovered such relics, but their proposals have not met with widespread acceptance.

These are just the issues pertaining to reconstruction. There is of course an array of questions concerning morphosyntax, semantics, and pragmatics. Space constraints forbid a proper treatment of these questions, so I will limit myself to highlighting a critical aspect of the augment that has yet to be sufficiently investigated, namely what morphosyntactic property (or perhaps properties) it realizes.[9] Given its appearance among finite past indicative verbal forms in classical Greek, for instance, it is plausible that the augment realized the feature [PAST], at least in this language, but this remains to be demonstrated.

To carry out ancestral state estimation, the augment was represented as a binary variable with values for presence and absence. The following augment languages are present in my dataset: Classical and Modern Greek, Classical Armenian, Vedic Sanskrit, Avestan, and Sogdian. These languages were assigned

[9]Willi (2018) has recently argued that the augment developed from the reduplicated syllable of the reduplicated aorist, which realized the value [PERFECTIVE]. At a minimum, this account is incomplete because it offers no analysis of the inflectional morphology of reduplicated aorists beyond the reduplicant. One is left to wonder, for instance, what morphosyntactic properties the stem itself realized.

to state one; the remaining languages in my sample were assigned to state zero. This representation of the data brings with it two drawbacks. First, in certain languages (e.g., Vedic Sanskrit and Avestan), the distribution of the augment is not categorical (i.e., it is not prefixed to every past-referring finite indicative verb form). Second, not all of the languages in which the augment is attested are in my dataset.

The posterior probabilities in Figure 10 support the view that the augment is a post-PIE innovation, but the posterior probability of the absence of the augment at the root is only fifty-six percent.[10] In other words, we cannot discount the possibility of the augment in Proto-Indo-European. The prehistory of the augment drives home the point that uncertainty in the topology can make for uncertainty in reconstruction. If the evidence for a clade containing only and all the augment languages were robust, the augment would straightforwardly be deemed an innovation of that node. The posterior probabilities in Figure 10 make it clear that our understanding of the verbal morphology of Proto-Indo-European and Proto-Nuclear-Indo-European is still very much an open question.

4 Conclusion

The use of consensus trees and their associated frequency heuristics for linguistic reconstruction should be abandoned, since the comparative method only makes sense in the light of phylogenetics. It remains to be seen how much of Proto-Indo-European rests on inferences from frequency heuristics and how much will consequently have to be revised. To infer unattested linguistic states in the face of phylogenetic uncertainty, I presented a Bayesian method developed in evolutionary biology. The results of my experiments accord with the communis opinio of the prehistory of the conjunction */ɸkʷe/ and support the view that the augment is a post-PIE innovation, although there is more uncertainty here. If nothing else, these results make it clear that Bayesian methods provide a powerful and exciting addition to the Indo-Europeanist's toolkit.

The data and code used for this paper are archived at `10.5281/zenodo.4302668`.

[10] I should add that there are other transition models that one could explore, some of which may offer different results. I plan to take up the question of model comparison on another occasion.

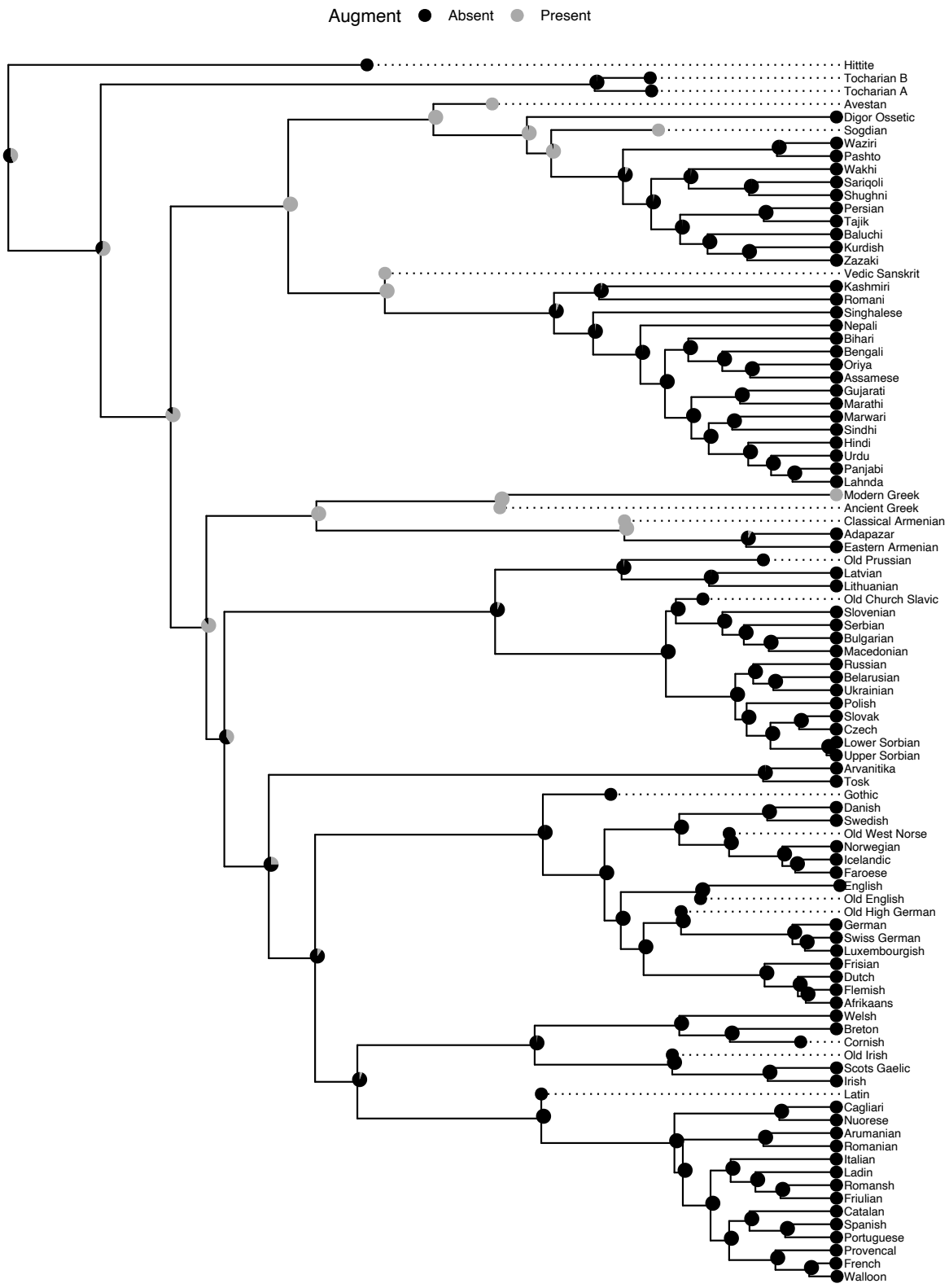

Figure 10: Posterior probabilities of the augment in Indo-European

Abbreviations

CDIAL Ralph Lilley Turner (1962–1966). *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press.

DELL Alfred Ernout and Antoine Meillet (1959). *Dictionnaire étymologique de la langue latine. Histoire des mots*. 4th ed. Paris: Klincksieck.

DMG² Michael Ventris and John Chadwick (1973). *Documents in Mycenaean Greek*. 2nd ed. Cambridge: Cambridge University Press.

EDG Robert S. P. Beekes (2010). *Etymological dictionary of Greek*. 2 vols. Leiden Indo-European Etymological Dictionary Series 10. Leiden: Brill.

EDHIL Alwin Kloekhorst (2008). *Etymological dictionary of the Hittite inherited lexicon*. Leiden Indo-European Etymological Dictionary Series 5. Leiden: Brill.

HED Jaan Puhvel (1997). *Hittite etymological dictionary. Words beginning with K*. Vol. 4. Berlin: de Gruyter.

KEWA Manfred Mayrhofer (1956–1980). *Kurzgefasstes etymologisches Wörterbuch des Altindischen*. 4 vols. Heidelberg: Winter.

LEW Alois Walde and Johann B. Hofmann (1938–1954). *Lateinisches etymologisches Wörterbuch*. 3rd ed. 3 vols. Heidelberg: Winter.

References

Adams, Douglas Q. and James P. Mallory (1997). Subgrouping. *Encyclopedia of Indo-European culture*. London: Fitzroy Dearborn, 550–556.

Anthony, David W. (2007). *The horse, the wheel, and language*. Princeton: Princeton University Press.

Bakker, Egbert J. (2005). *Pointing at the past. From formula to performance in Homeric poetics*. Cambridge, MA: Harvard University Press.

Bartolotta, Annamaria (2009). Root lexical features and inflectional marking of tense in Proto-Indo-European. *Journal of Linguistics* 45.3 (Nov. 2009), 505–532. DOI: 10.1017/S0022226709990016.

– (2017). On syntactic diagnostics as tests for telicity in ancient Indo-European languages. Evidence from Vedic and Greek. *Incontri Linguistici* 40, 39–63. DOI: 10.19272/201700801003.

Baum, David A. and Stacey D. Smith (2013). *Tree thinking. An introduction to phylogenetic biology*. Greenwood Village, CO: Roberts and Co.

Beekes, Robert S. P. (2010). *Etymological dictionary of Greek*. 2 vols. Leiden Indo-European Etymological Dictionary Series 10. Leiden: Brill.

– (2011). *Comparative Indo-European linguistics. An introduction*. 2nd ed. Amsterdam: John Benjamins.

Brugmann, Karl (1916). *Grundriss der vergleichenden Grammatik der indogermanischen Sprachen. Lehre von den Wortformen und ihrem Gebrauch*. 2nd ed. Vol. 2. 3 vols. 3. Strassburg: Trübner.

Cardona, George (2007). Sanskrit morphology. *Morphologies of Asia and Africa*. Ed. by Alan S. Kaye. Vol. 2. Winona Lake, IN: Eisenbrauns, 775–824.

Cathcart, Chundra Aroor (2018). Modeling linguistic evolution. A look under the hood. *Linguistics Vanguard* 4.1, 1–11. DOI: 10.1515/lingvan-2017-0043.

Cathcart, Chundra Aroor, Gerd Carling, Filip Larsson, Niklas Johansson, and Erich R. Round (2018). Areal pressure in grammatical evolution. An Indo-European case study. *Diachronica* 35.1, 1–34. DOI: 10.1075/dia.16035.cat.

Cathcart, Chundra Aroor, Andreas Hölzl, Gerhard Jäger, Paul Widmer, and Balthasar Bickel (2020). Numeral classifiers and number marking in Indo-Iranian. *Language Dynamics and Change* 11.2, 1–53. DOI: 10.1163/22105832-bja10013.

Chang, Will, Chundra Aroor Cathcart, David P. Hall, and Andrew J. Garrett (2015a). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91.1 (Mar. 2015), 194–244. DOI: 10.1353/lan.2015.0005.

– (2015b). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. Supplementary materials. *Language* 91.1 (Mar. 2015). DOI: 10.1353/lan.2015.0007.

Clackson, James P. T. (1994). *The linguistic relationship between Armenian and Greek*. Oxford: Blackwell.

Drinka, Bridget (2013). Phylogenetic and areal models of Indo-European relatedness. The role of contact in reconstruction. *Journal of Language Contact* 6.2, 379–410. DOI: 10.1163/19552629-00602009.

Dunkel, George E. (2014). *Lexikon der indogermanischen Partikeln und Pronominalstämme. Lexikon*. Vol. 2. Heidelberg: Winter.

Dunn, Michael, Tonya Kim Dewey, Carlee Arnett, Thórhallur Eythórsson, and Jóhanna Barðdal (2017). Dative sickness. A phylogenetic analysis of argument structure evolution in Germanic. *Language* 93.1 (Mar. 2017), e1–e22.

Ernout, Alfred and Antoine Meillet (1959). *Dictionnaire étymologique de la langue latine. Histoire des mots*. 4th ed. Paris: Klincksieck.

Euler, Wolfram (1979). *Indoiranisch-griechische Gemeinsamkeiten der Nominalbildung und deren indogermanische Grundlagen*. Innsbruck: Institut für Sprachwissenschaft.

Fortson, Benjamin W. IV (2010). *Indo-European language and culture. An introduction*. 2nd ed. Malden, MA: Blackwell.

Fox, Anthony (1995). *Linguistic reconstruction. An introduction to theory and method*. Oxford: Oxford University Press.

Gaitzsch, Torsten and Johann Tischler (2017). The homeland of the speakers of Proto-Indo-European. *Handbook of comparative and historical Indo-European linguistics*. Ed. by Jared S. Klein, Brian D. Joseph, and Matthias Fritz. Vol. 1. Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 41. Berlin: de Gruyter, 85–92. DOI: 10.1515/9783110261288.

Garrett, Andrew J. (1999). A new model of Indo-European subgrouping and dispersal. *Proceedings of the Twenty-Fifth Annual Meeting of the Berkeley Linguistics Society, February 12–15, 1999*. Ed. by Steve S. Chang, Lily Liaw, and Josef Ruppenhofer. Berkeley: Berkeley Linguistics Society, 146–156. DOI: 10.3765/bls.v25i1.1180.

Gąsiorowski, Piotr (1999). The tree of language. A cladistic look at the genetic classification of languages. *Dialectologia et Geolinguistica*, 39–57. DOI: 10.1515/dig.1999.1999.7.39.

Goldstein, David M. (2019). Language change and linguistic theory. The case of archaic Indo-European conjunction. *Transactions of the Philological Society* 117.1 (Mar. 2019), 1–34. DOI: 10.1111/1467-968X.12139.

Hajnal, Ivo (2009). Zur Rekonstruktion und Segmentierung des Indogermanischen. Wolfgang Meids Beitrag aus heutiger Sicht. MS., Universität Innsbruck.

Hale, Mark R. (2007). *Historical linguistics. Theory and method*. Oxford: Blackwell.

– (2015). The comparative method. Theoretical issues. *The Routledge handbook of historical linguistics*. Ed. by Claire Bowern and Bethwyn Evans. London: Routledge, 146–160. DOI: 10.4324/9781315794013.ch5.

Haynie, Hannah J. and Claire Bowern (2016). Phylogenetic approach to the evolution of color term systems. *Proceedings of the National Academy of Sciences of the United States of America* 113.48, 13666–13671. DOI: 10.1073/pnas.1613666113.

Hoffmann, Karl (1970). Das Kategoriensystem des indogermanischen Verbums. *Münchener Studien zur Sprachwissenschaft* 28, 19–41.
Höhna, Sebastian, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, and Fredrik Ronquist (2016). RevBayes. Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology* 65.4 (July 2016), 726–736. DOI: 10.1093/sysbio/syw021.
Klein, Jared S. (2007). Classical Armenian morphology. *Morphologies of Asia and Africa*. Ed. by Alan S. Kaye. Vol. 2. Winona Lake, IN: Eisenbrauns, 1051–1086.
Kloekhorst, Alwin (2008). *Etymological dictionary of the Hittite inherited lexicon*. Leiden Indo-European Etymological Dictionary Series 5. Leiden: Brill.
Lehmann, Winfred P. (1986). *A Gothic etymological dictionary*. Leiden: Brill.
– (1993). *Theoretical bases of Indo-European linguistics*. London: Routledge.
Lundquist, Jesse S. and Anthony D. Yates (2018). The morphology of Proto-Indo-European. *Handbook of comparative and historical Indo-European linguistics. An international handbook of language comparison and the linguistic reconstruction of Indo-European*. Ed. by Jared S. Klein, Brian D. Joseph, and Matthias A. Fritz. Vol. 3. Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 41. Berlin: de Gruyter, 2079–2195. DOI: 10.1515/9783110542431-043.
Mallory, James P. and Douglas Q. Adams (2006). *The Oxford introduction to Proto-Indo-European and the Proto-Indo-European world*. Oxford: Oxford University Press.
Malzahn, Melanie (2017). The second one to branch off? The Tocharian lexicon revisited. *Etymology and the European lexicon*. Ed. by Bjarne Simmelkjær Sandgaard Hansen, Benedicte Nielsen Whitehead, Thomas Olander, and Birgit Anette Olsen. Wiesbaden: Reichert, 281–292.
Masica, Colin P. (1991). *The Indo-Aryan languages*. Cambridge: Cambridge University Press.
Matasović, Ranko (2017). The sources for Indo-European reconstruction. *Handbook of comparative and historical Indo-European linguistics*. Ed. by Jared S. Klein, Brian D. Joseph, and Matthias Fritz. Vol. 1. Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 41. Berlin: de Gruyter, 20–25. DOI: 10.1515/9783110261288.
Mayrhofer, Manfred (1956–1980). *Kurzgefasstes etymologisches Wörterbuch des Altindischen*. 4 vols. Heidelberg: Winter.
McGrayne, Sharon Bertsch (2012). *The theory that would not die. How Bayes' rule cracked the Enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. New Haven: Yale University Press.
Meier-Brügger, Michael (2010). *Indogermanische Sprachwissenschaft*. 9th ed. Berlin: de Gruyter.
Meillet, Antoine (1908). *Les dialectes indo-européens*. Paris: Champion.
– (1937). *Introduction à l'étude comparative des langues indo-européennes*. 8th ed. Paris: Hachette.
Meiser, Gerhard (2002). *Historische Laut- und Formenlehre der lateinischen Sprache*. 2nd ed. Darmstadt: Wissenschaftliche Buchgesellschaft.
Melchert, H. Craig (2018). Hittite and Indo-European. Revolution and counter-revolution. *100 Jahre Entzifferung des Hethitischen—Morphosyntaktische Kategorien in Sprachgeschichte und Forschung. Arbeitstagung der lndogermanischen Gesellschaft Philipps-Universitat Marburg, 21. bis 23. September 2015*. Ed. by Elisabeth Rieken. Wiesbaden: Reichert, 289–294.
Miller, D. Gary (2014). *Ancient Greek dialects and early authors*. Berlin: de Gruyter.
Mumm, Peter-Arnold (2004). Zur Funktion des homerischen Augments. *Analecta homini universali dicata. Arbeiten zur Indogermanistik, Linguistik, Philologie, Politik, Musik und Dichtung: Festschrift für Oswald*

Panagl zum 65. Geburtstag. Ed. by Oswald Panagl, Thomas Krisch, Thomas Lindner, and Ulrich Müller. Vol. 1. Stuttgart: Verlag Hans-Dieter Heinz, 148–158.

Olander, Thomas (2018). Connecting the dots. The Indo-European family tree as a heuristic device. *Proceedings of the 29th Annual UCLA Indo-European Conference*. Ed. by David M. Goldstein, Stephanie W. Jamison, and Brent H. Vine. Bremen: Hempen, 181–202.

Pagel, Mark, Andrew Meade, and Daniel Barker (2004). Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* 53.5 (Oct. 2004), 673–684. DOI: 10.1080/10635150490522232.

Paradis, Emmanuel (2014). An introduction to the phylogenetic comparative method. *Modern phylogenetic comparative methods and their application in evolutionary biology. Concepts and practice*. Ed. by László Zsolt Garamszegi. Berlin: Springer, 3–18. DOI: 10.1007/978-3-662-43550-2.

Porzig, Walter (1954). *Die Gliederung des indogermanischen Sprachgebiets*. Heidelberg: Winter.

Puhvel, Jaan (1997). *Hittite etymological dictionary. Words beginning with K*. Vol. 4. Berlin: de Gruyter.

Rastorgueva, Vera Sergeevna and Joy I. Edelman (2003). *Этимологический словарь иранских языков. B–D*. Vol. 2. Москва: Восточная литература РАН.

Ringe, Donald A. (1998). Some consequences of a new proposal for subgrouping the IE family. *Proceedings of the Twenty-Fourth Annual Meeting of the Berkeley Linguistics Society. Special session on Indo-European subgrouping and internal relations*. Ed. by Benjamin K. Bergen, Madelaine C. Plauché, and Ashlee C. Bailey. Berkeley: Berkeley Linguistics Society, 32–46.

– (2017). *A linguistic history of English. From Proto-Indo-European to Proto-Germanic*. 2nd ed. Vol. 1. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199284139.001.0001.

Rix, Helmut ([1976] 1992). *Historische Grammatik des Griechischen*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Ronquist, Fredrik (2004). Bayesian inference of character evolution. *Trends in Ecology and Evolution* 19.9 (Sept. 2004), 475–481. DOI: 10.1016/j.tree.2004.07.002.

Sihler, Andrew L. (1995). *New comparative grammar of Greek and Latin*. Oxford: Oxford University Press.

Skjærvø, Prods Oktor (2007). Avestan and Old Persian morphology. *Morphologies of Asia and Africa*. Ed. by Alan S. Kaye. Vol. 2. Winona Lake, IN: Eisenbrauns, 853–940.

– (2009). Old Iranian. *The Iranian Languages*. London: Routledge, 43–195.

Skjaervø, Prods Oktor (2006). Iranian languages. *Encyclopedia of language and linguistics*. Ed. by Keith Brown. 2nd ed. Vol. 6. Oxford: Elsevier, 18–22.

Slade, Benjamin (2011). Formal and philological inquiries into the nature of interrogatives, indefinites, disjunction, and focus in Sinhala and other languages. PhD thesis. University of Illinois, Urbana-Champaign.

Stone, James V. (2013). *Bayes' rule. A tutorial introduction to Bayesian analysis*. Sheffield: Sebtel.

Szemerényi, Oswald J. L. (1996). *Introduction to Indo-European linguistics*. Oxford: Oxford University Press.

Thurneysen, Rudolf (1921). Allerlei Keltisches. *Zeitschrift für celtische Philologie* 13.1, 297–304.

Tichy, Eva (2009). *Indogermanistisches Grundwissen*. 3rd ed. Bremen: Ute Hempen.

Turner, Ralph Lilley (1962–1966). *A comparative dictionary of Indo-Aryan languages*. London: Oxford University Press.

Ventris, Michael and John Chadwick (1973). *Documents in Mycenaean Greek*. 2nd ed. Cambridge: Cambridge University Press.

Walde, Alois and Johann B. Hofmann (1938–1954). *Lateinisches etymologisches Wörterbuch*. 3rd ed. 3 vols. Heidelberg: Winter.

Watkins, Calvert (1985). Indo-European *-$k^{w}e$ 'and' in Hittite. *Sprachwissenschaftliche Forschungen. Festschrift für Johann Knobloch zum 65. Geburtstag*. Ed. by Hermann Ölberg and Gemot Schmidt. Innsbruck: Institut für Sprachwissenschaft, 491–497.

Weiss, Michael (2011). *Outline of the historical and comparative grammar of Latin*. 2nd ed. Ann Arbor: Beech Stave Press.

– (2015). The comparative method. *The Routledge handbook of historical linguistics*. Ed. by Claire Bowern and Bethwyn Evans. London: Routledge, 127–145. DOI: 10.4324/9781315794013.ch4.

– (2018). Tocharian and the west. *Priscis libentius et liberius novis. Indogermanische und sprachwissenschaftliche Studien. Festschrift für Gerhard Meiser zum 65. Geburtstag*. Ed. by Olav Hackstein and Andreas Opfermann. Hamburg: Baar, 373–381.

Widmer, Paul (2018). Indogermanische Stammbäume. Datentypen, Methoden. *100 Jahre Entzifferung des Hethitischen. Morphosyntaktische Kategorien in Sprachgeschichte und Forschung Akten der Arbeitstagung der Indogermanischen Gesellschaft vom 21. bis 23. September 2015 in Marburg*. Ed. by Elisabeth Rieken, Ulrich Geupel, and Theresa Maria Roth. Wiesbaden: Reichert, 373–388.

Willi, Andreas (2018). *Origins of the Greek verb*. Cambridge: Cambridge University Press. DOI: 10.1017/9781108164207.

Winter, Werner (1998). Lexical archaisms in the Tocharian languages. *The Bronze Age and early Iron Age peoples of eastern Central Asia. Archeology, migration and nomadism, linguistics*. Ed. by Victor H. Mair. Vol. 1. Washington, D. C.: Intitute for the Study of Man, 347–357.

Yang, Ziheng (2014). *Molecular evolution. A statistical approach*. Oxford: Oxford University Press.

Zahn, Ingo (2014). *Vergleichende indogermanische Formenlehre*. Hamburg: Kovač.

Zimmer, Stefan (2017). The culture of the speakers of Proto-Indo-European. *Handbook of comparative and historical Indo-European linguistics*. Ed. by Jared S. Klein, Brian D. Joseph, and Matthias Fritz. Vol. 1. Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 41. Berlin: de Gruyter, 75–85. DOI: 10.1515/9783110261288.