Федеральное государственное бюджетное образовательное учреждение высшего образования
«РОССИЙСКАЯ АКАДЕМИЯ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»

Касьян А.С., Живлов М.А., Старостин Г.С., Трофимов А.А.

Новый подход к индоевропейской классификации: фонетико-семантическая реконструкция базисных лексических единиц для промежуточных праязыковых состояний

Москва 2019

Аннотация. В настоящем препринте предлагается реконструкция индоевропейской филогении на основе тринадцати 110-словных списков для праязыков ИЕ-подгрупп (протогерманский, протославянский и т. д.) или древних языков соответствующих подгрупп (хеттский, древнегреческий и т. д.). Мы применяем более или менее формальные методы сбора лингвистических данных и последующей обработки (семантическая реконструкция, деривационная дрейфовая элиминация, оптимизация гомоплазий), недавно предложенные или разработанные для настоящего исследования. Мы используем последовательную филогенетическую процедуру и получаем консенсусное дерево, основанное на нескольких алгоритмах (байесовский анализ, максимальная экономность, присоединение соседей, отсутствие применения топологических ограничений). Полученная топология дерева и датировки полностью совместимы с традиционными представлениями экспертов. Нашей главной находкой является мультифуркация внутренней ИЕ клады на четыре ветви ок. 3357-2162 гг. до н.э.: (1) греко-армянскую, (2) албанскую, (3) итало-германо-кельтскую, (4) балто-славяно-индо-иранскую. Предлагаемый сценарий расхождения примиряет друг с другом различные мнения о внутреннем ИЕ ветвлении, высказывавшиеся ранее индоевропеистами.

Касьян А.С. старший научный сотрудник научно-исследовательской лаборатории востоковедения и компаративистики ШАГИ ИОН Российской академии народного хозяйства и государственной службы при Президенте РФ

Живлов М.А. старший научный сотрудник старший научный сотрудник научно-исследовательской лаборатории востоковедения и компаративистики ШАГИ ИОН Российской академии народного хозяйства и государственной службы при Президенте РФ

Старостин Г.С. заведующий старший научный сотрудник научно-исследовательской лаборатории востоковедения и компаративистики ШАГИ ИОН Российской академии народного хозяйства и государственной службы при Президенте РФ

Трофимов А.А. старший научный сотрудник научно-исследовательской лаборатории востоковедения и компаративистики ШАГИ ИОН Российской академии народного хозяйства и государственной службы при Президенте РФ

Данная работа подготовлена на основе материалов научно-исследовательской работы, выполненной в соответствии с Государственным заданием РАНХиГС при Президенте Российской Федерации на 2018 год

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Индоевропейская языковая семья в настоящее время самая большая в мире по географическому охвату и количеству языковых носителей. Она включает в себя несколько сотен живых и десятки древних языков (частично вымерших, частично имеющих прямых потомков в наши дни). В семье есть двенадцать подгрупп, которые единодушно признаются экспертами в этой области и подтверждаются любым методом формального анализа: анатолийская, тохарская, греческая, армянская, албанская, италийская (романская), кельтская, германская, славянская, балтийская, индийская, иранская.

Несмотря на более чем 150-летний опыт филогенетических ИЕ исследований - новаторское ИЕ дерево было опубликовано Шлейхером [1: 7] — единственный консенсус или почти консенсус, достигнутый индоевропеистами, касается внешнего статуса анатолийского и тохарского, а также наличия отдельной индо-иранской клады. Внутренние ветвления, происходившие между отделением тохарского и образованием вышеупомянутых молодых клад (греческой, германской т. д.), по-прежнему остаются предметом дискуссий среди экспертов. Мнения настолько противоречивы, что большинство индоевропеистов предпочитают вовсе не обсуждать ветвление внутренних ИЕ языков. В настоящей статье для удобства мы используем определение «внутренние ИЕ» для ИЕ языков после отделения анатолийского и тохарского (термин был введен в [2] и затем принят некоторыми другими индоевропеистами).

В последние десятилетия был опубликован ряд формальных филогений ИЕ языковой семьи. Они основаны либо на лексических единицах, т. е. на списке семантических концептов, так называемой лексикостатистике (например, [3, 4]) или на смешанных - фонологических, грамматических и лексических - наборах данных [5, 6].

Итоговые топологии и даты, предложенные в этих исследованиях, противоречат друг другу во многих деталях. Некоторые из полученных деревьев более уместны с точки зрения индоевропестов, например, публикации команды Ринджа. Другие менее убедительны, например, ИЕ деревья и датировки в [4, 7] несовместимы с экспертными мнениями в некоторых пунктах (мы считаем, что это вызвано использованием неточных вводных данных, см., например, в качестве некоторой критики лингвистическое приложение к [8]).

Учитывая недостатки предыдущих исследований, цель представленного исследования состоит в том, чтобы получить усовершенствованное дерево ИЕ языков, которое полностью соответствовало бы мнениям традиционных экспертов, будучи основанным на высоком качестве доказательства и новой методологии.

1 Материалы и методы

Сбор данных: Наш анализ основан на 110-словных списках Сводеша, принятых в проекте «Глобальная лексикостатистическая база данных» (starling.rinet.ru/new100). Так как подгруппы (такие как славянская, германская, армянская и т. д.) внутри ИЕ семьи являются бесспорными, мы предпочитаем использовать для каждой подгруппы реконструированный праязыковой список слов (например, прагерманский для германской группы) или список засвидетельствованного языка, который в целом может быть приравнен к протоязыку (например, ведический для индоарийской группы). Помимо этого, мы вводим прасамодийский как представителя уральской семьи — ближайшего известного родственника индоевропейской семьи. Мы используем следующие списки слов (см. Приложение для основной информации о подгруппах и обсуждение дат): см. Таблицу 1.

Таблица 1 - Хронологические ограничения на таксоны.

Таксон	Ограничения на байесовский анализ	Точные даты для анализа StarlingNJ
Древнехеттский	1650-1500 BC	1550 BC
Тохарский Б	400-900 AD	650 AD
Древнегреческий	375 BC	375 BC
Древнеармянский	400-500 AD	450 AD
Албанский	1950 AD	1950 AD
Архаическая латынь	200 BC	200 BC
Древнеирландский	700-900 AD	800 AD
Прабретонский	300-600 AD	450 AD
Прагерманский	500-300 BC	400 BC
Праславянский	1-300 AD	100 AD
Правост.-балтийский	400-1 BC	200 BC
Древнеиндийский (Атхарваведа)	1200-1000 BC	1100 BC

Продолжение таблицы 1

Таксон	Ограничения на байесовский анализ	Точные даты для анализа StarlingNJ
Праиранский	1500-1000 BC	1300 BC
Прасамодийский	950-750 BC	800 BC
Праиндоевропейский	3500–8500 BC	—

Мы считаем, что использование реконструированных списков слов для праязыков подгрупп вместо традиционного подхода с большим количеством списков слов из современных языков предпочтительнее по двум причинам.

Во-первых, в [9] для основанного на лексике байесовского вывода предлагается, что количество родственных классов (и, следовательно, символов, т. е. семантических концептов), необходимых для адекватной реконструкции языковой филогении, прямо пропорционально числу таксонов. Другими словами, чем больше набор языков, тем большее количество семантических концептов необходимо. По оценке Рамы и Вихмана, 100-словный список достаточен для набора из 30 или менее лектов, для 31-100 лектов требуется 200-словный список и т. д. Лингвистические данные анализа в [9] имеют в основе серьезные качественные различия, что может привести к некоторым искажениям их результатов, но их основная гипотеза о прямой связи между количеством таксонов и количеством символов интуитивно логична и кажется вполне правильной.

Вторая причина касается всех филогенетических методов, а не только байесовского. Поэтапная реконструкция уменьшает количество гомопластических развитий в наборе данных, делает набор данных менее шумным и, следовательно, делает всю модель менее сложной. Дополнительным источником шумных данных являются ненадежные источники для некоторых языков, что почти неизбежно в ситуации, когда задействовано много разных лектов различной степени документированности.

Другая сторона медали поэтапной реконструкции, однако, заключается в том, что можно некорректно реконструировать некоторую черту праязыка, например, поместить определенное прото-слово в список Сводеша, хотя исторически это слово не было основным выражением для данного семантического концепта в праязыке. Тем не менее, мы не считаем риск ошибок реконструкции слишком высоким, потому что, во-первых, мы придерживаемся строгой методологии семантической реконструкции (см. данный раздел ниже), во-вторых, праязыки, о которых идет речь, хронологически не очень глубоки: обычно мы имеем дело с дистанцией 2000-2500 лет до н.э.

Для засвидетельствованных языков 100-словные списки Сводеша собирались в соответствии с эксплицитными семантическими спецификациями [10] при использовании наиболее авторитетных лексикографических источников и, в случае необходимости, проверялись по текстовым корпусам. Такая строгость приводит к очень высокому лингвистическому качеству входных данных, что отличает наш проект от некоторых предыдущих попыток ИЕ филогении.

Большинство используемых списков слов доступно онлайн в проекте GLD (http://starling.rinet.ru/new100/main.htm); все списки предоставлены в Дополнении. Семантическая реконструкция для промежуточных праязыков отдельных групп внутри ИЕ семьи основана на относительно строгой методологии, предложенной в [11: 304-306]. Ее основными принципами являются: топология дерева, внешняя этимология, внутренняя выводимость, типология семантических сдвигов, исключение ареального эффекта. Каждую праформу и реконструированное значение мы сопровождаем подробными комментариями, которые объясняют наш выбор.

2 Построение матрицы

В работе мы сравниваем три последовательно применяемых метода маркирования родства:

1) Этап-1. Высококачественный набор данных с традиционным корневым родством (например, англ. wind — соответствие русскому ветер, эти формы в конечном счете представляют собой отдельные производные от одного и того же глагольного прото-корня). Этот набор данных находится вне фокуса настоящего исследования, он рассматривается только в Приложении, используясь в качестве ориентира для некоторых выводов.

2) Этап-2. Набор данных без деривационного дрейфа. Это набор данных о корневом родстве из Этапа-1, где формы, демонстрирующие так называемый деривационный дрейф, дополнительно помечаются как несвязанные (английский *wind* ≠ русский ветер), см. ниже о подробной и формальной процедуре выявления деривационного дрейфа. Это наш основной набор входных данных.

3) Этап-3. Набор данных, оптимизированный в отношении гомоплазий. Это набор очищенных от деривационного дрейфа данных из Этапа-2, в котором когнаты, нарушающие топологию дерева, дополнительно помечаются как не связанные (не только *wind* ≠ ветер, но и др. инд. agni ≠ лат. ignis), см. подробнее ниже и в [13]. Это тот набор данных, на основе которого построено итоговое ИЕ дерево.

На каждом этапе разрабатывается по два набора данных: собственно ИЕ (только индоевропейские списки слов) и ИЕ-самодийский (с добавлением прасамодийского списка слов). Таким образом, всего у нас получается шесть наборов данных.

Лексическая матричная компиляция является стандартной процедурой: в рамках единой концепции Сводеша этимологически родственные формы из разных языков помечаются одним и тем же индексом, т. е. включаются в один и тот же класс когнатов (см., например, [12: 93-94]). Наш подход имеет две особенности, требующие специальных пояснений.

Первая особенность нашей процедуры заключается в том, что мы помечаем заимствования как лакуны, а не как одиночные формы (формы с уникальным родственным индексом) - это стандартный для Московской школы способ, поскольку лексические замены путем заимствования не отражают эволюцию естественного языка и могут быть в большой степени обусловлены экстралингвистическими обстоятельствами. Как было предложено в [13: 225], исключительный случай, когда мы рассматриваем заимствования как одиночные формы, - это ситуация, когда у нас есть свидетельства в пользу того, что данное слово было заимствовано в целевой язык с несводешовским

значением, использовалось с таким значением в течение определенного времени, а затем приобрело сводешовскую функцию благодаря естественному семантическому и морфологическому развитию. Примеры, где в качестве одиночных форм разумно рассматривать заимствованные слова или слова с заимствованными корнями - это современное немецкое Kopf 'head' <древневерхненемецкое kopf ‘mug, bowl’ <латинское cupa, cuppa ‘cask, bowl’ или романское слово для значения «печень» (итальянское fegato, французское foie, etc.) <вульгарно-латинское ficato ‘fig-stuffed liver (a dish)’, образованное от латинского fīcus 'fig' <субстратный Средиземноморский термин для 'fig'.

В нашем текущем наборе данных албанский список слов является наиболее пораженным заимствованиями. В предварительном порядке мы рассматриваем такие случаи, как албанское flokë 'волосы (на голове)' <вульгарно-латинское *floccus* ‘локон, пучок’ или албанское *kripë* "соль" < болгарское *krupa* "комок соли" в качестве заимствований, помечая их в матрице знаком "?". Причина в том, что, во-первых, эти единицы могли проникнуть в албанский сразу со сводешовским значением, ср. сопутствующие семантические сдвиги при переходе слова из одного языка в другой: эстонское hunt "волк" <нижненемецкое hunt "собака" или средневаллийское ofydd ″поэт, литератор, любовник, возлюбленный, дорогой; мастер, чемпион" < латинский (Публий) Овидий (Назон). Во-вторых, когда мы имеем дело с многочисленными народно-латинскими и романскими заимствованиями в албанский язык, мы не знаем точно, какой лект был донором, это может быть недокументированный язык или диалект, где наблюдаемые сдвиги ("локон, пучок"> "волосы" и т. д.) уже произошли. Напротив, албанское *kokë* "голова / луковица / ягода / зерно" (<латинское *coccum* ‘berry’) помечается как одиночная форма, поскольку есть свидетельства того, что семантический сдвиг "ягода"> "голова" произошел уже на албанской почве.

В последнее время проблема заимствований обсуждалась Чангом и др. в [14: 12], которые проводят два вида филогенетического анализа ИЕ семьи. В первом они рассматривают все заимствования как одиночные формы. Во втором они включают заимствования как полноценные когнаты их лексических источников. Чанг и др. показывают, что при анализе хронологии исключение заимствований необоснованно. Их аргументы заключаются в следующем: во-первых, недавние заимствования сначала могут функционировать как маргинальные термины и лишь постепенно приобретать статус основного выражения для данного значения, таким образом воспроизводя эволюцию исконных слов (единственный упомянутый пример - французское заимствование *animal*, ставшее базовым термином после нескольких столетий употребления в английском языке).

С другой стороны, обнаружение заимствований более сложно для древних языков, чем для современных, поскольку в первом случае языки-доноры могут оставаться неизвестными нам, не будучи задокументированными в ходе их истории.

Оба аргумента Чанга и др. кажутся нам не вполне убедительными. Чаще всего заимствованное слово быстро приобретает статус базового термина под давлением доминирующего языка-донора. Такие случаи не отражают внутриязыковой лексической эволюции и таким образом затрудняют хронологическую оценку. Этот тип лексической замены особенно характерен для ситуации, когда мигрирующее слово уже имело сводешовский статус в языке-доноре. Примером может служить французское заимствование *mountain*, которое впервые появляется в среднеанглийском ок. 1200 и задокументировано как базовый термин уже в рассказах Чосера или даже раньше. Что касается примера *animal* в [14], то в рамках нашего подхода он также будет рассматриваться как одиночное слово, так как современное английское *animal* технически является тем же случаем внутренней семантической эволюции, что и немецкое *Kopf* или итальянское fegato, упомянутые выше.

С другой стороны, скрытые заимствования являются серьезной проблемой не только для древних языков, но и для многих современных языков, если на этих языках говорят на территории, плохо описанной (социо)лингвистически за последние несколько веков. Тем не менее, в некоторых случаях, когда язык-донор неизвестен, мы можем выявлять заимствования на основе их специфических фонологических или морфологических признаков.

В любом случае, вряд ли будет правильной идеей не сводить к минимуму неоднородное искажение сигнала в одной части набора данных, даже если такое искажение неизбежно в другой части набора данных.

Вторая особенность нашей процедуры (одно из нововведений настоящего исследования) заключается в том, что мы не помечаем как родственные формы с так называемым деривационным дрейфом. Согласно традиционному и почти повсеместно принятому подходу в качестве лексикостатистических совпадений рассматриваются такие формы, у которых основные полнозначные морфемы (а именно, корни) являются родственными друг другу, т. е. считаются восходящими к одному пракорню. Тем не менее, очевидна мысль, что по морфологическим основаниям истинные исторические когнаты могут быть замещены параллельными новыми образованиями, имеющими тот же корень, см. хорошее обсуждение в [14: 202-203], где феномен параллельной морфологической деривации называется «деривационным дрейфом». К сожалению, несмотря на формулирование проблемы, Чанг и др. не предпринимают попыток предложить критерии

для выявления деривационного дрейфа и применить их в своей модели. Формальные критерии деривационного дрейфа, используемые в нашем текущем исследовании, обсуждаются ниже.

На первый взгляд, любые две основы (из разных лектов), чьи корни родственны, тогда как аффиксальная структура различается, следует рассматривать как результат параллельной эволюции, поскольку эти основы не происходят от общей праосновы. Тем не менее, такой строгий критерий исключал бы множество случаев, когда вовлеченные основы должны рассматриваться как истинные когнаты исходя из здравого смысла. Возьмем в качестве примера единицу "сердце" из нашего набора данных. Следующие основы, обозначающие "сердце" в индоевропейских языках, восходят к одному и тому же ПИЕ корню (слова приводятся в nom. sg.): хеттское *kir* (<праанатолийское *kʸer <*kʸerd, корневое существительное), древнегреческое *kard-í-aː* (форма с нулевой ступенью, дополненная суффиксом *-i*), прагерманское **xert-oːn* (форма с полной ступенью, дополненная суффиксом *-on*). Наиболее вероятный сценарий, принятый экспертами в этой области, заключается в том, что исходное праиндоевропейское корневое существительное * kʸeːr / * kʸr̥d- (сохранившееся в анатолийском) впоследствии было модифицировано в отдельных подгруппах путем добавления различных суффиксов с целью выравнивания парадигмы. Невозможно не рассматривать пару хеттское *kir*/ греческое *kard-í-aː* как исторически истинные когнаты. То же касается пары хеттское *kir* / германское **xert-oːn*, которые также должны быть помечены как истинные когнаты. В отсутствие праанатолийского корневого существительного третья пара, греческое *kard-í-aː* / германское**xert-oːn*, действительно выглядела бы подозрительной, и можно было бы предложить анализировать kard-í-aː и *xert-oːn как независимые гомопластические производные от некоторого корня *kʸerd- с другим (неизвестным) значением, но мы видим, что это решение было бы неправильным.

Таким образом, необходимы более сложные алгоритмы для выявления деривационного дрейфа, т. е. таких случаев, когда две лексемы из разных лектов обладают одним и тем же значением и имеют общий корень, но фактически представляют собой параллельные эволюционные события с лексикостатистической точки зрения. Для настоящего анализа мы предлагаем два формальных критерия, которые могут обнаружить значительное число случаев деривационного дрейфа.

Первый критерий деривационного дрейфа. Если два элемента из сопоставляемых лектов имеют общий корень, но различаются аффиксальной структурой, и есть свидетельства в пользу того, что хотя бы одна из основ подверглась части речевых

изменений (например, существительное ↔ глагол), эти основы, скорее всего, демонстрируют гомопластическое развитие.

Примеры. В некоторых ИЕ лектах адъективные термины для значения "теплый" образованы от глагола * tep- "быть теплым / горячим (vel sim.)": древнеирландское tʸee {теё}, прабритонское * teːmː, праславянское *tep-l-. Однако в каждом случае изменение части речи "глагол → прилагательное" происходило с использованием разных суффиксов, а именно: * -nt- (древнеирландский), * -smo- (британский), * -lo- (славянский). Все три обсуждаемых формы представляют собой, скорее всего, полноценные лексикостатистические события, будучи результатом параллельного словообразования. Другим примером является глагол "умереть" в британских языках. Самый распространенный глагол для значения "умирать", засвидетельствованный во внутренних индоевропейских языках, - это *mer-: латинское mor-, древнеиндийское mar-, праславянское *mer- и т. д. Для прабритского можно восстановить глагол *marw- "умирать", который представляет собой отыменное образование от прабритского прилагательного *marw "мертвый". Последнее происходит от *mr̥-wo-, содержащее в конечном счете тот же корень *mer-, модифицированный суффиксом прилагательного. Британский глагол подвергся частеречному изменению "глагол → прилагательное → глагол", и интуитивно кажется вероятным, что сдвиг "мертвый" → "умирать" представляет собой полноценное лексикостатистическое событие.

Конечно, во многих языках мира противопоставление частей речи является слабым, если вообще существует. Естественно, что к таким языкам обсуждаемый критерий неприменим. Но для языков, обладающих явным противостоянием между частями речи, предлагаемый критерий достаточно силен. Кросс-лингвистически наиболее базовой оппозицией является "существительное: глагол", тогда как прилагательные склонны иметь сходство либо с существительными, либо с глаголами. Хотя праиндоевропейские прилагательные ближе к существительным, они достаточно отличаются от них, чтобы считаться отдельным классом слов. Таким образом, ПИЕ прилагательные градуируемы и демонстрируют сложную систему суффиксальной подстановки, известную как «Каландова система» [15: 2113-2118].

Второй критерий деривационного дрейфа. Если два элемента из сравниваемых лектов имеют общий корень, но модифицированы с помощью разных аффиксов, и есть свидетельства в пользу того, что эти основы были образованы от более простой основы, семантика которой сильно отличалась от значений сопоставляемых основ, такие две основы скорее всего представляют собой гомопластическое развитие.

Пример. В латинском, балтийском и кельтском языках обозначения для "человека"

образованы от индоевропейского термина, обозначающего "земля" (т. е. "человек" как "землянин"), но с разными суффиксами: *-on в латинском (hom-in -) и в прабалтийском (*žm-un-) и *-yo- в пракельтском (*gdon-yo-). Латинская и балтийская формы, с одной стороны, и кельтская форма, с другой стороны, представляют собой скорее всего два различных лексикостатистических события, будучи результатом параллельного словообразования.

В настоящем исследовании мы применяем оба критерия к входному набору данных, чтобы очистить его от большинства случаев деривационного дрейфа.

С технической точки зрения исходный лексический набор данных представляет собой матрицу, которая, во-первых, является многоуровневой, то есть единицы могут иметь более двух уровней, во-вторых, включает синонимы. Синонимия неизбежна для лексических данных, технически это означает более одного слова для одного сводешовского концепта. Эта многоступенчатая матрица используется в пакете Starling [16: 271-274], единственном программном обеспечении, способном обрабатывать входные матрицы, содержащие синонимы.

Вторая матрица, используемая для других филогенетических пакетов, является бинарной. Бинарная матрица преобразуется из многоуровневого набора данных путем кодирования наличия ("1") или отсутствия ("0") конкретного пракорня с заданным сводешовским значением в соответствующем языке [4, 12]. Единицы из списка Сводеша, не задокументированные для данного языка или замещенные заимствованиями, помечаются как "?".

3 Укоренение деревьев

Ручное укоренение необходимо, по крайней мере, для анализа методом максимальной экономности. Выбор внешнегруппового таксона является в нашем случае нетривиальным вопросом, поскольку у индоевропейской семьи нет лингвистических родственников, которые были бы достаточно близки, условно приняты и представляли бы, таким образом, наиболее подходящую внешнюю группу. В свете этого мы предпочитаем дублировать наш анализ, используя два разных выброса: хеттский (индоевропейская семья) и прасамодийский (одна из двух первичных ветвей уральской семьи).

Так, для первого анализа в качестве внешней группы был взят хеттский язык (в качестве ограничения он использовался только для деревьев с максимальной экономностью). Хеттский относится к анатолийской группе индоевропейской семьи, и эксперты практически единодушны в том, анатолийская ветвь - первая, отделившаяся от индоевропейской семьи [17, 18].

Во втором анализе в качестве внешней группы мы вводим прасамодийский (в качестве ограничения он использовался только для деревьев с максимальной экономностью). Самодийская группа [19] состоит из нескольких близкородственных языков и представляет одну из двух основных ветвей в уральской лингвистической семье. Среди известных языков и групп прауральский, по-видимому, является ближайшим сестринским таксоном для праиндоевропейского, хотя это родство относительно дальнее и сопоставимо с таковым современных ИЕ языков (скажем, между современными немецким и греческим языками). Несмотря на то, что индо-уральская гипотеза имеет долгую историю (см. обзор и статистические данные в пользу ИЕ-уральской клады в [11]), она далеко не общепринята. Тем не менее, общая тенденция среди индоевропеистов позитивна, ср., например, недавнюю конференцию «Прекурсоры праиндоевропейского: индо-хеттская и индо-уральская гипотезы», состоявшуюся в Лейденском университете 9-11 июля 2015.

Следующие 10 прасамодийских единиц мы считаем этимологическими когнатами их праиндоевропейских эквивалентов: **ɔ-r-* “пить”, **tɒ-* “давать”, **mə̆-n* “я”, **kɨy-tV-* “лежать”, **nim* “имя”, **ta-* “тот”, **tə̆-* “этот”, **tə̆-n* “ты”, **wet* “вода”, **ke-* “кто”. Восемь из них подробно рассматриваются в [11]. Девятый корень, **tɒ-* “давать” (< прауральское **toɣi-)*, очевидно следует сравнивать с ИЕ **do꞉-* или **deh*$_3$- “давать”, с **-ɣ-* в качестве эквивалента ИЕ “ларингала”. Десятую единицу **kɨy-tV-* “лежать” (< прауральское **kuyi-*) можно надежно сравнить с ИЕ **k^{y}ey-* “лежать”, потому что последующий **y* препятствует правилу Иллич-Свитыча индо-уральское **Ku* > ИЕ **K^{w}* [20].

Построение дерева

Мы в целом следуем процедуре, предложенной в [13: 255-257]. Она состоит из нескольких последовательных шагов:

1) После составления высококачественного лексического набора данных (Этап-1, wind = ветер) и исключения деривационного дрейфа (Этап-2, wind ≠ ветер) деревья строятся с помощью нескольких филогенетических алгоритмов (Starling neighbor-joining, Bayesian MCMC, Maximum Parsimony).

2) Компилируется консенсусное дерево.

3) Исходя из топологии полученного консенсусного дерева и реконструированных состояний предковых единиц, мы исследуем входной набор данных на предмет гомопластических развитий, т.е. ищем единицы, которые несовместимы с топологией консенсусного дерева. Лексические единицы, опознанные как составляющие такие гомопластические соответствия, обозначаются как этимологически несвязанные или, если мы уверены в направлении влияния, целевая единица может быть помечена как заимствование. Эта процедура называется гомопластической оптимизацией (Этап-3, agni ≠ ignis). Стоит обратить внимание, что для гомопластической оптимизации требуется не только предопределенное дерево, но и семантическая реконструкция слов из списка Сводеша для праиндоевропейского и промежуточных праязыков. Для семантической реконструкции мы используем методологию, предложенную в [11: 304-306]. См. в Приложении лингвистические замечания по отдельным случаям гомоплазии в нашем наборе данных.

4) Оптимизированные в отношении гомоплазии деревья перестраиваются с помощью индивидуальных алгоритмов (StarlingNJ, Bayesian MCMC, MP).

5) Полученные деревья, оптимизированные по гомоплазии, суммируются как дерево консенсуса, которое является конечным результатом нашего исследования.

Лексикостатистические деревья были получены с помощью следующих филогенетических методов.

– Присоединение соседей по Starling - Starling neighbor-joining, далее StarlingNJ [16: 163–167]. Деревья по StarlingNJ были получены в программном обеспечении Starling v.2.5.3 [16: 271-274] из лексикостатистической базы данных, представляющей собой многоэтапную матрицу с разрешенной синонимией. Разрешенная синонимия означает, что, когда в данном языке одно и то же поле списка Сводеша занято более чем одним словом, то есть несколькими синонимами, каждое слово из такого поля сравнивается с каждым словом из того же поля в другом языке, так что сравниваются все возможные пары слов между двумя

языками: если есть хотя бы одна пара соответствий, все поле рассматривается как соответствие. Для узловых датировок применялся так называемый «экспериментальный метод», согласно которому каждая сводешовская единица обладает индивидуальным относительным индексом стабильности [21]. Был проведен непараметрический тест начальной загрузки (10 000 псевдорепликаций). Иерархическая агломеративная кластеризация дает корневое дерево по определению (последнее слияние - это корень, оно совпадает со средней точкой при предположении о почти равномерной скорости замещения). Даты узлов были установлены строгими часами, см. [22] о шкале калибровки и более подробно. Соседние узлы объединяются в единый узел, если временное расстояние между ними составляет 300 лет или меньше (300 лет соответствует мутации около 1,5 слов в лекте - разумная ошибка расчета, хотя этот временной интервал во многом произволен на текущей стадии исследования). Деревья были визуализированы в Starling, а затем вручную перерисованы для презентации.

– Симуляция метода Монте-Карло для марковских цепей в рамках байесовского подхода (далее Bayesian MCMC), см. [23: 68–69]. Деревья были получены в программном обеспечении MrBayes v.3.2.6 [24] из бинарной матрицы, описанной выше. Мы использовали модель F81, где datatype = restriction, coding = noabsencesites, rates = gamma, brlenspr = clock:fossilization, clockvarpr = TK02 (автосоотнесенные свободные часы), см. полный набор параметров MrBayes в Приложении и далее см. подробное обсуждение параметров MrBayes, подходящих для лингвистической филогении, в Yanovich forthc. Для собственно ИЕ набора данных диапазон возраста корня строго предопределен как 10 500-5 500 лет до н.э. Для набора ИЕ-самодийских данных диапазон возраста корня задан как offsetexp (10000,20000) с верхней границей 10 000 лет до н.э. и средним значением 20 000 лет до н.э., см. обсуждение датировки в Приложении. Не было установлено внешних групп или других топологических ограничений. Программа запускалась 4 раза с использованием 4 параллельных цепей Маркова. В каждом прогоне было получено 10 000 000 генераций деревьев, образцы были взяты через каждые 500 генераций. Для каждого прогона первые 25% порожденных деревьев были отбракованы как пробные. Для датированных консенсусных деревьев программой были определены корни. Деревья были визуализированы в программном обеспечении FigTree (v.1.4.3).

– Невзвешенный метод максимальной экономности (maximum parsimony, далее MP), см. [23: 66-67]. Деревья были произведены в программном обеспечении

TNT (издание Willi Hennig Society of TNT, v.1.5, март 2017, см. [25]) из двоичной матрицы, описанной выше, с помощью алгоритма ветвей и границ («имплицитного перечисления»). Обязательная бинаризация узлов была запрещена («Свернуть деревья после поиска»). Хеттский и прасамодийский были помечены как внешняя группа для собственно ИЕ и ИЕ-самодийского наборов данных, соответственно. Когда был получен набор оптимальных деревьев равной ценности (плюс субоптимальные деревья с шагом 1, если возвращается единственное оптимальное дерево), было произведено консенсусное дерево пр принципу большинства, для которого был выполнен непараметрический тест начальной загрузки (1000 псевдорепликаций). Деревья не датированы. Деревья были визуализированы в программном обеспечении FigTree v.1.4.3).

4 Результаты

Набор данных Этапа-1 (корневое родство): Для каждого набора данных – списков слов с корневым родством (Этап-1), списков слов, очищенных от деривационного дрейфа (Этап-2), списков слов, оптимизированных по гомоплазии (Этап-3) – были получены следующие деревья с прасамодийским таксоном и без него:

– Рис. 1, метод StarlingNJ с присоединенными соседними узлами.

– Рис. 2, метод байесовского MCMC.

– Рис. 3, метод максимальной экономии.

Деревья, основанные на списках слов с корневым родством (wind = veter), не сильно противоречат нашим ожиданиям. Анатолийский и тохарский корректно распознаются как последовательно отделившиеся. Недавние общепринятые клады - островная кельтская, балто-славянская, индо-иранская и греко-армянская - корректно распознаны в большинстве деревьев. Тем не менее, общий результат не может считаться полностью удовлетворительным, поскольку деревья противоречат друг другу в некоторых важных узлах. Такие расхождения существуют не только между деревьями, полученными разными методами, но и между деревьями, полученными одним методом на собственно ИЕ и ИЕ-самодийском наборах данных.

Набор данных Этапа-2 (без деривационного дрейфа): Результаты анализа набора данных без деривационного дрейфа (wind ≠ ветер, agni = ignis) более перспективны. Деревья по StarlingNJ и по Байесу топологически почти идентичны. Оба метода предполагают два последовательных выброса (анатолийский и тохарский) и разделение внутренней ИЕ семьи на четыре ветви независимо от того, используется ли собственно ИЕ набор данных или же ИЕ-самодийский: (1) греко-армянская, (2) албанская (3)) итало-германо-кельтская, (4) балто-славяно-индоиранская ветви.

Разница между методами StarlingNJ и Байеса заключается в том, что дерево по StarlingNJ членит западноевропейскую кладу как [[италийский, германский] кельтский], что может иметь исторический смысл, тогда как байесовское дерево показывает тройной раскол [италийский, германский, кельтский].

Консенсусные MP-деревья MP по принципу большинства обычно создают одну и ту же топологию, хотя содержат больше бинарных узлов, чем деревья, построенные другими методами. Следует обратить внимание, что дерево MP, основанное на собственно ИЕ наборе данных, предлагает такую же детальную структуру для западноевропейской клады: [[италийский, германский] кельтский].

Единственное существенное расхождение между этими методами касается италийской, германской и кельтской групп в дереве MP, основанном на ИЕ-самодийском

наборе данных. Эти группы не образуют четкой клады (сначала кельтский, а затем итало-германский последовательно отделяются от ствола). Поскольку этот результат не вполне ясен, мы предпочитаем игнорировать его при компиляции консенсусного дерева.

Два метода дают датированные деревья: StarlingNJ (строгие даты) и Bayesian MCMC (плотность максимальной вероятности (highest probability density, далее *HPD*) 95% для времени расхождения и средних дат расхождения). Разница суммируется в Таб. 2. Для байесовского метода мы ссылаемся на дерево, основанное на собственно ИЕ наборе данных (его даты лишь незначительно отличаются от дат ИЕ-самодийского дерева).

Таблица 12 - Расхождения в датах, полученных для набора данных без деривационного дрейфа Этапа-2 (wind ≠ ветер, agni = ignis). 95% HPD и среднее значение для Байесовского MCMC, строгие даты для StarlingNJ.

	Байесовский MCMC	StarlingNJ
Отделение анатолийского	4314–3450 BC (mean 3747 BC)	5080 BC
Отделение тохарского	3821–2099 BC (mean 2974 BC)	4700 BC
Распад внутри-ИЕ	3572–2145 BC (mean 2802 BC)	4100 BC
Распад греко-армянского	2747–1264 BC (mean 1986 BC)	3460 BC
Распад итало-германо-кельтского	2825–1443 BC (mean 2128 BC)	3500 BC
Распад островного кельтского	605 BC – 138 AD (mean 217 BC)	1500 BC
Распад балто-славянских и индо-иранских	2933–1847 BC (mean 2366 BC)	3570 BC
Распад балто-славянского	1807–882 BC (mean 1331 BC)	2390 BC
Распад индо-иранского	2100–1447 BC (mean 1763 BC)	2230 BC

Как следует из таблицы 2, для набора данных Этапа-2 без деривационного дрейфа даты по StarlingNJ существенно глубже, чем байесовские. Кроме того, полученные в настоящее время даты StarlingNJ глубже, чем даты StarlingNJ, полученные для индоевропейской семьи в некоторых предыдущих исследованиях. Например, отделение анатолийского датируется 4340 годом до нашей эры, отделение тохарского датируется 3870 годом до нашей эры (на основе 50-словных списках для реконструированных праязыков ИЕ подгрупп). Альтернативно, в [26: 85] эти бифуркации датируются 4670 годом до н.э. и 3810 г. до н.э. соответственно (на основе 110-словных списков засвидетельствованных языков). Хронологические расхождения между нашими текущими

расчетами по Starling и представленными ранее объясняются различиями в подготовке входных данных. Строгие семантические спецификации и некоторые общие принципы компиляции списка Сводеша были добавлены в наш арсенал только в 2010 году [10], а техники оптимизации гомоплазии (такие как элиминация деривационного дрейфа и т. д.) были введены совсем недавно. Напротив, алгоритм датирования, реализованный в Starling, был откалиброван гораздо раньше на основе «традиционных» списков слов Сводеша, т.е. списков, составленных без такого строгого семантического и прагматического контроля, не говоря уже об оптимизации гомоплазии.

В свою очередь, полученные байесовские даты показывают противоположную тенденцию: некоторые из них представляются слишком поздними. Действительно, диапазоны 95% HPD, приведенные в Таб. 2, в целом не противоречат нашим ожиданиям, но по крайней мере некоторые из средних дат несколько моложе, чем признается традиционными экспертами, например, индоиранский распад обычно датируется ранее 2000 года до н. э. [27] в отличие от 1753 г. до н.э. (средняя дата), что предлагается нашим анализом. С другой стороны, конец археологической культуры Синташты, иногда связываемый с праиндоиранскими языковыми носителями, датируется началом 18-го в. до н. э. [28]. Средняя дата распада островного кельтского - 221 г. до н.э., хотя топоним Βρεττανικη «Британия» со специфическим прабритонским развитием *k^w> p (> b) упоминался Пифеем из Массалии уже в 325 г. до н.э. (согласно Страбону). NB: лингвистам следует избегать использования байесовских средних (или медианных) дат в качестве строгих, так как в конце концов ожидается, что только полученные 95% интервалов HPD должны иметь исторический смысл.

Все эти вопросы, касающиеся методов датирования, требуют детального изучения, которое должно быть предметом дальнейших исследований. Для консенсусного дерева в текущем исследовании (рис.1) мы принимаем байесовские даты как более гибкие.

Для набора данных без деривационного дрейфа (Этап-2, wind ≠ ветер, agni = ignis) мы суммируем полученные топологии по StarlingNJ, Bayesian MCMCM и MP в качестве консенсусного дерева (рис. 1). Консенсус строгий, за исключением парафилетической топологии италийской, германской и кельтской подгрупп в дереве MP, основанном на ИЕ-самодийском наборе данных; следует обратить внимание на то, что дерево MP для собственно ИЕ набора данных лишено этого недостатка, будучи в соответствии с деревьями StarlingNJ и Bayesian MCMC. На самом деле консенсусное дерево Этапа-2 (рис.1) идентично байесовскому дереву.

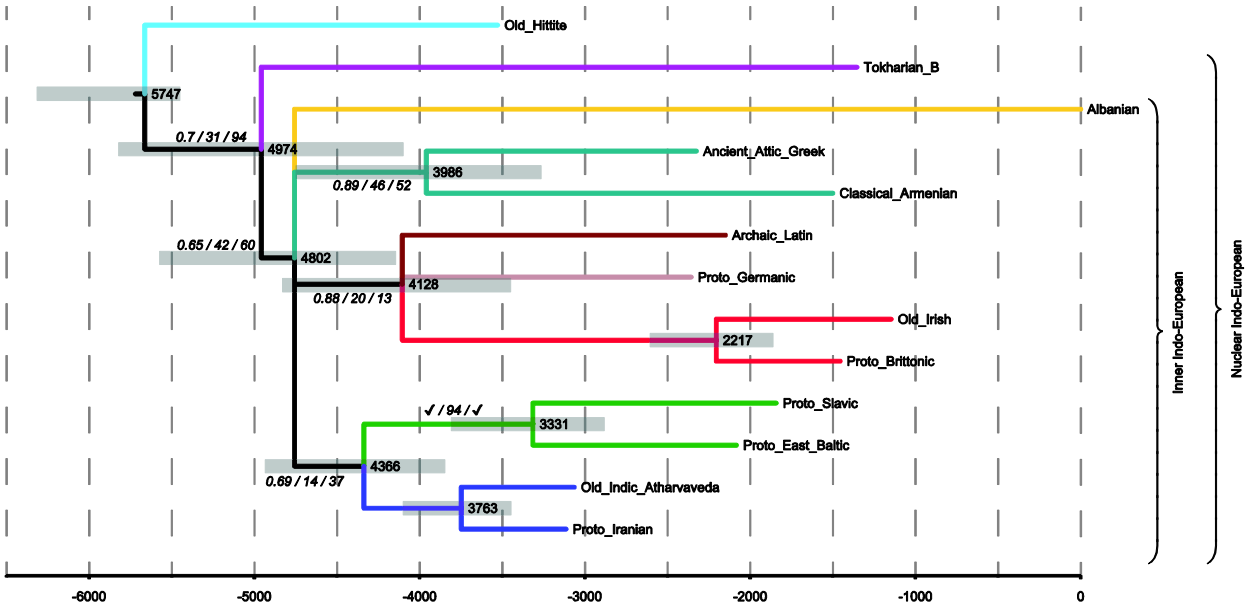

Рисунок 1 - Строго (кроме итало-германо-кельтской парафилии в одной из топологий MP) консенсусное дерево ИЕ семьи, основанное на наборе данных Этапа-2 без деривационного дрейфа (wind ≠ ветер, agni = ignis). Дерево суммирует шесть деревьев, полученных отдельными методами. Даты получены с помощью байесовского MCMC-анализа: серые полосы представляют собой 95% плотность максимальной вероятности (HPD) для времен расхождения; справа от каждого узла дается среднее время расхождений. Значения шкалы представляют годы до настоящего времени (yBP). Значения статистической поддержки показаны курсивом рядом с ветвями в следующей последовательности: Bayesian MCMC / StarlingNJ / MP («✓» означает, что P ≥ 0,95 по отдельному методу; не показывается для узлов с P ≥ 0,95 во всех методах). Традиционные подгруппы обозначаются цветными ветвями.

Набор данных Этапа-3 (оптимизировано по гомоплазиям):

Следующим шагом в подготовке вводных данных является гомопластическая оптимизация [13]. Мы обозначаем этот шаг как Этап-3 (agni ≠ ignis). Исходя из консенсусного дерева Этапа-2 (рис.1) и нашей методологии семантической реконструкции [11: 304-306], мы рассматриваем случаи, когда два или более корня находятся в перекрестной конфигурации, т.е. по крайней мере один из них нарушает топологию дерева: так называемые несовместимые единицы. Если у нас есть свидетельства в пользу того, что один из конкурирующих корней представляет собой архаизм, тогда как другой, вероятно, является параллельной инновацией в отдельных подгруппах, мы отмечаем рефлексы второго корня как неродственные. Примером может служить древнеиндийское agni и латинское ignis, которые оба обозначают "огонь" и являются непосредственными этимологическими когнатами и, таким образом, помечаются одним индексом родства (а.и. классом когнатов) в наборе данных Этапа-2, но получают разные индексы в наборе данных Этапа-3 (см. в Приложении лингвистический комментарий).

Следующие сводешовские концепты затрагиваются гомопластической оптимизацией в наборе данных Этапа-3: "живот", "грудь", "приходить", "есть", "огонь", "слышать", "врать", "человек", "много", "ночь", "человек", "видеть", "сидеть", "стоять".

В следующих случаях гомоплазия не может быть устранена из-за скудости данных и невозможности предложить надежную семантическую реконструкцию: "один", "видеть", "это", "этот", "зуб", "далеко", "змея", "год".

Наконец, существует ряд сводешовских концептов, где мы реконструируем пра-ИЕ супплетивную парадигму, которая может быть упрощена одинаковым образом независимо друг от друга в разных ветвях, что приводит к перекрестной конфигурации: "идти", "видеть", "идти", "Я", "мы". Как можно видеть, они представляют собой либо базовые глаголы, либо личные местоимения, т.е. категории, для которых супплетивизм кросс-лингвистически является типичным.

Деревья, основанные на наборе данных Этапа-3 (с оптимизацией по гомоплазиям) (*wind ≠ ветер, agni ≠ ignis*) и полученные индивидуальными методами, представлены в Дополнении. Деревья в целом не противоречат друг другу (в частности, итало-германо-кельтская клада теперь наблюдается во всех деревьях) и совместимы с деревьями, полученными на основе набора данных Этапа-2 (wind ≠ ветер, agni = ignis). Новый результат заключается в том, что албанский теперь объединен с балто-славяно-индо-иранским в отдельную кладу в обоих байесовских деревьях. С одной стороны, это не противоречит традиционным взглядам, поскольку мы слишком мало знаем об истории албанского. С другой стороны, статистическая поддержка для клады [албанский [балто-славянский, индоиранский]] очень слабая: 0,54 и 0,59 в зависимости от того, включен или нет в набор данных прасамодийский. Клада [албанский [балто-славянский, индоиранский]] также наблюдается в дереве StarlingNJ с бинарными узлами, хотя временная шкала слишком коротка, и албанский оказывается отдельной ветвью в дереве StarlingNJ, когда соседние узлы соединяются при расстоянии между ними ≤ 300 лет. Напротив, албанский оказывается первым, отделившимся от внутреннего ИЕ дерева в деревьях MP.

Даты, полученные для набора данных Этапа-3, оптимизированных по гомоплазиям, суммированы в табл. 3. Они схожи с датами для набора данных Этапа-2 (табл. 2) даты StarlingNJ снова выглядят слишком ранними, тогда как верхний предел и среднее значение байесовских интервалов 95% HPD кажутся слишком поздними.

Таблица 3 - Несоответствия в датах, полученных для набора данных, оптимизированных для гомоплазии Stage-3 (*wind ≠ ветер, agni ≠ ignis*). 95% HPD и среднее значение для байесовского MCMC, строгие даты для StarlingNJ.

	Байесовский MCMC	StarlingNJ
Отделение анатолийского	4139–3450 BC (mean 3686 BC)	5110 BC
Отделение тохарского	3727–2262 BC (mean 3011 BC)	4710 BC
Распад внутри-ИЕ	3357–2162 BC (mean 2717 BC)	4150 BC
Распад греко-армянского	2676–1407 BC (mean 2015 BC)	3460 BC
Распад итало-германо-кельтского	2655–1537 BC (mean 2080 BC)	3540 BC
Распад островного кельтского	596 BC – 95 AD (mean 243 BC)	1570 BC
Распад балто-славянских и индо-иранских	2723–1790 BC (mean 2241 BC)	3570 BC
Распад балто-славянского	1686–855 BC (mean 1250 BC)	2450 BC
Распад индо-иранского	2044–1458 BC (mean 1740 BC)	2230 BC

Для набора данных Этапа-3, оптимизированных по гомоплазии (wind ≠ veter, agni ≠ ignis), мы суммируем топологии, полученные методами StarlingNJ, Bayesian MCMCM и MP, как строгое консенсусное дерево Рис. 2. Мы принимаем байесовские даты как более гибкие.

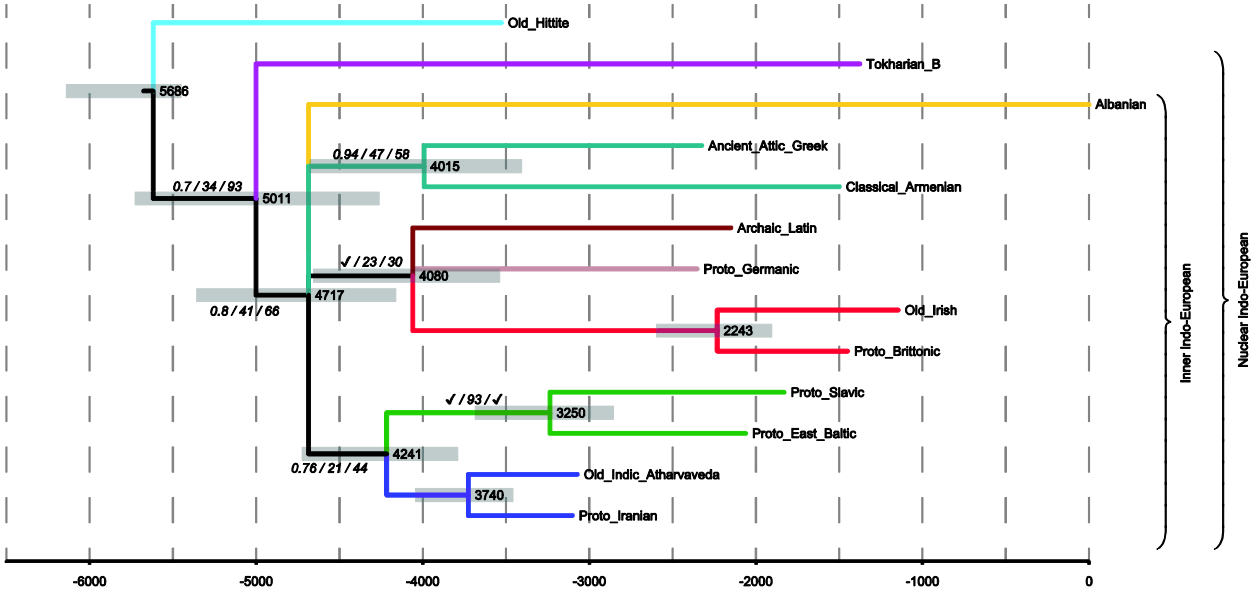

Рисунок. 2 - Строгое консенсусное дерево ИЕ семьи на основе набора данных Этапа-3 с оптимизацией по гомоплазиям (wind ≠ veter, agni ≠ ignis). Дерево суммирует шесть деревьев, полученных индивидуальными методами. Даты получены с помощью анализа Bayesian MCMC: серые полосы представляют собой плотность максимальной вероятности 95% (HPD) для времени расхождения; справа от каждого узла указывается среднее время расхождения. Значения шкалы представляют годы до настоящего времени (yBP). Значения статистической поддержки указаны курсивом рядом с ветвями в следующей последовательности: Bayesian MCMC / StarlingNJ / MP («✓» означает, что P ≥ 0,95 по отдельному методу, не указано для узлов с P ≥ 0,95 во всех методах). Традиционные подгруппы обозначаются цветными ветвями.

5 Обсуждение

Мы разрабатываем три набора данных, которые являются последовательными преобразованиями друг друга, и не используем никаких топологических ограничений для анализа.

Деревья, полученные для набора данных Этапа-1 с традиционным корневым родством (*wind = ветер, agni = ignis*) не имеют серьезных противоречий традиционным взглядам (за исключением албанско-балто-славянско-индо-иранской клады).

Тем не менее деревья, полученные для набора данных Этапа-2 без деривационного дрейфа (wind ≠ ветер, agni = ignis консенсусное дерево на Рис.1) намного лучше соответствуют традиционным представлениям экспертов, Это говорит о том, что предложенная выше формальная процедура исключения деривационного дрейфа является мощным и важным методом, помогающим улучшить итоговую филогению.

Деревья, полученные для набора данных Этапа-3, оптимизированных по гомоплазии (wind ≠ veter, agni ≠ ignis, консенсусное дерево на Рис. 2), сообщают мало нового в сравнении с деревьями Этапа-2, но в соответствии с теоретическими ожиданиями оптимизация гомоплазий делает итоговую топологию более надежной.

Все методы порождают схожие топологии независимо от того, включен прасамодийский список или нет. Во-первых, анатолийский и тохарский всегда распознаются как два последовательных выброса, что означает, что внутренние индоевропейские языки образуют четкую кладу. Во-вторых, все поздние подгруппы корректно распознаются как отдельные клады: греко-армянская (отсутствует в одном из MP-деревьев), ирландско-британский (т.е. островной кельтский), балто-славянский, индоиранский. Все эти характеристики идеально соответствуют традиционным представлениям экспертов об ИЕ семье. Следует отметить, что некоторые лингвисты отрицают балто-славянский узел, предлагая объяснять сходство между балтийским и славянским за счет интенсивных доисторических контактов. Однако, исходя из теории языковых контактов, мы не видим оснований для контактного сценария (интересно, что, насколько нам известно, авторы, которые отвергают близкую генетическую связь между балтийским и славянским, никогда не предлагают свою филогенетическую концепцию в виде дерева, предпочитая описывать ее словами).

Наряду с вышеупомянутыми поздними кладами, которые единодушно выделяются экспертами данной области, наша итоговая топология (рис.2) предлагает две группы более высокого уровня: итало-германо-кельтскую и балто-славяно-индо-иранскую. Эти группы иногда обсуждаются индоевропеистами. Интересно, что два из трех алгоритмов -

StarlingNJ и Maximum parsimony - членят западноевропейскую кладу как [[италийский, германский] кельтский], что не может быть исключено с исторической точки зрения, хотя байесовский анализ MCMC дает трехчастное расхождение [италийский, германский, кельтский].

Наиболее интересным и важным результатом является мультифуркация внутреннего ИЕ на четыре ветви: (1) греко-армянскую, (2) итало-германо-кельтскую, (3) балто-славяно-индо-иранскую, (4) албанскую. Это может показаться парадоксальным, но на самом деле такое расхождение основных внутренних ИЕ ветвей также идеально соответствует традиционным представлениям экспертов. Действительно, большинство индоевропеистов, если не все из них, согласны с внешним статусом анатолийского и тохарского и с наличием поздних клад, таких как индоиранская или балто-славянская. Но что находится в середине ИЕ дерева? Несмотря на более чем столетнее интенсивное развитие индоевропейских исследований, до сих пор нет консенсуса или, по крайней мере, ведущего мнения о том, как могло бы выглядеть раннее ветвление внутренних ИЕ языков. Это отсутствие консенсуса вытекает из отсутствия надежных общих инноваций, разделяемых некоторыми подмножествами внутренних ИЕ ветвей, что могло бы помочь выявить раннюю топологию данной клады. В такой ситуации сценарий мультифуркации, предложенный в нашем анализе, является наиболее естественным и вероятным решением.

Неудивительно, что наиболее проблематичным таксоном является албанский, который перескакивает по дереву, занимая различные позиции внутри внутренней ИЕ клады: от первого выброса до третьего члена балто-славянско-индоиранской клады в зависимости от используемого набора данных и метода. Основными причинами такой нестабильности являются, во-первых, большое количество неисконных лексических единиц в албанском базовом словаре и, во-вторых, наше недостаточное знание албанской исторической фонологии: из-за скудности албанского исконного словаря мы можем упускать из виду некоторые нетривиальные фонологические правила. В результате некоторые албанские основы, рассматриваемые здесь как этимологически изолированные, на самом деле могут быть истинными когнатами соответствующих ИЕ терминов.

Хронологические интервалы, полученные с помощью байесовского MCMC анализа и обобщенные в табл. 3, не противоречат представлениям экспертов. Например, первоначальная бифуркация ИЕ, отделение анатолийского, находится в пределах 4139-3450 г. до н.э., что совместимо с традиционными оценками.

Мы избегаем здесь обсуждения вопроса об ИЕ прародине, поскольку считаем, что полученные даты на самом деле могут мало сообщить о географическом распределении доисторических индоевропейцев.

6 Выводы

Подготовка входных данных — как сбор данных, так и последующая обработка, например, элиминация деривационного дрейфа и оптимизация гомоплазий — играют решающую роль в лингвистической филогении. В частности, мы считаем, что неточные лингвистические данные могут быть виной неверных топологических результатов, полученных в некоторых предшествующих исследованиях по индоевропейской филогении. Первоначальная мультифуркация внутренних индоевропейских языков на четыре первичные ветви, насколько нам известно, подтверждается формальными методами впервые. Это согласуется с имеющимися лингвистическими свидетельствами и примиряет друг с другом различные мнения об ИЕ филогенезе, высказывавшиеся индоевропеистами начиная с Августа Шлейхера в середине XIX века до современных авторитетных справочников.

Мы выражаем благодарность Игорю Яновичу (DFG Center for Advanced Study “Words, Bones, Genes and Tools”, Университет г. Тюбингена) и Валерию Запорожченко (Медико-генетический научный центр РАМН) за помощь в вычислительном анализе.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Schleicher, August. Compendium der vergleichenden Grammatik der indogermanischen Sprachen. 1861. Weimar: H. Böhlau.
2. Jasanoff, Jay H. Hittite and the Indo-European verb. 2003. Oxford: Oxford University Press.
3. Rexová, Kateřina, Daniel Frynta, Jan Zrzavý. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. 2003. Cladistics 19, P. 120–127.
4. Gray R. D., Atkinson Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. 2003. Nature, № 426, P. 435-439.
5. Ringe D, Warnow T, Taylor A. Indo-European and computational cladistics. 2002. Trans Philol Soc. 100, P. 59–129.
6. Nakhleh L, Warnow T, Ringe D, Evans SN. A comparison of phylogenetic reconstruction methods on an Indo-European dataset. 2005. Trans Philol Soc. 103, P. 171–192.
7. Bouckaert, R., P. Lemey, M. Dunn, S. J. Greenhill, A. V. Alekseyenko, A. J. Drummond, R. D. Gray, M. A. Suchard, Q. D. Atkinson. Mapping the Origins and Expansion of the Indo-European Language Family. 2012. Science 337, P. 957–960.
8. Kushniarevich, A., Utevska, O., Chuhryaeva, M., Agdzhoyan, A., Dibirova, K., Uktveryte, I., Möls, M., Mulahasanovic, L., Pshenichnov, A., Frolova, S., Shanko, A., Metspalu1, E., Reidla, M., Tambets, K., Tamm, E., Koshel, S., Zaporozhchenko, V., Atramentova, L., Kučinskas, V., Davydenko, O., Goncharova, O., Evseeva, I., Churnosov, M., Pocheshchova, E., Yunusbayev, B., Khusnutdinova, E., Marjanović, D., Rudan, P., Rootsi, S., Yankovsky, N., Endicott, P., Kassian, A., Dybo, A., The Genographic Consortium, Tyler-Smith, C., Balanovska, E., Metspalu1, M., Kivisild, T., Villems, R., Balanovsky, O. Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. PLOS One, September 2, 2015.
9. Rama, Taraka, Søren Wichmann. Towards identifying the optimal datasize for lexically-based Bayesian inference of linguistic phylogenies. In: Emily M. Bender, Leon Derczynski, Pierre Isabelle (eds.). Proceedings of the 27th International Conference on Computational Linguistics, 1578–1590. Santa Fe: Association for Computational Linguistics, 2018.
10. Kassian, A., G. Starostin, A. Dybo,, V. Chernov. The Swadesh wordlist. an attempt at semantic specification. Journal of Language Relationship 4, 46–89, 2010.
11. Kassian, Alexei, Mikhail Zhivlov, George Starostin. Proto-Indo-European-Uralic comparison from the probabilistic point of view. Journal of Indo-European Studies 43(3-4): 301-347, 2015.

12. Atkinson Q. D., Gray R. D. How old is the Indo-European language family? Progress or more moths to the flame? In: Phylogenetic methods and the prehistory of languages (eds Forster P., Renfrew C.), Cambridge, UK: McDonald Institute for Archaeological Research, P. 91–109, 2006.

13. Kassian, Alexei S. Linguistic homoplasy and phylogeny reconstruction. The cases of Lezgian and Tsezic languages (North Caucasus). Folia Linguistica Historica 38: 217–262, 2017.

14. Chang, Will & Cathcart, Chundra & Hall, David & Garrett, Andrew. "Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis." Language, vol. 91 no. 1, P. 194-244, 2015.

15. Lundquist, Jesse & Yates, Anthony D. The morphology of Proto-Indo-European. In: Jared Klein, Brian Joseph, Matthias Fritz (Eds.), Handbook of Comparative and Historical Indo-European Linguistics, P. 2079–2195. Berlin, Boston: De Gruyter, 2015.

16. Бурлак С. А., Старостин С. А. Сравнительно-историческое языкознание. Москва, Academia, 2005, 432 с.

17. Gamkrelidze, Thomas. V., Ivanov, Vjačeslav V. Indo-European and the Indo-Europeans. A Reconstruction and Historical Analysis of a Proto-Language and a Proto-Culture. Transl. by Johanna Nichols. Ed. by Werner Winter. Mouton de Gruyter. Berlin/New York, 1995.

18. Winter, Werner. Lexical archaisms in the Tocharian languages. In: H. H. 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, 1996.

19. Hajdú, Péter. Die Samojedischen Sprachen. In: Sinor, Denis (ed.), The Uralic Languages. Leiden: Brill, P. 3-40, 1988.

20. Живлов М. А. Отражение ностратических огубленных гласных в индоевропейском. В: Памяти В. М. Иллич-Свитыча: материалы круглого стола. Москва: Институт славяноведения РАН, 2017.

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

22. Novotná, Petra & Blažek, Václav. On lexicostatistic classification of the Frisian dialects. Linguistica Brunensia: Sborník prací filozofické fakulty brněnské univerzity A55, P. 115-132, 2007.

23. Vladimir Makarenkov, Dmytro Kevorkov, Pierre Legendre. Phylogenetic Network Construction Approaches. In: Dilip K. Arora, Randy M. Berka, Gautam B. Singh (eds.). Applied Mycology and Biotechnology. Vol. 6: Bioinformatics. Elsevier, 2006. P. 61—98.

24. Huelsenbeck, J.P., Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17, 2001, P. 754-755.

25. Goloboff, Pablo A., Santiago A. Catalano. TNT, version 1.5, with a full implementation of phylogenetic morphometrics. Cladistics 32: 221-238, 2016.

26. Blažek, V. 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: 82–109, 2007.

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

28. Епимахов А. В. К вопросу о радиоуглеродной аргументации ранней датировки алакульских древностей. Вестник археологии, антропологии и этнографии, Тюмень, № 3(34), С. 60-67, 2016.