Alexei S. Kassian*, Mikhail Zhivlov, George Starostin, Artem A. Trofimov, Petr A. Kocharov, Anna Kuritsyna and Mikhail N. Saenko

Rapid radiation of the inner Indo-European languages: an advanced approach to Indo-European lexicostatistics

https://doi.org/10.1515/ling-2020-0060
Received March 31, 2020; accepted August 21, 2020; published online June 18, 2021

Abstract: In this article we present a new reconstruction of Indo-European phylogeny based on 13 110-item basic wordlists for protolanguages of IE subgroups (Proto-Germanic, Proto-Slavic, etc.) or ancient languages of the corresponding subgroups (Hittite, Ancient Greek, etc.). We apply reasonably formal techniques of linguistic data collection and post-processing (onomasiological reconstruction, derivational drift elimination, homoplastic optimization) that have been recently proposed or specially developed for the present study. We use sequential phylogenetic workflow and obtain a consensus tree based on several algorithms (Bayesian inference, maximum parsimony, neighbor joining; without topological constraints applied). The resulting tree topology and datings are entirely compatible with established expert views. Our main finding is the multifurcation

***Corresponding author: Alexei S. Kassian**, School of Advanced Studies in the Humanities, The Russian Presidential Academy of National Economy and Public Administration, Prospect Vernadskogo, 84, Bldg. 2, 119571, Moscow, Russia, E-mail: a.kassian@gmail.com
Mikhail Zhivlov, Institute for Oriental and Classical Studies, National Research University Higher School of Economics, Moscow, Russia; and Institute for Oriental and Classical Studies, Russian State University for the Humanities, Moscow, Russia, E-mail: zhivlov@yandex.ru
George Starostin, Institute for Oriental and Classical Studies, National Research University Higher School of Economics, Moscow, Russia; and Santa Fe Institute, Santa Fe, NM, USA, E-mail: gstarst1@gmail.com
Artem A. Trofimov, School of Advanced Studies in the Humanities, The Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia, E-mail: artemii.trofimov@gmail.com
Petr A. Kocharov, Department of Indo-European Studies and Areal Linguistics, Institute for Linguistic Studies of the Russian Academy of Sciences, Saint Petersburg, Russia, E-mail: peter.kocharov@gmail.com
Anna Kuritsyna, Institute for Oriental and Classical Studies, National Research University Higher School of Economics, Moscow, Russia, E-mail: linnansaari@gmail.com
Mikhail N. Saenko, Department of Slavic Linguistics, Institute of Slavic Studies of the Russian Academy of Sciences, Moscow, Russia, E-mail: veraetatis@yandex.ru

of the Inner IE clade into four branches *ca.* 3357–2162 BC: (1) Greek-Armenian, (2) Albanian, (3) Italic-Germanic-Celtic, (4) Balto-Slavic–Indo-Iranian. The proposed radiation scenario may be reconciled with diverse opinions on Inner IE branchings previously expressed by Indo-Europeanists.

Keywords: historical semantics; Indo-European languages; Indo-European phylogeny; lexicostatistics; linguistic phylogeny; onomasiological reconstruction; semantic reconstruction

1 Introduction[1]

Indo-European (IE) is currently the biggest language family in the world in terms of geographical coverage and number of native speakers. It includes several hundred living languages and dozens of ancient languages (some extinct, some having direct descendants to the present day). The family has 12 subgroups, unanimously accepted by experts in the field and recognized by whatever method of formal analysis: Anatolian, Tocharian, Greek, Armenian, Albanian, Italic (Romance), Celtic, Germanic, Slavic, Baltic, Indic, Iranian (Kapović 2017).

Despite more than 150 years of IE phylogenetic studies – the pioneering IE tree was published by Schleicher (1861: 7) – the only consensus or near consensus, beyond the 12 aforementioned subgroups, reached among Indo-Europeanists concerns the outlier status of Anatolian and the existence of distinct Indo-Iranian and Balto-Slavic clades. Internal branchings that may have occurred between the Anatolian split-off and the formation of the aforementioned recent clades (Germanic, Albanian etc.) still remain a matter of debate among experts. The opinions are so controversial that the majority of Indo-Europeanists prefer not to discuss Inner IE branchings at all. In the present abstract, for the sake of convenience, we use the label "Nuclear IE" for the bulk of IE languages without the Anatolian outlier and "Inner IE" for IE languages without Anatolian and Tocharian (the term "Inner IE" was introduced by Jasanoff (2003) and later adopted by some other Indo-Europeanists, see Olander 2019 for nomenclature overview).

A number of formal phylogenies for the IE family have been published in the last decades. These are based either on lexical characters reflexed as a fixed list of semantic concepts (so-called lexicostatistics, e.g., Rexová et al. 2003; Gray and Atkinson 2003; Blažek 2007; Bouckaert et al. 2012; Chang et al. 2015; also Müller

1 Alexei S. Kassian, Mikhail Zhivlov, and George Starostin were responsible for the study design, and all coauthors contributed to the linguistic data elaboration. Alexei S. Kassian, who conducted the computational analysis, also prepared the manuscript with input from Mikhail Zhivlov and George Starostin.

et al. 2013), or on mixed – phonological, grammatical and lexical – datasets (Nakhleh et al. 2005; Ringe et al. 2002).

The resulting topologies and dates proposed in these studies contradict each other in many details. Some of the trees seem more appropriate from the point of view of traditional Indo-European linguistics, e.g., publications by Ringe's team. Others are less convincing; thus IE trees and datings in (Bouckaert et al. 2012; Gray and Atkinson 2003) are not compatible with expert views, based on extensive linguistic and interdisciplinary data (e.g., Anthony and Ringe 2015; Mallory 1989), in some important points; we believe that this is caused by use of inaccurate input data, see, e.g., the linguistic supplement in Kushniarevich et al. (2015) for some linguistic criticism and Pereltsvaig and Lewis (2015) for general critical assessment of linguistic and geographic data involved. An important additional shortcoming of previous studies on IE phylogeny is that the formal algorithms and computational methods have not been previously tested on any language groups with the gold tree standard, i.e., groups with a general consensus as to their phylogeny.

In view of the deficiencies of the previous studies, the goal of the presented research is to check whether we can obtain a refined IE language tree that would (1) not be in conflict with established historical facts and widely shared expert views, (2) be based on linguistic evidence of high quality, (3) be produced by applying innovative methods that have already been successfully tested on language groups with the gold tree standard.

2 Materials and methods

2.1 Data collection

Our analysis is based on the 110-item Swadesh wordlist as it is currently defined in the *Global Lexicostatistical Database* project (Starostin 2011). Since the subgroups (such as Slavic, Germanic, Albanian and so on) within the IE family are uncontroversial, for each subgroup we prefer to use a reconstructed proto-language wordlist (e.g., Proto-Germanic for the Germanic group) or, where available, a list for an attested language which can be roughly equated with a proto-language (e.g., Vedic for the Indo-Aryan group), although it must be noted that these two types of objects are not fully conceptually equivalent. For the role of a non-IE outlier (needed for maximum parsimony analysis), we optionally introduce Proto-Samoyed as a representative of the Uralic family (the leading candidate for the role of the closest relative of Indo-European in scholarly works that take a positive stance towards defining the external connections of Indo-European [Cowgill 1986: 13; Kortlandt 2010]). See Kassian et al. (2015a) for the extensive treatment of potential IE-Uralic

etymologies within the Swadesh list and particularly Kassian et al. (2015a: 323–325) for a discussion of the implausibility of contact-based explanation for the IE-Uralic lexical matches; additionally, all plausible IE-Samoyed Swadesh etymologies are listed in Section 2.3 below and explicitly discussed in the Supplement (https://doi.org/10.5281/zenodo.4046607). Since, however, the Indo-Uralic hypothesis is not universally accepted, the trees based on the dataset with Proto-Samoyed are only offered in Supplement as an optional solution (in fact discrepancies between the trees with and without Proto-Samoyed are minor and do not affect our results, see Sections 3.2 and 3.3).

Overall, the following wordlists are included in the study (see the Supplement for basic information on subgroups and discussion about dates) (Table 1)

Table 1: List of involved taxa and chronological constraints (see Supplement Table S2 for details).

Taxa	Constraints for Bayesian analysis	Strict dates for StarlingNJ analysis
Old Hittite	1650–1500 BC	1550 BC
Tocharian B	400–900 AD	650 AD
Ancient Attic Greek	375 BC	375 BC
Classical Armenian	400–500 AD	450 AD
Albanian	1950 AD	1950 AD
Archaic Latin	200 BC	200 BC
Old Irish	700–900 AD	800 AD
Proto-Brittonic	300–600 AD	450 AD
Proto-Germanic	500–300 BC	400 BC
Proto-Slavic	1–300 AD	100 AD
Proto-East Baltic	400–1 BC	200 BC
Old Indic (Atharvaveda)	1200–1000 BC	1100 BC
Proto-Iranian	1500–1000 BC	1300 BC
Proto-Samoyed	950–750 BC	800 BC
(Proto-Indo-European)	3500–8500 BC	–

We believe that the use of reconstructed wordlists for intermediate proto-languages instead of the more traditional approach that requires a great number of wordlists from modern languages is preferable for two reasons. First, it is proposed in Rama and Wichmann (2018) for lexically-based Bayesian inference that the number of cognate classes (and therefore characters, i.e., semantic concepts) needed for an adequate reconstruction of linguistic phylogeny is directly proportional to the number of taxa: the larger the set of languages, the greater the required number of semantic concepts. Rama and Wichmann's estimation is that a 100-item wordlist is enough for a set of 30 or less lects; for 31–100 lects, a 200-item wordlist would be needed, etc. The linguistic data that Rama and Wichmann's analysis is based upon vary in quality, making their results somewhat skewed, but

their main hypothesis about the direct relationship between the number of taxa and the number of characters is intuitively logical and seems quite correct.

The second reason is relevant for all phylogenetic methods, not only for the Bayesian one. Step-by-step reconstruction decreases the amount of homoplastic developments within the dataset, reducing the amount of noise in the data and making the whole model less complicated. An additional reason for noisy data are unreliable sources for some languages, almost inevitable in a situation when a lot of various lects with different degrees of documentation are involved. In some cases, Swadesh items reconstructed for a recent proto-language can be considered more reliable than the same Swadesh items reported for a poorly described daughter language.

The disadvantage of step-by-step reconstruction is that it is possible to incorrectly reconstruct a certain feature for a proto-language, e.g., assign the Swadesh item status to a certain proto-word even if historically this word was not really the main expression for the given semantic concept in the proto-language. Nevertheless, we do not find the risk of reconstruction errors quite high, because, (1) we adhere to the strict methodology of onomasiological reconstruction (for which see this section below), (2) the proto-languages in question are not very deep chronologically: typically, we deal with the distance of 2000–2500 years BP.

For the attested languages, the 110-item Swadesh wordlists were collected according to explicit semantic specifications (Kassian et al. 2010) using the most authoritative lexicographic sources and checked, if necessary against text corpora. Special attention has been paid to the procedure of data collection, including nuanced "first-hand" philological analysis of lexical semantics in texts, as a way to overcome inaccuracies that are found in previous research on IE phylogeny; the underlying hypothesis is that quality of input data, even more than the actual phylogenetic method involved, reflects upon the accuracy of the resulting tree (Kassian 2015).

All the lexical lists used are offered in the Supplement. Additionally, the majority of the wordlists are available online at the GLD project (http://starling.rinet.ru/new100/main.htm).

Onomasiological reconstruction for intermediate proto-languages of individual groups within the IE family is based on the relatively strict methodology proposed in Kassian et al. (2015a: 304–306) and in Starostin (2016). Its main principles are: tree topology, external etymology, internal derivability, typology of semantic shifts, areal effect exclusion. We accompany each proto-form and reconstructed meaning with detailed comments which explain our choice (see the Supplement).

One should keep in mind that any number of characters, be it a 40-, 100-, 200- or 1,000-item set, is essentially arbitrary. The sometimes-discussed insufficiency of the 110-item wordlist is not in itself an obstacle to the application of the kinds of

quantitative methods employed in our study. Furthermore, expansion of the wordlist to 200 items or more runs into the obvious problem that, as items with less and less historical stability and more and more semantic vagueness come into consideration, it becomes progressively more difficult to select a basic word for documented languages and justify the optimal onomasiological reconstructions for intermediate protolanguages that are crucial for our methodology.

2.2 Matrix construction

For the present study we compare three sequentially applied methods of cognacy marking:

- Stage-1. High-quality dataset with traditional root cognacy (e.g., English *wind* is a match of Russian *veter* 'id.', these forms eventually representing separate derivatives from the same verbal proto-root 'to blow'). This dataset is outside the focus of the present study, it is treated only in Supplement, being used as a reference point for some conclusions.
- Stage-2. Dataset without derivational drift. This is the Stage-1 root cognacy dataset, modified so that forms showing the so-called derivational drift are marked as unrelated (English *wind* ≠ Russian *veter*); see this section below on details and on the formal procedure of derivational drift detection. This is our basic input dataset (Figure 1).
- Stage-3. Homoplasy-optimized dataset. This is the Stage-2 derivational drift-free dataset, in which cognates that violate the tree topology are marked as unrelated (not only *wind* ≠ *veter*, but also Old Indic *agni* 'fire' ≠ Latin *ignis* 'id.'), see Sections 2.4 and 3.3 below and Kassian (2017) for details. This is the dataset upon which the final IE tree is built (Figure 2).

At each stage two datasets are elaborated: proper IE (Indo-European wordlists only) and IE-Samoyed (with the Proto-Samoyed wordlist introduced). Thus, we have six datasets in total.

Compilation of the lexical matrix is a standard procedure: for a single Swadesh concept, etymologically cognate forms from different languages are marked with the same index, i.e., included in the same cognate class (see, e.g., Starostin 2007a [1995]; Atkinson and Gray 2006: 93–94). However, our approach has two features requiring special comments.

The first feature of our procedure is that we mark loanwords as *lacunae*, not as singletons (forms with unique cognate index) – according to a principle standardized in the *Moscow School*, where it is assumed that lexical replacement by means of borrowing is a fundamentally different process from internal replacement. Penetration of loanwords introduces a disturbance factor into the

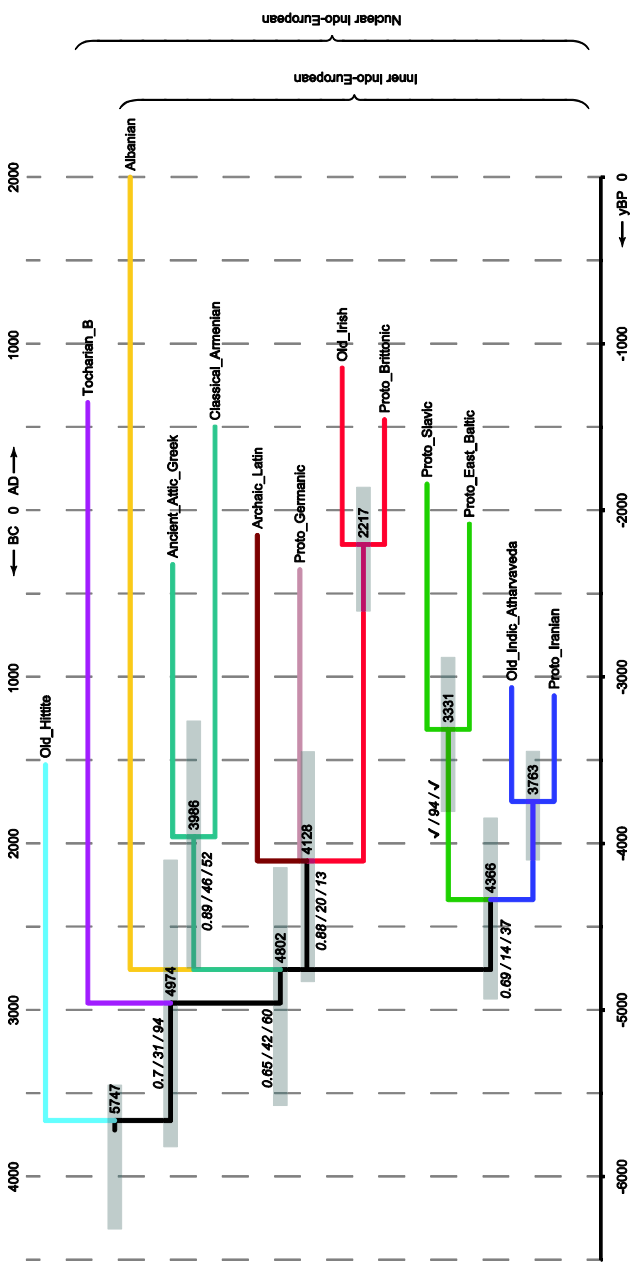

Figure 1: Manually constructed strict consensus tree of the IE family based on the Stage-2 derivational drift-free dataset (*wind* ≠ *veter, agni* = *ignis*). The tree summarizes three trees obtained by individual methods for the proper IE dataset (Figures S1d, S2c, S3c). Dates are obtained by the Bayesian MCMC analysis (Figure S2c): gray bars represent the 95% highest probability density (HPD) for the divergence times; mean divergence times are given to the right of each node. Scale values represent years before present (yBP). Statistical support values are shown in italic near the branches in the following sequence: Bayesian MCMC/StarlingNJ / MP ("✓" means that $p \geq 0.95$ in an individual method; not shown for nodes with $p \geq 0.95$ in all methods). Traditional subgroups are identified by color branches.

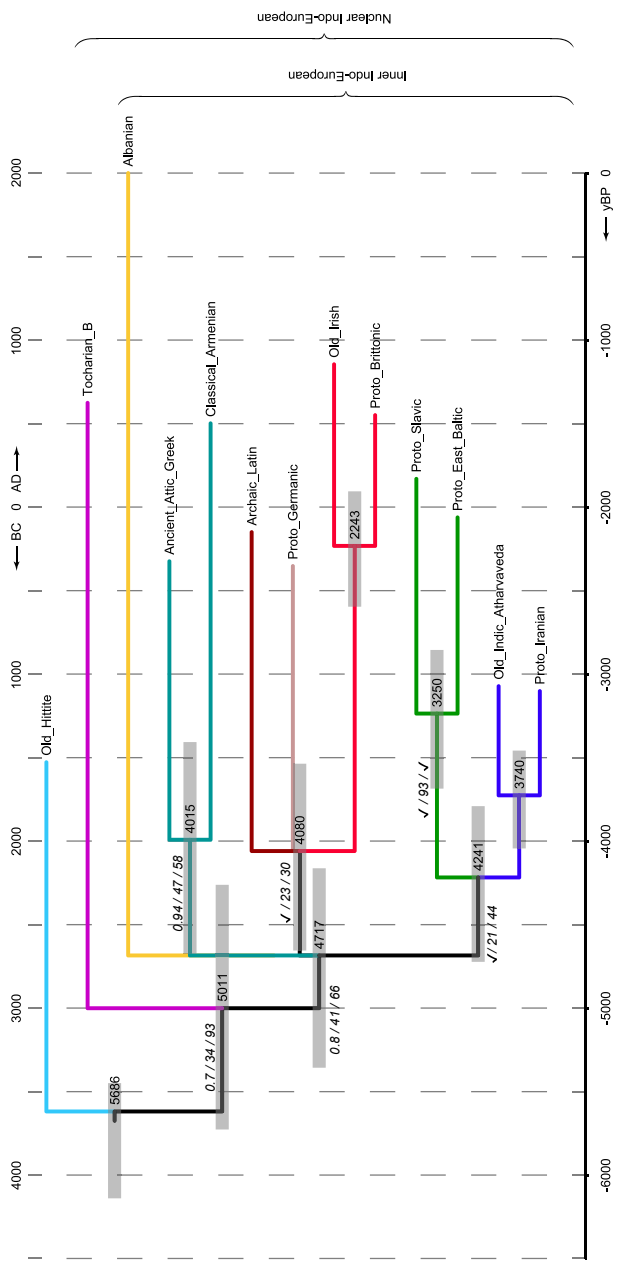

Figure 2: Manually constructed strict consensus tree of the IE family based on the Stage-3 homoplasy-optimized dataset (*wind* ≠ *veter*, *agni* ≠ *ignis*). The tree summarizes three trees obtained by individual methods for the proper IE dataset (Figures S1f, S2e and S3e). Dates are obtained by the Bayesian MCMC analysis (Figure S2e): gray bars represent the 95% highest probability density (HPD) for the divergence times; mean divergence times are given to the right of each node. Scale values represent years before present (yBP). Statistical support values are shown in italic near the branches in the following sequence: Bayesian MCMC / StarlingNJ / MP ("✓" means that p ≥ 0.95 in an individual method; not shown for nodes with p ≥ 0.95 in all methods). Traditional subgroups are identified by color branches.

ideal process of lexical evolution, since the number of lexical loans can be drastically affected by unpredictable social circumstances (Starostin 2007b [1989], 2000, 2013: 135; Kassian 2017: 224). As proposed in Starostin (2013: 135) and Kassian (2017: 225), exceptions when we should treat loanwords as singletons are situations when there is evidence that the word was borrowed into the target language with a non-Swadesh meaning, persisted as such for a certain period of time and later acquired the Swadesh meaning due to natural semantic and morphological development. Examples where it is reasonable to treat loanwords or words with borrowed roots as singletons are Modern German *Kopf* 'head' < Old High German *kopf* 'mug, bowl' < Latin *cupa, cuppa* 'cask, bowl' or the Romance word for 'liver' (Italian *fegato*, French *foie*, etc.) < Vulgar Latin **fecatum* 'fig-stuffed liver (a dish)', derived from Latin *fīcus* 'fig' < substrate Mediterranean term for 'fig'.

In our current dataset, the Albanian wordlist is the one most corroded by loans. Provisionally we treat such cases as Albanian *flokë* '(head) hair' < Vulgar Latin *floccus* 'lock, flock' or Albanian *kripë* 'salt' < Bulgarian *krupa* 'lump of salt' as loans, marking them with "?" in the matrix. The reason is that, first, these items could have penetrated into Albanian already with the Swadesh meanings, cf. attendant semantic shifts when a word is transferred from one language to another: Estonian *hunt* 'wolf' < Low German *hunt* 'dog' or Middle Welsh *ofydd* '(love) poet, littérateur; lover, sweetheart, darling; master, champion' < Latin *(Publius) Ovidius (Naso)*. Second, when dealing with numerous Vulgar Latin and Romance loans in Albanian, we do not know exactly which lect was the donor; it might be an undocumented language or dialect where the observed shifts ('lock, flock' > 'hair' etc.) had already taken place. In contrast, Albanian *kokë* 'head; bulb; berry; grain' (<Latin *coccum* 'berry') is marked as a singleton, since there is evidence that the semantic shift 'berry' > 'head' took place already on Albanian ground.

The problem of loanwords has recently been discussed by Chang et al. (2015: 212) who perform two kinds of phylogenetic analysis for the IE family. For the first one they treat all loanwords as singletons. For the second one they include loanwords as full-fledged cognates with their lexical sources. Chang et al. (2015: 205) suggest that loan exclusion cannot be sufficiently motivated for the purposes of chronological analysis. Their reasons are as follows: recent loanwords can at first function as marginal terms and only gradually acquire the status of a basic expression for the given meaning, thus reproducing the evolution of inherited words (the only example cited is the French loanword *animal* which became a basic term after several hundred years of its usage in English). Furthermore, detection of loanwords is more difficult for ancient languages than for modern languages, since in the former case the donor languages might lack proper historical documentation, remaining undetectable for our studies.

Neither of Chang et al.'s arguments seem fully convincing to us. In most cases where one can historically trace loanwords that occupy the status of basic terms under pressure of the dominant donor language, this process occurs over a very limited time period. Such cases do not reflect standard patterns of language-internal lexical evolution and thus upset any chronological estimations. This type of lexical replacement is especially characteristic of situations when the migrating word already had Swadesh status in the donor language. An example is the French loanword *mountain*, which appears in Middle English for the first time *ca.* 1200 AD and is already documented as a basic term in Chaucer's works or even earlier. As for Chang et al.'s example *animal*, this would be treated as a singleton under our approach as well, since Modern English *animal* technically represents the same case of internal semantic evolution as German *Kopf* or Italian *fegato* mentioned above.

On the other hand, hidden loanwords are a serious problem not only for ancient languages but for many modern languages as well, especially if these languages are spoken in a territory whose (socio)linguistic conditions over the last several centuries have not been properly ascertained. Nevertheless, even in some cases when the donor language is unknown, we are still able to detect loanwords due to their specific phonological or morphological traits. However it does not seem preferable, to intentionally avoid minimizing non-uniform signal distortion in one part of the dataset if such distortion remains technically inevitable for another portion of the dataset.

The second feature of our procedure (one of the novelties of the present study) is that forms with the so-called derivational drift are not marked as cognates in the input matrix. The traditional and almost universally accepted approach is to treat as lexicostatistical matches those forms whose main meaningful morphemes (*scil.* roots) are cognate to each other, i.e., are thought to go back to a single proto-root. Nevertheless, the idea that, based on morphological grounds, true historical cognates may be distinguished from parallel new formations that share the same root is evident; see the insightful discussion in Chang et al. (2015: 202–203) where the phenomenon of parallel morphological derivation is called "derivational drift". Regrettably, despite having stated the problem, Chang et al. (2015) do not make attempts to propose criteria for derivational drift detection and to implement them into the model. We propose two formal criteria for derivational drift as discussed below in this section.

Upon first sight, any two stems (from different lects) whose roots are cognate but whose affixal structure is different are to be treated as resulting from parallel evolution, since these stems do not originate from a common proto-stem. Nevertheless, such a strict criterion should rule out lots of cases where the stems involved have to be treated as true cognates under a common-sense approach. As an example, we may discuss the item 'heart' from our dataset. The following stems denoting 'heart'

in Indo-European languages go back to the same PIE root (words are quoted in nom. sg.): Hittite *kir* (<Proto-Anatolian **ḱer* < **ḱerd*, a root noun), Ancient Greek *kard-í-a:* (zero grade form extended with the *i*-suffix), Proto-Germanic **xert-o:n* (full grade form extended with the *on*-suffix). The most likely scenario, accepted by experts in the field, is that the original Proto-Indo-European root noun nom.-acc. **ḱe:r* (<**ḱerd*) / obl. **ḱr̥d-* (the paradigm is retained in Anatolian) was subsequently modified in individual subgroups by adding various suffixes in order to make the paradigm more transparent (Wodtko et al. 2008: 417–423). It is impossible not to regard the pair Hittite *kir* / Greek *kard-í-a:* as historically true cognates. The same is true for the pair Hittite *kir* / Germanic **xert-o:n* which is also to be marked as reflecting true cognacy. In the absence of the Proto-Anatolian root noun, the third pair, Greek *kard-í-a:* / Germanic **xert-o:n*, would indeed look suspicious, so that one might propose to analyze *kard-í-a:* and **xert-o:n* as independent homoplastic derivatives from a certain root **ḱerd-* with a different (unknown) meaning, but data such as Anatolian shows that this solution would be incorrect.

Consequently, more sophisticated algorithms are needed to detect derivational drift, i.e., cases when two lexemes from different lects possess the same meaning and share the same root, but actually represent parallel evolutionary events from the lexicostatistical point of view. For the present analysis, we propose two formal criteria capable of uncovering a substantial number of cases of derivational drift.

(1) First criterion of derivational drift: if two stems from compared lects share the same root, but differ in their affixal structure, and there is evidence that at least one of the stems has undergone a part of speech change (e.g., noun ↔ verb), these stems most likely represent a homoplastic development.

Here are some examples. In some IE lects, adjectival terms for 'warm' are derived from the verb **tep-* 'to be warm/hot (*vel sim.*)': Old Irish *teë*, Proto-Brittonic **te:m:*, Proto-Slavic **tep-l-*. In each case, however, the part of speech change "verb → adjective" has occurred with different suffixes used, namely **-nt-* (Old Irish), **-smo-* (Brittonic), **-lo-* (Slavic). The three forms in question most likely represent full-fledged lexicostatistical replacements, being the result of parallel word formation. Another example is the verb 'to die' in Brittonic languages. The most common verb for 'to die' attested in Inner Indo-European languages is **mer-*: Latin *mor-*, Old Indic *mar-*, Proto-Slavic **mer-* and so on. For Proto-Brittonic, one can reconstruct the verb **marw-* 'to die' which represents a denominative formation from the Proto-Brittonic adjective **marw* 'dead'. The latter originates from **mr̥-wo-*, ultimately containing the same root **mer-*, but further modified with an adjectival suffix. The Brittonic verb has undergone the part of speech change "verb → adjective → verb" and it is intuitively likely that shifts such as 'dead' → 'to die', involving changes in parts of speech,

represent a type of lexicostatistical event that is substantively different from cases of morphological extension such as 'heart' above.

Although the described criterion would be inapplicable to languages in which the morphological opposition between different parts of speech is weak or non-existent, for languages that display such an opposition, including Indo-European, the proposed criterion could affect a significant number of input forms. Crosslinguistically the most basic opposition is "noun: verb", whereas adjectives tend to be either noun-like or verb-like. Although Proto-Indo-European adjectives are noun-like, they are sufficiently different from nouns to be considered a separate word class: PIE adjectives are gradable, and they already in Proto-Indo-European exhibited a complicated system of suffix substitution known as "the Caland system" (Lundquist and Yates 2018; Nussbaum 1976: 2113–2118).

(2) Second criterion of derivational drift: if two stems from compared lects share the same root, but are modified with differing affixes, and there is evidence that these stems have been derived from a simpler stem whose semantics was quite different from the meanings of the stems compared, these two stems most likely represent homoplastic development.

For example, in Latin, Baltic and Celtic, expressions for 'person' are derived from the Indo-European term for 'earth' (i.e., 'person' as 'earthling'), but with different suffixes: *-*on* in Latin (*hom-in-*) & Proto-Baltic (**žm-un-*) and *-*yo-* in Proto-Celtic (**gdon-yo-*). The Latin and Baltic forms, on the one hand, and the Celtic form, on the other hand, most likely represent two distinct lexicostatistical replacements, being the result of parallel word formation. A particular case of derivational drift to be noted here is valency-changing derivation, e.g., the meaning 'to kill' is frequently expressed by causatives from the verb 'to die' in the world's languages. If these causative stems represent etymologically different morphological patterns, it seems reasonable to treat such verbs 'to kill' as parallel, i.e., homoplastic formations.

For the present study we apply both criteria to the input dataset to make it free from reliably identifiable cases of derivational drift, i.e., we assign distinct cognate classes for the cases affected by two aforementioned criteria.

From the technical point of view, the original lexical dataset represents a matrix which is multistate, i.e., characters may have more than two states, and which includes synonyms. Although one has to aim to minimize the amount of synonyms in the input dataset, in practice synonymy is inevitable for lexical data, this implies the possibility of having more than one word for a single Swadesh concept. This multistate matrix is used in the Starling package (Starostin 2007c [1993]; Burlak and Starostin 2005: 271–274), the only phylogeny-inferring software able to process input matrices containing synonyms (Kassian 2017: 219–220).

The second matrix that is used for other phylogenetic packages is binary. The binary matrix is converted from the multistate dataset by coding for the presence ("1") or absence ("0") of the specific proto-root with the given Swadesh meaning in the language in question (Atkinson and Gray 2006; Gray and Atkinson 2003). Swadesh items that are not documented for the given language or superseded by loanwords are marked as "?".

2.3 Rooting the trees

Manual rooting is needed at least for one of the quantitative methods adopted in the present study: the maximum parsimony analysis. The choice of an outgroup taxon is a non-trivial issue in our case, since the Indo-European family lacks any linguistic relatives that would be close enough to be conventionally accepted and represent the most appropriate outgroup. In light of this, we prefer to duplicate our analysis by using two different outliers: Hittite (Indo-European family) and Proto-Samoyed (one of two primary branches of the Uralic family).

For the first analysis, Hittite was taken as an outgroup (as a constraint it was used only for maximum parsimony trees). Hittite belongs to the Anatolian group of the Indo-European family, and there is a near consensus that Anatolian is the first outlier in the family (Anthony and Ringe 2015; Gamkrelidze and Ivanov 1995; Winter 1996).

For the second analysis, we introduce Proto-Samoyed as an outgroup (as a constraint it was likewise used only for maximum parsimony trees). The Samoyed group (Hajdú 1988) consists of several closely related languages and represents one of the two principal branches within the Uralic linguistic family. Among known languages and groups, Proto-Uralic seems to be the closest sister taxon for Proto-Indo-European, although this relationship is relatively distant, comparable to that of modern IE languages (say, as between Modern German and Modern Greek). Despite the fact that the Indo-Uralic hypothesis has a long history (see Kassian et al. 2015a, 2015b for an overview and for statistical evidence in favor of the IE-Uralic clade), it is far from being commonly accepted. Nevertheless, the general trend among Indo-Europeanists is positive, cf., e.g., the recent conference "The Precursors of Proto-Indo-European: The Indo-Hittite and Indo-Uralic Hypotheses", held at Leiden University, 9–11 July 2015 (Zhivlov and Zhivlova 2015).

We consider the following 10 Proto-Samoyed items to be etymologically cognate to their Proto-Indo-European counterparts: **ḛ-r-* 'to drink', **tå-* 'to give', **mə-n* 'I', **kʲy-tV-* 'to lie', **nim* 'name', **ta-* 'that', **tə-* 'this', **tə-n* 'thou', **wet* 'water', **ke-* 'who'. Eight of them are treated in detail in Kassian et al. 2015a, 2015b. The ninth root, **tå-* 'to give' (<Proto-Uralic **toyi-*) is an obvious comparandum for IE **do:-* or

*deh_3- ‘to give’ with *-ɣ- as a counterpart of the IE “laryngeal”. The 10th item *kɨy-tV- ‘to lie’ (<Proto-Uralic *kuyi-) can be safely compared with IE *ḱey- ‘to lie’.

2.4 Tree building

In the present study we generally follow the workflow proposed in Kassian (2017: 255–257) and successfully tested on three language groups with the gold topological standard (Lezgian, Tsezic and Balto-Slavic). The workflow developed for the present study consists of several sequential steps:

(i) Once the high-quality lexical dataset is compiled (Stage-1, *wind* = *veter*) and derivational drift is eliminated (Stage-2, *wind* ≠ *veter*), the trees are built by means of several phylogenetic algorithms (Starling neighbor-joining, Bayesian MCMC, Maximum Parsimony).

(ii) A strict consensus tree is compiled, i.e., a summarizing tree which only contains the binary nodes that are present in all input trees, whereas any topological disagreement between input trees results as a multifurcation in the strict consensus tree.

(iii) Proceeding from the topology of the resulting consensus tree and reconstructed ancestral character states, we examine the input dataset for homoplastic developments, i.e., search for characters that are incompatible with the topology of the consensus tree. Lexical items identified as constituting such homoplastic matches are marked as etymologically unrelated or, if we can identify the direction of the influence, the target item can be marked as a borrowing (that is, we do not discard incompatible characters from the matrix, but simply allocate them to different cognate classes). This procedure is called homoplastic optimization (Stage-3, *agni* ≠ *ignis*). Note that homoplastic optimization requires not only a predefined tree, but also onomasiological reconstruction of Swadesh words for Proto-Indo-European and intermediate protolanguages. For the onomasiological reconstruction we use the methodology proposed in Kassian et al. (2015a: 304–306). See Supplement for linguistic notes on individual cases of homoplasy in our dataset.

(iv) Homoplasy-optimized trees are rebuilt by means of previously listed individual algorithms (StarlingNJ, Bayesian MCMC, MP).

(v) Obtained homoplasy-optimized trees are summarized as a strict consensus tree which constitutes the final result of our study.

Lexicostatistical trees were produced by means of the following phylogenetic methods:

– Starling neighbor-joining, hence StarlingNJ (Burlak and Starostin 2005; Kassian 2015). The StarlingNJ trees were produced in the Starling software v.2.5.3 (Starostin 2007c [1993]; Burlak and Starostin 2005: 271–274) from the lexicostatistical database which represents a multistate matrix with synonymy allowed. The allowed synonymy means that when the same Swadesh slot is occupied by more than one word in a given language, i.e., by several synonyms, each word from this slot is compared to each word from the same slot in another language, so that all possible pairs of words between two languages are compared: if there is at least one matching pair, the whole slot is treated as a match. For node dating, the so-called "experimental method" was applied, according to which every Swadesh item possesses an individual relative index of stability (Starostin 2007d, 2010). The non-parametric bootstrap test was performed (10,000 pseudoreplicates). The hierarchical agglomerative clustering produces a rooted tree by definition (the last merger is the root; it coincides with the midpoint under the assumption of a nearly uniform replacement rate). The dates of the nodes were established by strict clock, see Starostin (2007b [1989], 2000), Blažek and Novotná (2007) and Balanovsky et al. (2011) on scale calibration and for further details. Neighboring nodes are joined in a single node if the temporal distance between them is 300 years or less (300 years corresponds to the mutation of ca. 1.5 words in a lect – a reasonable calculation error, although this temporal interval is essentially arbitrary at the current stage of research.
– Markov chain Monte Carlo simulation under a Bayesian framework (hence Bayesian MCMC), see Makarenkov et al. (2006: 68–69). The trees were produced in the MrBayes software v.3.2.6 (Ronquist et al. 2012) from the binary matrix described above. We used the covarion F81 model with datatype = restriction, coding = noabsencesites, rates = gamma, covarion = yes, brlenspr = clock:fossilization, clockvarpr = TK02 (autocorrelated relaxed clock), see Supplement for the full set of MrBayes options and further Yanovich (2020) for an ample discussion on MrBayes parameters that are suitable for linguistic phylogeny. For the proper IE dataset, the range of the root age is strictly predefined as 10,500–5,500 yBP. For the IE-Samoyed dataset, the range of the root age is predefined as offsetexp(10,000, 20,000) with the upper limit 10,000 yBP and mean 20,000 yBP, see Supplement for the discussion on dating. No outgroup or other topological constraints were set. The program was run 4 times using four concurrent Markov chains. Each run produced 10,000,000 tree generations with samples taken every 500 generations. For each run, the first 25% tree generations were discarded as a burn-in. The dated consensus trees were rooted by the program. The trees were visualized in the FigTree software (v.1.4.3).

– Unweighted maximum parsimony method (hence MP), see Makarenkov et al. (2006: 66–67). The trees were produced in the TNT software (Willi Hennig Society edition of TNT, v.1.5, March 2017, see Goloboff and Catalano 2016) from the binary matrix described above by the branch-and-bound ("Implicit enumeration") algorithm. Obligatory binarization of nodes was prohibited ("Collapse trees after the search"). Hittite and Proto-Samoyed were marked as an outgroup for the proper IE and IE-Samoyed datasets respectively. After a set of optimal trees of equal cost (plus suboptimal trees with step 1, if a single optimal tree is returned) was obtained, the majority-rule consensus tree was produced for which the non-parametric bootstrap test was performed (1,000 pseudoreplicates). The trees are not dated. The trees were visualized in the FigTree software (v.1.4.3).

3 Results

3.1 Stage-1 dataset (root cognacy)

For each dataset – root cognacy wordlists (Stage 1), derivational drift-free wordlists (Stage 2), homoplasy-optimized wordlists (Stage 3) – the following trees are obtained with and without Proto-Samoyed taxon:

– Figure S1 (see Supplement), StarlingNJ method with neighboring nodes joined.
– Figure S2 (see Supplement), Bayesian MCMC method.
– Figure S3 (see Supplement), maximum parsimony method.

The trees based on the root cognacy wordlists (*wind* = *veter*; Figures S1a, S1b, S2a, S2b, S3a and S3b) do not significantly contradict our expectations. Anatolian and Tocharian are correctly detected as sequential outliers. Recent generally accepted clades – (Insular) Celtic, Balto-Slavic, Indo-Iranian and additionally Greek-Armenian – are correctly detected in most trees. Nevertheless, the overall result cannot be considered totally satisfactory, since the trees conflict with each other in some important nodes. Such discrepancies exist not only between trees obtained by different methods, but also between trees obtained by the same method on proper IE and IE-Samoyed datasets.

3.2 Stage-2 dataset (derivational drift-free)

The results of the analysis of the derivational drift-free dataset (*wind* ≠ *veter*, *agni* = *ignis*) are more promising. The StarlingNJ (Figure S1c–S1d) and Bayesian

(Figure S2c–S2d) trees are almost identical topologically. Both methods suggest two sequential outliers (Anatolian and Tocharian) and a four-way multifurcation of Inner IE regardless of whether the proper IE dataset or the IE-Samoyed one is used: (1) Greek-Armenian, (2) Albanian, (3) Italic-Germanic-Celtic, (4) Balto-Slavic–Indo-Iranian. The difference between the StarlingNJ and Bayesian methods is that the StarlingNJ tree resolves the West European clade as [[Italic, Germanic] Celtic], whereas the Bayesian tree shows a three-way split [Italic, Germanic, Celtic].

The majority-rule consensus MP trees (Figure S3c and S3d) generally produce the same topology, although they contain more binary nodes than trees provided by other methods. Note that the MP tree based on the proper IE dataset suggests the same detailed structure for the West European clade: [[Italic, Germanic] Celtic].

The only substantial discrepancy between the methods concerns the Italic, Germanic and Celtic groups on the MP tree based on the IE-Samoyed dataset (Figure S3d): these groups do not form a distinct clade (first Celtic and then Italic-Germanic sequentially split from the trunk). This result is not entirely clear.

Two methods produce dated trees: StarlingNJ (strict dates) and Bayesian MCMC (95% highest probability density for the divergence times and the mean dates of divergence). The difference is summarized in Table 2. For the Bayesian method we refer to the tree based on the proper IE dataset (its dates differ only slightly from those of the IE-Samoyed tree).

As follows from Table 2, the StarlingNJ dates are substantially deeper than the Bayesian ones for the Stage-2 derivational drift-free dataset. Moreover, the currently obtained StarlingNJ dates are deeper than StarlingNJ dates reported for the Indo-European family in some previous studies. E.g., the Anatolian split-off is dated back to 4340 BC, the Tocharian split-off is dated back to 3870 BC in Kassian (2009: 424) (based on 50-item wordlists for reconstructed protolanguages of IE subgroups).

Table 2: Discrepancies in dates obtained for the Stage-2 derivational drift-free dataset (*wind ≠ veter, agni = ignis*). 95% HPD and mean for Bayesian MCMC, strict dates for StarlingNJ. See Figure 1 for the tree representation.

	Bayesian MCMC (Figure S2c)	**StarlingNJ (Figure S1c, S1d)**
Anatolian split-off (root)	4314–3450 BC (mean 3747 BC)	5080 BC
Tocharian split-off	3821–2099 BC (mean 2974 BC)	4700 BC
Inner IE break-up	3572–2145 BC (mean 2802 BC)	4100 BC
Greek-Armenian break-up	2747–1264 BC (mean 1986 BC)	3460 BC
Italic-Germanic-Celtic break-up	2825–1443 BC (mean 2128 BC)	3500 BC
Insular Celtic break-up	605 BC – 138 AD (mean 217 BC)	1500 BC
Balto-Slavic–Indo-Iranian break-up	2933–1847 BC (mean 2366 BC)	3570 BC
Balto-Slavic break-up	1807–882 BC (mean 1331 BC)	2390 BC
Indo-Iranian break-up	2100–1447 BC (mean 1763 BC)	2230 BC

Alternatively, in Blažek (2007: 85) these bifurcations are dated back to 4670 BC and 3810 BC respectively (based on 110-item wordlists of attested languages). The chronological discrepancies between our current Starling calculations and the previously reported ones are due to differences in input dataset preparation. The strict semantic specifications and some general principles of Swadesh wordlist compilation were added to our arsenal only in 2010 (Kassian et al. 2010) and homoplasy optimization techniques (such as derivational drift elimination etc.) were introduced just recently. By contrast, the dating algorithm implemented in Starling was calibrated much earlier on the basis of "traditional" Swadesh wordlists, i.e., wordlists compiled without such a strict semantic and pragmatic control, not to mention homoplasy optimization.

In turn, the obtained Bayesian dates show an opposite tendency: some of them seem too recent. Indeed the 95% HPD ranges quoted in Table 2 generally do not contradict our expectations, but at least some of the mean dates are somewhat younger than accepted by experts in the field, e.g., the Indo-Iranian break-up is usually dated before 2000 BC (Mallory 1989: 38–39; Kulikov 2017: 205) as opposed to 1753 BC (a mean date) suggested by our analysis. The mean date of the Insular Celtic break-up is 221 BC, although the toponym Πρεττανική 'Britain' with the specific Proto-Brittonic development $*k^w > p$ was reported by Pytheas of Massalia as early as 325 BC (*apud* Strabo) (Koch 2006). It should be noted that linguists should avoid regarding a Bayesian mean (or median) dates as absolute, since only 95% HPD intervals make historical sense.

All these issues concerning dating techniques require detailed investigation that must be a matter of further research. For the consensus tree of the current study (Figure 1) we accept Bayesian dates as being more flexible.

For the proper IE derivational drift-free dataset (Stage-2; *wind* ≠ *veter*, *agni* = *ignis*), we summarize the resulting StarlingNJ (Figure S1c, S1d), Bayesian MCMC (Figure S2c) and MP (Figure S3c) topologies as a strict consensus tree (Figure 1). The trees obtained for the IE-Samoyed dataset (Figures S2d, S3d) are the same with the exception of the paraphyletic topology of Italic, Germanic and Celtic subgroups in the MP tree (Figure S3d); note that the MP tree for the proper IE dataset (Figure S3c) avoids this shortcoming being in accordance with the StarlingNJ and Bayesian MCMC trees. In fact, the consensus Stage-2 tree (Figure 1) is identical to the Bayesian tree Figure S2c.

3.3 Stage-3 dataset (homoplasy-optimized)

The next step in the input dataset preparation is homoplastic optimization (Kassian 2017). We label it as Stage 3 (*agni* ≠ *ignis*). Proceeding from the consensus

Stage 2 tree (Figure 1) and our methodology of onomasiological reconstruction (Kassian et al. 2015a: 304–306), we examine cases when two or more roots are in "criss-crossed" configuration, i.e., at least one of them violates the tree topology (so-called incompatible characters). If we have evidence that one of the competing roots should represent a retention, whereas the other one is likely a parallel innovation in separate subgroups, we mark the reflexes of the second root as unrelated. An example is Old Indic *agni* and Latin *ignis*, both 'fire', which are direct etymological cognates and thus marked with the same cognate index (*scil*. cognate class) in the Stage-2 dataset, but receive different indexes in the Stage 3 dataset. Whether **ŋ̊gnis* was a special poetic designation of fire or meant specifically 'flame' or something similar in the protolanguage, it probably must have become the basic everyday word for 'fire' independently in Old Indic and Latin.

The following Swadesh concepts are affected by homoplastic optimization in the Stage 3 dataset: 'belly', 'breast', 'to come', 'to eat', 'fire', 'to hear', 'to lie', 'man', 'many', 'night', 'person', 'to see', 'to sit', 'to stand' (see Supplement for linguistic comments). For other Swadesh concepts homoplasy cannot be resolved as the evidence is too scanty to allow a reliable onomasiological reconstruction: 'one', 'to see', 'that', 'this', 'tooth', 'far', 'snake', 'year', therefore those concepts remain the same as in the Stage 2 dataset. Finally there are several Swadesh concepts for which a Proto-IE suppletive paradigm is reconstructed that could be simplified independently in the same manner in different branches, resulting in a crisscrossing configuration: 'to come', 'to see', 'to go', 'I', 'we' (kept the same as in the Stage 2 dataset). As one can see, these represent either basic verbs or personal pronouns, i.e., categories for which suppletion is normal crosslinguistically. (See the Supplement for linguistic comments on each of the aforementioned cases of tree topology violation)

The trees based on the Stage 3 homoplasy-optimized dataset (*wind* ≠ *veter, agni* ≠ *ignis*) and obtained by individual methods are offered in Supplement as Figures S1e–S1f, S2e–S2f, S3e–S3f. The trees generally do not contradict each other (in particular the Italic-Germanic-Celtic clade is now observed in all trees) and are compatible with trees based on the Stage 2 dataset (*wind* ≠ *veter, agni* = *ignis*). The new result is that Albanian is now joined with Balto-Slavic–Indo-Iranian in a distinct clade in both Bayesian trees (Figure S2e, S2f). On one hand, this does not contradict traditional views, since we know too little about the history of Albanian. On the other hand, statistical support for the clade [Albanian [Balto-Slavic, Indo-Iranian]] is very weak: 0.54 and 0.59 depending on whether or not Proto-Samoyed is included in the dataset. The clade [Albanian [Balto-Slavic, Indo-Iranian]] is also observed in the StarlingNJ tree with binary nodes (Figure S1e), although the temporal span is too short and Albanian turns out to be a separate branch in the StarlingNJ tree when neighboring nodes are joined if the distance between them

is ≤ 300 years (Figure S1f). By contrast, Albanian is the first Inner IE outlier in the MP trees (Figure S3e, S3f).

Dates obtained for the Stage-3 homoplasy-optimized dataset are summarized as Table 3. They are similar to those for the Stage 2 dataset (Table 2); again, the StarlingNJ dates seem too early, whereas the upper limit and the mean of the Bayesian 95% HPD intervals seem too recent. One could hypothesize, however, that the dates in Table 3 could be biased due to incorrect dating of reconstructed proto-languages. We repeated the Bayesian analysis of the Stage-3 homoplasy-optimized dataset (*wind ≠ veter, agni ≠ ignis*) excluding all reconstructed taxa, so that only the following languages were involved: Hittite, Tocharian B, Ancient Greek, Classical Armenian, Albanian, Archaic Latin, Old Irish, Old Indic. The newly obtained Bayesian dates are compatible with those of Table 3 and Figure 2, although predictably the resulting time intervals became wider due to reduction of input data: Initial Proto-IE break-up 4511–3450 BC, quaternary Inner IE break-up 3728–1973 BC, Latin-Irish break-up 2703–1111 BC.

For the IE proper Stage-3 homoplasy-optimized dataset (*wind ≠ veter, agni ≠ ignis*), we summarize the resulting StarlingNJ (Figure S1e, S1f), Bayesian MCMC (Figure S2e) and MP (Figure S3e) topologies as a strict consensus tree Figure 2. We adopt Bayesian dates as more flexible ones. The trees obtained for the IE-Samoyed dataset (Figures S2f, S3f) are the same.

As one can see, some clades in Figure S2e (and correspondingly Figure 2) possess relatively low posterior probabilities obtained by the Bayesian analysis, meaning that there are alternative branchings for these taxa which cannot be ignored and should be discussed. The alternative branchings are visualized as the DensiTree plots: Figures S4–S6. The following cases must be noted:

Table 3: Discrepancies in dates obtained for the Stage-3 homoplasy-optimized dataset (*wind ≠ veter, agni ≠ ignis*). 95% HPD and mean for Bayesian MCMC, strict dates for StarlingNJ. See Figure 2 for the tree representation.

	Bayesian MCMC (Figure S2e)	**StarlingNJ (Figure S1e, S1f)**
Anatolian split-off (root)	4139–3450 BC (mean 3686 BC)	5110 BC
Tocharian split-off	3727–2262 BC (mean 3011 BC)	4710 BC
Inner IE break-up	3357–2162 BC (mean 2717 BC)	4150 BC
Greek-Armenian break-up	2676–1407 BC (mean 2015 BC)	3460 BC
Italic-Germanic-Celtic break-up	2655–1537 BC (mean 2080 BC)	3540 BC
Insular Celtic break-up	596 BC – 95 AD (mean 243 BC)	1570 BC
Balto-Slavic–Indo-Iranian break-up	2723–1790 BC (mean 2241 BC)	3570 BC
Balto-Slavic break-up	1686–855 BC (mean 1250 BC)	2450 BC
Indo-Iranian break-up	2044–1458 BC (mean 1740 BC)	2230 BC

(i) Nuclear IE (all without Hittite), $p = 0.7$. As follows from the MrBayes table of bipartitions (Table S1), the main competitive hypothesis is that Hittite and Tocharian form a distinct clade, its posterior probability is relatively high: $p = 0.25$. It is not entirely clear how to interpret such a result, it can hardly be explained by the long branch attraction effect.

(ii) Inner IE (all without Hittite and Tocharian), $p = 0.8$. The main competitive hypothesis (Table S1) is that Greek-Armenian and Tocharian form a distinct clade with $p = 0.15$.

(iii) The Balto-Slavic–Indo-Iranian clade has a low probability $p = 0.76$ in the Bayesian tree Figure S2e due to the unstable position of Albanian which can be inserted within this clade in some sampled trees available in the MrBayes output files *.runX.t: either [[Balto-Slavic, Albanian], Indo-Iranian] $p = 0.18$ or [Balto-Slavic, [Indo-Iranian, Albanian]] $p = 0.06$. Since Albanian is raised to a higher level in the manual consensus tree Figure 2, the sum of the aforementioned probabilities can be taken as the cumulative probability of the Balto-Slavic–Indo-Iranian clade with or without Albanian: $p = 1$.

(iv) Multifurcation of the Inner IE clade into four (Figure S2c) or three (Figure S2e) main branches means that none of the bifurcations between these branches is revealed in at least a half of the sampled trees. If we exclude Albanian as an unstable "roaming" taxon, the competitive binary clades between Greek-Armenian, Italic-Germanic-Celtic and Balto-Slavic–Indo-Iranian (any clade can be with or without Albanian) are: (1) [Greek-Armenian] + [Italic-Germanic-Celtic], $p = 0.47$; (2) [Italic-Germanic-Celtic] + [Balto-Slavic–Indo-Iranian], $p = 0.31$; (3) [Greek-Armenian] + [Balto-Slavic–Indo-Iranian], $p = 0.09$.

(v) In the case of hard, i.e., historically true polytomy, it is expected that dates of competitive bifurcations are identical or close to each other. As visualized by the DensiTree plots (Figures S4–S6), MrBayes time intervals for three competitive Inner IE bifurcations – (1) [Greek-Armenian] + [Italic-Germanic-Celtic]; (2) [Italic-Germanic-Celtic] + [Balto-Slavic–Indo-Iranian]; (3) [Greek-Armenian] + [Balto-Slavic–Indo-Iranian] – overlap each other to a great extent ("roaming" Albanian is disregarded). The same is true for the dated StarlingNJ trees (Figure S1c, S1e): the time gap between the Greek-Armenian split-off and the subsequent Italic-Germanic-Celtic split-off is *ca.* 50 years. This suggests that we are dealing with historically real multifurcation of the Inner IE clade.

4 Discussion

We evaluate three datasets that are sequential transformations of each other, using no topological constraints for the analysis.

- Trees obtained for the Stage 1 dataset with traditional root cognacy (*wind = veter, agni = ignis*; Figures S1a–S1b, S2a–S2b, S3a–S3b) do not have serious conflicts with traditional views (except for the Albanian–Balto-Slavic–Indo-Iranian clade in Figure S2a–S2b).
- However, trees obtained for the Stage-2 derivational drift-free dataset (*wind ≠ veter, agni = ignis*; Figure S1c–S1d, S2c–S2d, S3c–S3d; consensus tree Figure 1) fit established expert views even better. This suggests that the formal procedure of derivational drift elimination proposed above is a powerful and important technique which helps to improve the resulting phylogeny.
- Trees obtained for the Stage-3 homoplasy-optimized dataset (*wind ≠ veter, agni ≠ ignis*; Figure S1e–S1f, S2e–S2f, S3e–S3f; consensus tree Figure 2) add little to previously generated Stage 2 trees, but, in accordance with theoretical expectations, homoplastic optimization makes the resulting topology more robust.

All methods produce similar topologies (Figure S1–S13) regardless of whether the Proto-Samoyed list is included or not. First, Anatolian and Tocharian are always recognized as two sequential outliers, which means that the Inner Indo-European languages form a distinct clade. Second, all recent subgroups are correctly recognized as distinct clades: Greek-Armenian (missing from one of the MP-trees), Irish-Brittonic (i.e., Insular Celtic), Balto-Slavic, Indo-Iranian. All these features ideally fit traditional expert views on the IE family (although it must be noted that some scholars currently find available linguistic evidence for a Greek-Armenian clade insufficient and thus refuse to accept the Greek-Armenian node as a valid historical unity: Clackson 1994; Kim 2018).

In addition to the aforementioned recent clades, which are established by experts in the field, our resulting topology (Figure 2) suggests two higher level groupings: Italic-Germanic-Celtic and Balto-Slavic–Indo-Iranian. Such groupings are sometimes hypothesized by Indo-Europeanists (e.g., Schrijver 1991: 418–419 for the former and Kortlandt 2016 for the latter). It is interesting that two of the three algorithms – StarlingNJ (Figure S1e and S1f) and Maximum parsimony (Figure S3e, S3f) – resolve the West European clade as [[Italic, Germanic] Celtic] which does not contradict any historical facts, although the Bayesian MCMC analysis results in a three-way divergence [Italic, Germanic, Celtic] (Figure S2e, S2f).

In the corresponding section of the Supplement, we offer an overview of lexical innovations of the aforementioned Inner IE clades: Greek-Armenian, Balto-Slavic–Indo-Iranian, Italic-Germanic-Celtic.

The most interesting and important result of the current study is the multifurcation of Inner IE into four branches: (i) Greek-Armenian, (ii) Italic-Germanic-Celtic, (iii) Balto-Slavic–Indo-Iranian, (iv) Albanian. Such a radiation of the main

Inner IE branches, although formally innovative, is not at all incompatible with established expert views. Indeed, the majority of Indo-Europeanists, if not all of them, agree with the outlier status of Anatolian and Tocharian and with the existence of recent clades such as Indo-Iranian or Balto-Slavic. But what is in the middle of the IE tree? Despite more than two hundred years of intense development of Indo-European studies, there is no consensus or mainstream opinion on what the early Inner IE branchings could look like. The lack of an evident solution in terms of a tree structure leads many Indo-Europeanists to reject the tree model altogether or to accept a total fan-like model for the disintegration of IE which does not make much sense from the historical point of view. This lack of consensus follows from the lack of reliable and consistent common innovations shared by some subset(s) of Inner IE branches that could help to reveal the early topology of this clade. In such a situation the multifurcation scenario suggested by our analysis is the most natural and likely solution.

It must be noted that some historical linguists believe that any multifurcation is simply a poorly resolved sequence of bifurcations. This follows the widespread view of modern biologists, which is mostly based on the computational limitations of existing software packages. Our opinion is that this maxim is not correct for the history of languages. On the contrary, one can expect that a fan-like disintegration of a population and its language into several branches may have frequently happened in human history. Moreover, if such a disintegration represents, in the strict sense, a sequence of bifurcations with short time intervals (say, one or two human generations), we believe that from the historical point of view it is more reasonable to treat this as a single event with multifurcation.

Unsurprisingly, the most problematic taxon is Albanian, which skips across the tree, occupying various positions within the Inner IE clade: from the first outlier to the third member of the Balto-Slavic–Indo-Iranian clade, depending on the dataset and the method used (Figures S1–S3). There are underlying reasons for such instability. First there are a huge number of non-inherited items in the Albanian basic vocabulary, which are thus excluded from our dataset, as scarceness of data leads to lack of resolution. A second reason is our insufficient knowledge of Albanian historical phonology. Due to the scarceness of Albanian inherited vocabulary we may be failing to detect some non-trivial phonological rules. As a result, certain Albanian stems that are treated here as etymologically isolated may actually be true cognates of the corresponding Proto-IE terms.

Chronological intervals obtained by Bayesian MCMC analysis and summarized in Table 3 do not contradict the expert views. For example, the initial IE bifurcation, the Anatolian split-off, falls within the range 4139–3450 BC that is compatible with traditional estimations, e.g., 4000–2500 BC in Beekes (2011: 51), 5000–3000 BC

in Meier-Brügger (2010: 194), 3600–3500 BC in Garrett (2006: 146), although Fortson (2004: 39) speaks of the more recent date of ca. 3100 BC.

Moreover, Bayesian chronological intervals (Table 3) do not contradict radiocarbon datings of archaeological cultures that may be associated with the spread of Indo-European languages. Thus, the split-off of Tocharian can be identified with the migration that gave rise to the Afanasievo culture (Anthony and Ringe 2015: 208; Danilenko 1974: 138, 142, 157; Mallory 1997; Semenov 1987). The current C14 dates for Afanasievo place it in the interval from the 29th to 25th centuries BC (Svyatko et al. 2009: 257). The 29th century BC date for the rise of Afanasievo aligns well with the Bayesian date for the Tocharian split-off, namely 3727–2262 BC (mean 3011 BC).

The end of the Sintashta archaeological culture, frequently associated with Proto-Indo-Iranian speakers (Anthony 2007: 408–411; Kuz'mina 2007; Parpola and Carpelan 2005: 129), is dated the beginning of the 18th century BC (Epimakhov and Krause 2013; Hanks et al. 2007). Cf. the Bayesian dates for the break-up of Proto-Indo-Iranian: 2044–1458 BC (mean 1740 BC).

According to ancient DNA data, it is likely that the population of the Sintashta culture is at least partially descended from that of the Corded Ware culture.

> Although we cannot formally test whether the Sintashta derives directly from an eastward migration of Corded Ware peoples or if they share common ancestry with an earlier steppe population, the presence of European Neolithic farmer ancestry in both the Corded Ware and the Sintashta, combined with the absence of Neolithic farmer ancestry in the earlier Yamnaya, would suggest the former being more probable. (Allentoft et al. 2015: 169).

Since the Corded Ware culture is usually associated (non-exclusively) with the ancestors of Balto-Slavic peoples (Anthony 2007: 367; Mallory and Adams 1997), it seems reasonable to suppose that the Balto-Slavic–Indo-Iranian break-up may be correlated with the end of Corded Ware. According to the current view, "[t]he years between 2300 and 2100 BC were a period during which the Corded Ware culture ended in most regions, especially in the southern part of its domain (basins of the Danube, Upper Rhine, Elbe, and Vistula). Only in the Russian Plain did it last until 2000 BC." (Czebreszuk 2004a: 469). These datings align relatively well with the Bayesian dates for the Balto-Slavic–Indo-Iranian break-up: 2723–1790 BC (mean 2241 BC).

Finally, it is not excluded that the Italic-Germanic-Celtic unity should be associated with the Bell Beaker culture. Similarly Mallory (2013) proposes a connection to the Bell Beaker culture due to its chronological depth and not with Proto-Celtic per se, but generally with an ancestor of "North-West" Indo-European languages. Recent studies on ancient DNA have confirmed that the spread of this culture in most places (with the significant exception of Iberia) was associated with a real migration rather than simply a dissemination of a "cultural package" (Olalde et al. 2018).

The latest dates for this culture extend into the first centuries of the second millennium BC. (Czebreszuk 2004b: 482). Cf. the Bayesian dates for the Italic-Germanic-Celtic break-up: 2655–1537 BC (mean 2080 BC).

That having been said, we avoid any further discussion on the IE homeland issue, since it would be a serious simplification to assume direct one-to-one correlations between our current results and the geographic distribution of early Proto-Indo-European speakers.

5 Conclusions

Input data preparation – both the initial gathering of data and post-processing such as derivational drift elimination and homoplastic optimization – plays a crucial role in linguistic phylogeny. We believe that inaccurate linguistic data might be at least partially responsible for odd topological results obtained in some previous Indo-European phylogenetic studies, and, conversely, that maximally accurate data will yield similar results regardless of the phylogenetic method used (Kassian 2015).

To the best of our knowledge, this is the first time that the initial multifurcation of the Inner Indo-European languages into four primary branches has been confirmed by formal computational methods. This result is consistent with available linguistic evidence and reconciles the multiple opinions on IE phylogeny expressed by Indo-Europeanists all the way from August Schleicher in the mid-19th century to modern-day authoritative handbooks.

Supplement

The online supplement includes wordlists (*.xls), linguistic comments on individual Swadesh words (*.pdf), phylogenetic trees obtained by individual methods (including trees with the Samoyed outgroup), and all technical files used for and produced by phylogenetic packages that render our experiment reproducible. https://doi.org/10.5281/zenodo.4046607.

Acknowledgements: We wish to thank Igor Yanovich (DFG Center for Advanced Study "Words, Bones, Genes and Tools", Universität Tübingen) and Valery Zaporozhchenko (Research Centre for Medical Genetics, Russian Academy of Sciences) for assistance in computational analyses.

Research funding: The research for this article was conducted within the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2020-908).

References

Allentoft, Morten E., Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, Hannes Schroeder, Torbjorn Ahlstrom, Lasse Vinner, Anna-Sapfo Malaspinas, Ashot Margaryan, Tom Higham, David Chivall, Niels Lynnerup, Lise Harvig, Justyna Baron, Philippe Della Casa, Paweł Dąbrowski, Paul R. Duffy, Alexander V. Ebel, Andrey Epimakho, Karin Margarita Frei, Mirosław Furmanek, T. Gralak, Andrey Gromov, Stanislaw Gronkiewicz, Gisela Grupe, Tamás Hajdu, Radosław Jarysz, V.I. Khartanovich, Alexandr Khokhlov, Viktória Kiss, Jan Kolář, Aivar Kriiska, Irena Lasak, Cristina Longhi, George McGlynn, Algimantas Merkevicius, Inga Merkyte, Mait Metspalu, Ruzan Mkrtchyan, Vyacheslav Moiseyev, Laszlo Paja, György Pálfi, Dalia Anna Pokutta, Lukasz Pospieszny, Doug Price, Lehti Saag, Mikhail Sablin, Natalia Shishlina, Vaclav Smrcka, Vasilii Soenov, Vajk Szeverényi, Gusztav Toth, Synaru V. Trifanova, Liivi Varul, Magdolna Vicze, Levon Yepiskoposyan, Vladislav Zhitenev, Ludovic Orlando, Thomas Sicheritz-Ponten, Søren Brunak, Rasmus Nielsen, Kristian Kristiansen & Eske Willerslev. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522(7555). 167–172.

Anthony, David W. 2007. *The horse, the wheel, and language: How bronze-age riders from the Eurasian steppes shaped the modern world*. Princeton, NJ: Princeton University Press.

Anthony, David W. & Don Ringe. 2015. The Indo-European homeland from linguistic and archaeological perspectives. *Annual Review of Linguistics* 1(1). 199–219.

Atkinson, Quentin D. & Russell D. Gray. 2006. How old is the Indo-European language family?: Illumination or more moths to the flame?. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 91–109. Cambridge: McDonald Institute for Archaeological Research.

Balanovsky, Oleg, Khadizhat Dibirova, Dybo Anna, Oleg Mudrak, Svetlana Frolova, Elvira Pocheshkhova, Marc Haber, Daniel Platt, Theodore Schurr, Wolfgang Haak, Marina Kuznetsova, Magomed Radzhabov, Olga Balaganskaya, Alexey Romanov, Tatiana Zakharova, F David, Hernanz Soria, Pierre Zalloua, Sergey Koshel, Merritt Ruhlen, Colin Renfrew, R. Spencer Wells, Chris Tyler-Smith & Balanovska Elena & the Genographic Consortium. 2011. Parallel evolution of genes and languages in the Caucasus region. *Molecular Biology and Evolution* 28(10). 2905–2920.

Beekes, Robert S. P. 2011. Comparative Indo-European linguistics: An introduction. In Cor de Vaan Michiel Arnoud (ed.), 2nd edn. Amsterdam & Philadelphia: John Benjamins.

Blažek, Václav. 2007. From August Schleicher to Sergei Starostin. On the development of the tree-diagram models of the Indo-European languages. *Journal of Indo-European Studies* 35(1). 82–109.

Blažek, Václav & Petra Novotná. 2007. Glottochronology and its application to the Balto-Slavic languages. *Baltistica* 42. 185–210.

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russel D. Gray, Marc A. Suchard & Quentin D. Atkinson. 2012. Mapping the origins and expansion of the Indo-European language family. *Science* 337. 957–960.

Burlak, Svetlana A. & Sergei A. Starostin. 2005. *Sravnitel'no-istoričeskoe jazykoznanie [Comparative-historical linguistics]*. Moscow: Academia.

Chang, Will, Chundra Cathcart, David Hall & Andrew Garrett. 2015. Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis. *Language* 91(1). 194–244.

Clackson, James P. T. 1994. *The linguistic relationship between Armenian and Greek (Publications of the Philological Society 30)*. Oxford: Blackwell.

Cowgill, Warren. 1986. Einleitung. In Manfred Mayrhofer (ed.), *Indogermanische Grammatik*, vol. 1. Heidelberg: Carl Winter.

Czebreszuk, Janusz. 2004a. Corded Ware from east to west. In Peter Bogucki, J Pam & Crabtree (eds.), *Ancient Europe 8000 bc–1000 AD: An encyclopedia of the Barbarian World*, vol. 1, 467–475. New York: Charles Scribner's Sons.

Czebreszuk, Janusz. 2004b. Bell Beakers from west to east. In Peter Bogucki, J Pam & Crabtree (eds.), *Ancient Europe 8000 bc–1000 AD: An encyclopedia of the Barbarian World*, vol. 1, 476–485. New York: Charles Scribner's Sons.

Danilenko, Valentin N. 1974. *Èneolit Ukrainy [The Chalcolithic in Ukraine]*. Kiev: Naukova dumka.

Epimakhov, Andrey V. & Rüdiger Krause. 2013. Relative and absolute chronology of the settlement Kamennyi Ambar. In Rüdiger Krause & Ludmila N. Koryakova (eds.), *Multidisciplinary investigations of the Bronze Age settlements in the Southern Trans-Urals (Russia)*, 129–146. Bonn: Verlag Dr. Rudolf Habelt.

Fortson, Benjamin W. 2004. *Indo-European language and culture: An introduction (Blackwell Textbooks in Linguistics 19)*. Malden, MA: Blackwell.

Gamkrelidze, Thomas & Vyacheslav V. Ivanov. 1995. Indo-European and the Indo-Europeans: A reconstruction and historical analysis of a proto-language and a proto-culture. In Werner Winter (ed.), *[trans. by Johanna Nichols]*, vol. 2. Berlin & New York: Mouton de Gruyter.

Garrett, Andrew. 2006. Convergence in the formation of Indo-European subgroups: phylogeny and chronology. In Peter Forster & Colin Renfrew (eds.), *Phylogenetic methods and the prehistory of languages*, 139–151. Cambridge: McDonald Institute for Archaeological Research.

Goloboff, Pablo A. & Santiago A. Catalano. 2016. TNT version 1.5, including a full implementation of phylogenetic morphometrics. *Cladistics* 32(3). 221–238.

Gray, Russell D. & Quentin D. Atkinson. 2003. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426. 435–439.

Hajdú, Péter. 1988. Die Samojedischen Sprachen. In Denis Sinor (ed.), *The Uralic languages: Description, history and foreign influences*, 3–40. Leiden: Brill.

Hanks, Bryan K., Andrey V. Epimakhov & A. Colin Renfrew. 2007. Towards a refined chronology for the Bronze Age of the southern Urals, Russia. *Antiquity* 81(312). 353–367.

Jasanoff, Jay H. 2003. *Hittite and the Indo-European verb*. Oxford: Oxford University Press.

Kapović, Mate. 2017. Indo-European languages: Introduction. In Mate Kapović (ed.), *The Indo-European languages*, 2nd edn., 1–9. London: Routledge.

Kassian, Alexei S. 2009. Hattic as a Sino-Caucasian language. *Ugarit-Forschungen* 41. 309–447.

Kassian, Alexei S. 2015. Towards a formal genealogical classification of the Lezgian languages (North Caucasus): Testing various phylogenetic methods on lexical data. *PLOS ONE* 10(2). e0116950.

Kassian, Alexei S. 2017. Linguistic homoplasy and phylogeny reconstruction: The cases of Lezgian and Tsezic languages (North Caucasus). *Folia Linguistica* 51(s38). https://doi.org/10.1515/flih-2017-0008.

Kassian, Alexei S., Starostin George, Dybo Anna & Vasily Chernov. 2010. The Swadesh wordlist. An attempt at semantic specification. *Journal of Language Relationship* 4. 46–89.

Kassian, Alexei S., George S. Starostin & Mikhail A. Zhivlov. 2015a. Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies* 43(3/4). 301–347.

Kassian, Alexei S., George S. Starostin & Mikhail A. Zhivlov. 2015b. Lexicostatistics, probability, and other matters. *Journal of Indo-European Studies* 43(3/4). 376–392.

Kim, Ronald I. 2018. Greco-Armenian: The persistence of a myth. *Indogermanische Forschungen* 123(1). 247–272.

Koch, John T. 2006. Pytheas. In John T. Koch (ed.), *Celtic culture: A historical encyclopedia, 1472*. Santa Barbara, CA: ABC-CLIO.

Kortlandt, Frederik. 2010. *Studies in Germanic, Indo-European and Indo-Uralic*. Amsterdam: Rodopi.

Kortlandt, Frederik. 2016. Balto-Slavic and Indo-Iranian. *Baltistica* 51(2). 355–364.

Kulikov, Leonid. 2017. Indo-Iranian. In Mate Kapović (ed.), *The Indo-European languages*, 2nd edn., 205–213. London: Routledge.

Kushniarevich, Alena, Olga Utevska, Marina Chuhryaeva, Anastasia Agdzhoyan, Khadizhat Dibirova, Ingrida Uktveryte, Märt Möls, Lejla Mulahasanovic, Andrey Pshenichnov, Svetlana Frolova, Andrey Shanko, Ene Metspalu, Maere Reidla, Kristiina Tambets, Erika Tamm, Sergey Koshel, Valery Zaporozhchenko, Lubov Atramentova, Vaidutis Kučinskas, Oleg Davydenko, Olga Goncharova, Irina Evseeva, Michail Churnosov, Elvira Pocheshchova, Bayazit Yunusbayev, Elza Khusnutdinova, Damir Marjanović, Pavao Rudan, Siiri Rootsi, Nick Yankovsky, Phillip Endicott, Alexei Kassian, Dybo Anna, Genographic Consortium, Chris Tyler-Smith, Balanovska Elena, Mait Metspalu, Toomas Kivisild, Richard Villems & Oleg Balanovsky. 2015. Genetic heritage of the Balto-Slavic speaking populations: A synthesis of autosomal, mitochondrial and Y-chromosomal data. *PLOS ONE* 10(9). 1–19.

Kuz'mina, Elena Efimovna. 2007. In James P. Mallory (ed.), *The origin of the Indo-Iranians*. Leiden: Brill.

Lundquist, Jesse & Anthony D. Yates. 2018. The morphology of Proto-Indo-European. In Jared S. Klein, Brian D. Joseph & Matthias Fritz (eds.), *Handbook of comparative and historical Indo-European linguistics (Handbücher Zur Sprach- Und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science 41.3*, vol. 3, 2079–2195. Berlin & Boston: De Gruyter Mouton.

Makarenkov, Vladimir, Dmytro Kevorkov & Pierre Legendre. 2006. Phylogenetic network construction approaches. In Dilip K. Arora, Randy M. Berka & Gautam B. Singh (eds.), *Applied mycology and biotechnology*, vol. 6, 61–97. Amsterdam: Elsevier.

Mallory, James P. 1989. *In search of the Indo-Europeans. Language, archaeology and myth*. London: Thames and Hudson.

Mallory, James P. 1997. Afanasevo culture. In James P. Mallory & Douglas Q. Adams (eds.), *Encyclopedia of Indo-European culture*, 4–6. London: Fitzroy Dearborn.

Mallory, James P. 2013. The Indo-Europeanization of Atlantic Europe. In John T. Koch & Barry Cunliffe (eds.), *Celtic from the West 2: Rethinking the Bronze Age and the arrival of Indo-European in Atlantic Europe*, 17–40. Oxford: Oxbow Books.

Mallory, James P. & Douglas Q. Adams. 1997. Corded ware culture. In James P. Mallory & Douglas Q. Adams (eds.), *Encyclopedia of Indo-European culture*, 127–128. London: Fitzroy Dearborn.

Meier-Brügger, Michael. 2010. *Indogermanische Sprachwissenschaft (De Gruyter Studium)*, 9th edn. Berlin New York: De Gruyter.

Müller, André, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Eric W. Holman, Sebastian Sauppe, Pamela Brown, Harald Hammarström, Oleg Belyaev, Johann-Mattis List, Dik Bakker, Dmitri Egorov, Matthias Urban, Robert Mailhammer, Matthew S. Dryer, Evgenia Korovina, David Beck, Helen Geyer, Pattie Epps, Anthony Grant & Pilar Valenzuela. 2013. *ASJP world language trees of lexical similarity* Version 4. https://asjp.clld.org/download (accessed 7 October 2018).

Nakhleh, Luay, Tandy Warnow, Don Ringe & Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. *Transactions of the Philological Society* 103. 171–192.

Nussbaum, Alan. 1976. *Caland's "law" and the Caland system*. Cambridge, MA: Harvard University dissertation.

Olalde, Iñigo, Selina Brace, Morten E. Allentoft, Ian Armit, Kristian Kristiansen, Thomas Booth, Nadin Rohland, Swapan Mallick, Anna Szécsényi-Nagy, Alissa Mittnik, Eveline Altena, Mark Lipson, Iosif Lazaridis, Thomas K. Harper, Nick Patterson, Nasreen Broomandkhoshbacht, Yoan Diekmann, Zuzana Faltyskova, Daniel Fernandes, Matthew Ferry, Eadaoin Harney, Peter de Knijff, Megan Michel, Jonas Oppenheimer, Kristin Stewardson, Alistair Barclay, Kurt Werner Alt, Corina Liesau, Patricia Ríos, Concepción Blasco, Jorge Vega Miguel, Roberto Menduiña García, Azucena Avilés Fernández, Eszter Bánffy, Maria Bernabò-Brea, David Billoin, Clive Bonsall, Laura Bonsall, Tim Allen, Lindsey Büster, Sophie Carver, Laura Castells Navarro, Oliver E. Craig, T Gordon, 33 Cook, Barry Cunliffe, Denaire Anthony, Kirsten Egging Dinwiddy, Natasha Dodwell, Michal Ernée, Christopher Evans, Milan Kuchařík, Joan Francès Farré, Chris Fowler, Michiel Gazenbeek, Rafael Garrido Pena, María Haber-Uriarte, Elżbieta Haduch, Hey Gill, Nick Jowett, Timothy Knowles, Ken Massy, Saskia Pfrengle, Philippe Lefranc, Olivier Lemercier, Arnaud Lefebvre, César Heras Martínez, Virginia Galera Olmo, Ana Bastida Ramírez, Joaquín Lomba Maurandi, Tona Majó, Jacqueline I. McKinley, Kathleen McSweeney, Balázs Gusztáv Mende, Alessandra Modi, Gabriella Kulcsár, Viktória Kiss, András Czene, Róbert Patay, Endrődi Anna, Kitti Köhler, Tamás Hajdu, Tamás Szeniczey, János Dani, Zsolt Bernert, Maya Hoole, Olivia Cheronet, Denise Keating, Petr Velemínský, Miroslav Dobeš, Francesca Candilio, Fraser Brown, Raúl Flores Fernández, Ana-Mercedes Herrero-Corral, Sebastiano Tusa, Emiliano Carnieri, Luigi Lentini, Antonella Valenti, Alessandro Zanini, Clive Waddington, Germán Delibes, Elisa Guerra-Doce, Benjamin Neil, Brittain Marcus, Mike Luke, Richard Mortimer, Jocelyne Desideri, Marie Besse, Günter Brücken, Mirosław Furmanek, Agata Hałuszko, Maksym Mackiewicz, Artur Rapiński, Stephany Leach, Ignacio Soriano, Katina T. Lillios, João Luís Cardoso, Michael Parker Pearson, T Piotr Włodarczak, Douglas Price, Pilar Prieto, Pierre-Jérôme Rey, Roberto Risch, Manuel A. Rojo Guerra, Aurore Schmitt, Joël Serralongue, Ana Maria Silva, Smrčka Václav, Luc Vergnaud, João Zilhão, David Caramelli, Thomas Higham, Mark G. Thomas, Douglas J. Kennett, Harry Fokkens, Volker Heyd, Alison Sheridan, Karl-Göran Sjögren, Philipp W. Stockhammer, Johannes Krause, Pinhasi Ron, Wolfgang Haak, Ian Barnes, Carles Lalueza-Fox & David Reich. 2018. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* 555(7695). 190–196.

Olander, Thomas. 2019. Indo-European cladistic nomenclature. *Indogermanische Forschungen* 124(1). 231–244.

Parpola, Asko & Christian Carpelan. 2005. The cultural counterparts to Proto-Indo-European, Proto-Uralic and Proto-Aryan: Matching the dispersal and contact patterns in the linguistic and archaeological record. In Edwin F. Bryant & Laurie L. Patton (eds.), *The Indo-Aryan controversy: Evidence and inference in Indian history*, 107–141. London: Routledge.

Pereltsvaig, Asya & Martin W. Lewis. 2015. *The Indo-European controversy: Facts and fallacies in historical linguistics*. Cambridge: Cambridge University Press.

Rama, Taraka & Søren Wichmann. 2018. Towards identifying the optimal datasize for lexically-based Bayesian inference of linguistic phylogenies. In *Proceedings of the 27th International Conference on Computational Linguistics*. In Emily M. Bender, Derczynski Leon & Pierre Isabelle (eds.), 1578–1590. Santa Fe, NM: Association for Computational Linguistics.

Rexová, Kateřina, Daniel Frynta & Zrzavý Jan. 2003. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics* 19. 120–127.

Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1). 59–129.

Ronquist, Fredrik, Maxim Teslenko, Paul van der Mark, Daniel L. Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A. Suchard & John P. Huelsenbeck. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* 61(3). 539–542.

Schleicher, August. 1861. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Weimar: H. Böhlau.

Schrijver, Peter. 1991. *The reflexes of the proto-Indo-European laryngeals in Latin (Leiden Studies in Indo-European)*. Amsterdam: Rodopi.

Semenov, Vladimir A. 1987. Drevnejamnaja kul'tura – afanas'evskaja kul'tura i problemy prototoxarskoj migracii na vostok [Pit Grave culture – Afanasievo culture and the problems of the eastward migration of Proto-Tocharians]. In L. M. Pletneva (ed.), *Smeny kul'tur i migracii v Zapadnoj Sibiri*, 17–19. Tomsk: Izd-vo TGU.

Starostin, George S. 2010. Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship* 3. 79–116.

Starostin, George S. 2013. Lexicostatistics as a basis for language classification: increasing the pros, reducing the cons. In Heiner Fangerau, Hans Geisler, Thorsten Halling, F William & Martin (eds.), *Classification and evolution in biology, linguistics and the history of science: Concepts – methods – visualization (KulturAnamnesen 5)*, 125–146. Stuttgart: Franz Steiner Verlag.

Starostin, George S. 2016. From wordlists to proto-wordlists: reconstruction as 'optimal selection. *Faits de langues* 47(1). 177–200.

Starostin, George S. (ed.). 2011. Global lexicostatistical database. Available at: http://starling.rinet.ru/new100/main.htm.

Starostin, Sergei A. 2000. Comparative-historical linguistics and lexicostatistics. In Colin Renfrew, April McMahon & Larry Trask (eds.), *Time depth in historical linguistics*, vol. 1, 223–265. Cambridge: The McDonald Institute for Archaeological Research.

Starostin, Sergei A. 2007a. *The historical position of Bai. Trudy po jazykoznaniju [Works on linguistics]*, 580–590. Moscow: Jazyki slavjanskix kul'tur.

Starostin, Sergei A. 2007b. *Sravnitel'no istoričeskoe jazykoznanie i leksikostatistika [Comparative-historical linguistics and lexicostatistics]. Trudy po jazykoznaniju [Works on linguistics]*, 407–447. Moscow: Jazyki slavjanskix kul'tur.

Starostin, Sergei A. 2007c. *Rabočaja sreda dlja lingvista [Linguist's workspace]. Trudy po jazykoznaniju [Works on linguistics]*, 481–496. Moscow: Jazyki slavjanskix kul'tur.

Starostin, Sergei A. 2007d. *Opredelenie ustojčivosti bazisnoj leksiki [Defining the stability of basic lexicon]. Trudy po jazykoznaniju [Works on linguistics]*, 827–839. Moscow: Jazyki slavjanskix kul'tur.

Svyatko, Svetlana V., James P. Mallory, Eileen M. Murphy, Andrey V. Polyakov, Paula J. Reimer & Rick J. Schulting. 2009. New radiocarbon dates and a review of the chronology of prehistoric populations from the Minusinsk Basin, Southern Siberia, Russia. *Radiocarbon* 51(1). 243–273.

Winter, Werner. 1996. Lexical archaisms in the Tocharian languages. In Hans Henrich Hock (ed.), *Historical, Indo-European, and lexicographical studies: A festschrift for Ladislav Zgusta on the occasion of his 70th birthday*, 183–194. Berlin & New York: Mouton de Gruyter.

Wodtko, Dagmar S., Britta Sofie Irslinger & Carolin Schneider. 2008. *Nomina im indogermanischen Lexikon*. Heidelberg: Winter.

Yanovich, Igor. 2020. Phylogenetic linguistic evidence and the Dene-Yeniseian homeland. *Diachronica* 37(3). 410–446.

Zhivlov, Mikhail A. & Nina Yu. Zhivlova. 2015. The precursors of Proto-Indo-European: The Indo-Hittite and Indo-Uralic hypotheses, Leiden University, 9–11.07.2015. *Journal of Language Relationship* 13(3). 281–288.