Лингвистическая филогения с человеческим лицом: быстрое разделение индоевропейских языков в раннем бронзовом веке

Алексей Касьян

В конце 2018 г. наша группа завершила многолетний проект по филогении индоевропейской языковой семьи. После нескольких туров рецензирования статья с описанием результатов была принята в Diachronica, который является ведущим мировым журналом по сравнительно-историческому языкознанию. [**Апдейт**: в виду своей важности статья была перенесена в журнал по общему языкознанию *Linguistics*, где вышла в июне 2021 г.]

Alexei S. Kassian, Mikhail Zhivlov, George Starostin, Artem A. Trofimov, Petr A. Kocharov, Anna Kuritsyna, Mikhail N. Saenko. 2021. Rapid radiation of the Inner Indo-European languages: an advanced approach to Indo-European lexicostatistics. Linguistics 59 (4). https://doi.org/10.1515/ling-2020-0060 (Open Acces).

Хотя мои основные научные интересы связаны с формализацией методов выяснения языкового родства, сразу скажу, что лично я рассматриваю индоевропейский проект как некоторый побочный результат тех исследований, которые веду я и наша группа.

Лингвистика сейчас переживает бурную цифровизацию (с эволюцией нейросетей этот процесс будет развиваться всё дальше), и лингвистическая компаративистика — не исключение. В компаративистике цифровизация касается нескольких аспектов: в частности, это алгоритмы автоматического определения когнатов между словарями сравниваемых языков и алгоритмы выяснения степени родства между сравниваемыми языками, т.е. построение генеалогического дерева.

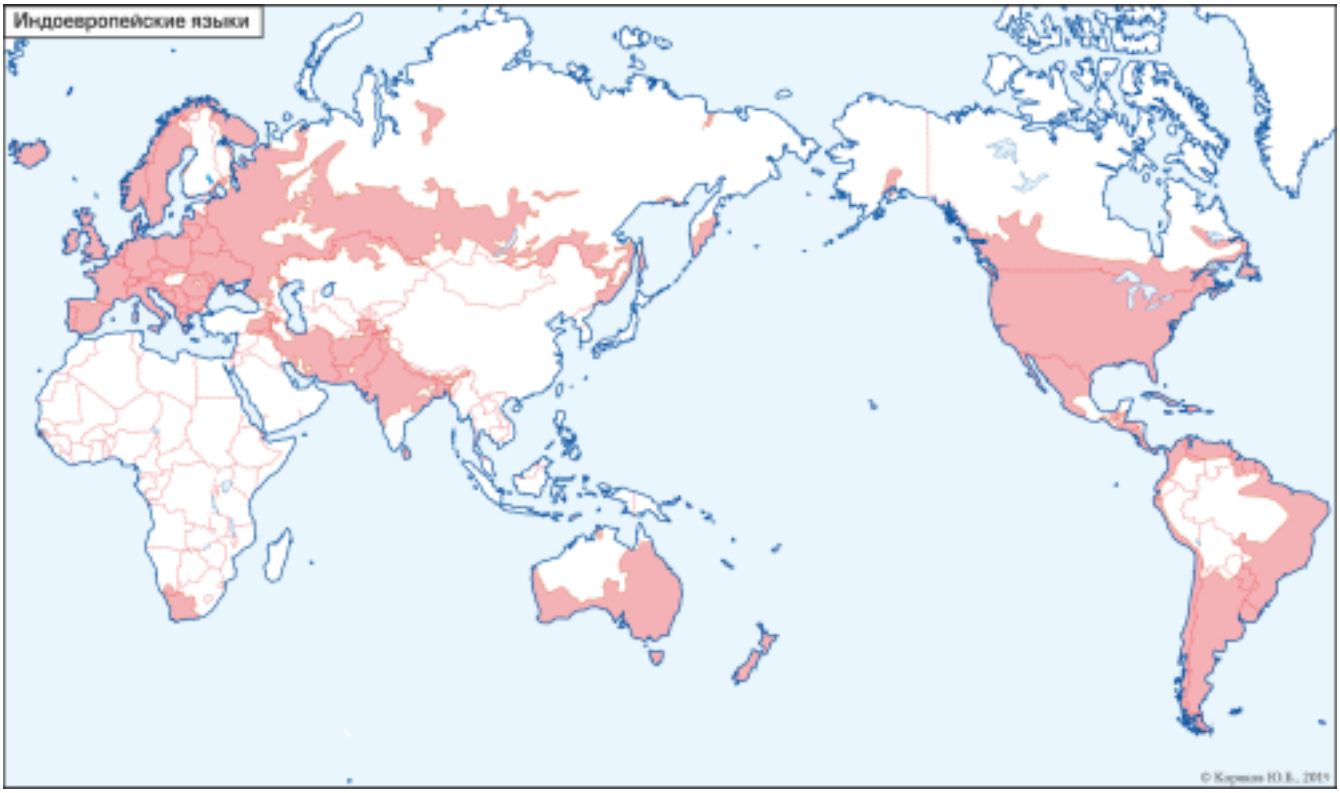

Рис. 1. Карта современных индоевропейских языков (автор: Юрий Коряков). На сегодняшний день индоевропейская семья занимает первое место по площади распространения и количеству носителей.

Что касается формализованных генеалогических связей между языками, первые важнейшие шаги в этом направлении были сделаны Моррисом Сводешем в середине XX в. В силу некоторых методологических изъянов та пионерская версия лексикостатистики не нашла большой популярности среди лингвистов. Начиная с 80-х годов наш выдающий лингвист Сергей Анатольевич Старостин развивал идеи Сводеша и преодолел, по крайней мере, часть теоретических проблем (Starostin 2007). К сожалению, Старостин публиковался преимущественно на русском, поэтому не оказал в полной мере того влияния на мировую компаративистику, которое мог бы оказать. В начале XXI в. группа Дона Ринджа выступила с обновленной концепцией построения лингвистических деревьев (Ringe, Warnow & Taylor 2002): на довольно качественном лингвистическом датасете методом наибольшей совместимости они предложили некоторую древесную классификацию и.-е. языков. Но опять же из-за довольно существенных методологических изъянов (специально подобранный датасет и кольцевая логика построения дерева) концепциях Ринджа не нашла большого количества сторонников.

Мощное «второе дыхание» лингвистическая филогения получила в начале 10-х гг. XXI в., когда этой областью стали заниматься люди с биологическим образованием и биологическим образом мышления, поскольку используемые техники и алгоритмы более-менее напрямую заимствованы из биологии (генетики). С одной стороны, для нас это крайне полезно и продуктивно — посмотреть, как мыслят и работают представители более точной науки. С другой стороны, отсутствие понимания лингвистической специфики у этих авторов привело к серьезным ляпам и явно ошибочным выводам, по крайней мере, в некоторых знаковых публикациях последних десяти лет, в первую очередь это статьи группы Расселла Грея и Квентина Аткинсона, выходившие в ведущих журналах вроде Nature и Science (ср., например, обширную критику в блогах Аси Перельцвайг и Мартина Льюиса: «The Malformed Language Tree of Bouckaert and His Colleagues» или «Quentin Atkinson's Nonsensical Maps of Indo-European Expansion» и др.). Далее началось своего рода состязание между биоинформатиками, кто сможет предложить более заковыристую и зубодробительную модель для описания языковой эволюции, где ценность имеет математический аппарат, а не лингвистические данные и не итоговый филогенетический результат (который в иных статьях оказывался самым диким), напр., Blanchard et al. 2011. Со стороны это может выглядеть как дискредитация формального подхода, и сегодня это подталкивает многих компаративистов, особенно старшего поколения, вообще к отрицанию лексикостатистики и всех подобных приемов классификации языков.

В этой связи следует сформулировать принципиальную разницу между сравнительно-историческим языкознанием и генетикой — разницу, которую представители смежных наук обычно плохо осознают:

1. в генетике собрано и накоплено очень много качественных данных, а актуальная задача заключается в разработке всё более сложных алгоритмов анализа, чтобы вычленить полезный сигнал.
2. в лингвистике же остро стоит проблема с нехваткой качественных входных данных (поскольку мало кто хочет тратить на время на сбор и обработку первичного материала), а вопросы матаппарата явно отходят на второй план.

Однако не всё так плохо, потому что сначала наша московская группа, а теперь и всё большее число коллег за рубежом активно пропагандируют важность качества лингвистических данных, подаваемых на вход модели. Надеюсь, через несколько лет это станет базовым принципом языковой филогении.

Итак, моя основная деятельность связана с разработкой принципов сбора лингвистического материала, изобретением техник и приемов последующей обработки этого материала и, наконец, применением существующих компьютерных алгоритмов к лингвистическому материалу для получения генеалогического дерева.

Все эти новшества, разумеется, следует тестировать на языковых группах с уже известной структурой: мы автоматизированно строим дерево, сравниваем насколько оно (не) противоречит мнению традиционных экспертов и далее делаем вывод, насколько адекватными оказались наши теоретические идеи. Как показывает практика, разработки нашей группы демонстрируют очень хорошие практические результаты. Тестирование на языковых группах с заранее известной филогенией показывает, что наши формальные деревья совпадают с традиционной классификацией: лезгинская группа (Kassian 2015), цезская группа (Kassian 2017), славянская группа (лингвистическая часть в Kushniarevich et al. 2015).

Что касается индоевропейских языков, то у них нет никакой устоявшейся генеалогической классификации, поэтому в орбиту основных интересов нашей группы индоевропейская семья не входит. Однако, поскольку мы накопили очень много нового в том, что касается языкового филогенетического анализа, а индоевропейские языки — интересная для публики тема, то мы решили, что могли бы применить наши знания и умения и к этой языковой семье. И можно сказать, что результаты получились очень удачными.

Рис. 2. Фантазия художника на тему неолитических земледельцев. Быт праиндоевропейцев мог выглядеть так.

Работа над проектом шла следующим образом. Первым делом мы собрали 110-словные сводешевские списки для основных древних и многих современных языков и.-е. семьи. Делалось это по строгой методологии и семантическим спецификациям, описанным в нашей более ранней статье (Kassian et al. 2010). Несмотря на кажущуюся легкость, это совсем не простая задача: на составление одного списка у квалифицированного лингвиста может уйти две-три недели.

Далее мы применили прием поэтапной реконструкции. Как известно, в и.-е. семье абсолютно консенсусно выделяется ряд неглубоких групп, таких как славянская, германская, албанская и т.п. Если у группы хорошо зафиксирован древний язык, который (пусть и с натяжкой) может рассматриваться как праязык данной группы, то мы брали сводешевский список для

этого древнего языка: например, для всей индийской группы это ведийский санскрит. А если такого языка не обнаруживается, то мы на основании синхронных списков реконструировали сводешевский список для праязыка данной группы. Так мы реконструировали 110-словные списки для праславянского, прабалтийского, прагерманского, праиранского, прабриттского.

Использования именно небольшого числа прасписков вместо большого числа синхронных списков имеет два сильных преимущества.

1. С математической точки зрения, чем больше таксонов (языков) мы исследуем, тем больше требуется признаков (сводешевских слов) для построения правдоподобного дерева. Скажем, для 30 таксонов может быть достаточным 100-словник, а для 30-100 таксонов уже лучше использовать 200-словник. При этом, чем дальше мы отдаляемся от сводешевского 100-словника, тем менее стабильные и менее семантически ясные концепты нам будут попадаться, т.е. для какого-либо языка составить 200-словник — это задача не в два, а в несколько раз более сложная, чем сбор 100-словника. В конечном итоге всё упирается в квалифицированные человеко-часы, которых, разумеется, не хватает.
2. Чем дальше наши списки отстоят от корня дерева (от праязыка), тем больше в них накапливается гомопластичных (параллельных) эволюционных событий. А чем больше у нас входных списков, тем больше в них будет ошибок в силу человеческого фактора. Всё это добавляет шум в датасет и усложняет модель.

Конечно, у нашего метода ступенчатой реконструкции есть своя оборотная сторона: реконструируя прасписки, мы можем банально ошибиться и взять в прасписок совсем не то слово, которое в данном языке выражало данный сводешевский концепт. Мы, однако, оцениваем вероятность ошибиться в конкретных концептах как не слишком высокую и не считаем, что этот риск перевешивает две проблемы синхронных списков, описанные выше. Дело в том, что, во-первых, мы реконструируем списки для довольно неглубоких групп, их хронологический возраст обычно составляет 2000-2500 лет (скажем, славянской группе ок. 2000 лет, германская группа глубже, но не принципиально глубже). Во-вторых, что важнее, мы используем строгую методологию семантической (ономасиологической) реконструкции, недавно разработанную нашей группой (Kassian, Starostin & Zhivlov 2015). В этой методологии сформулированы пять критериев, позволяющих выбрать для того или иного сводешевского концепта наиболее вероятную праоснову. Эти критерии таковы:

1. Топология дерева. Мы стремимся сократить число эволюционных событий на дереве.
2. Внешняя этимология, подсказывающая нам исходную семантику при сравнении нескольких лексических кандидатов.
3. Морфологическая производность. Морфологически прозрачное производное имеет больше шансов оказаться инновацией, чем непроизводная основа.
4. типология семантических сдвигов. Переход между некоторыми значениями обычен в обоих направлениях (напр., ‘трава’ ↔ ‘зеленый’), а в некоторых парах переход возможен только в одну сторону (напр., ‘светить’ → ‘луна’).
5. исключение ареального эффекта. Если лексическая изоглосса захватывает соседние языки, она может быть результатом контактов.

В итоге в нашем датасете оказалось 13 списков, представляющих все известные группы и.-е. семьи (астериском помечены реконструированные прасписки):

Таксон	**Датировка**
хеттский	1650–1500 до н.э.
тохарский B	400–900 н.э.
древнегреческий	375 до н.э.
классический армянский	400–500 н.э.
албанский (современный)	1950 н.э.
латынь	200 до н.э.
древнеирландский	700–900 н.э.
*прабриттский	300–600 н.э.
*прагерманский	500–300 до н.э.
*праславянский	1–300 н.э.
*прабалтийский	400–1 до н.э.
ведийский санскрит	1200–1000 до н.э.
*праиранский	1500–1000 до н.э.
*прасамодийский (представитель уральской семьи, добавлен для укоренения дерева)	950–750 до н.э.

Разметив формы с этимологически родственными корнями между списками, мы получили традиционную лексикостатистическую матрицу с корневыми когнациями (т.е. когда основы из разных языков, имеющие этимологически

общий корень, помечаются как родственные друг другу). Например, в этой матрице герм. **wenda-* =слав. **vętrъ*, а скр. *agni* = лат. *ignis*. Назовем эту матрицу Этап-1.

На основе этой матрицы мы строим деревья, причем не одним методом (как обычно делают), а тремя разными методами: метод ближайших соседей (специально модифицированный для лингвистических исследований), байесовский метод и метод максимальной парсимонии. Особенности этих методов — отдельная объемная тема, в которую сейчас нет нужды углубляться. Нам важно, что каждый из этих методов имеет свои сильные и слабые стороны, поэтому мы используем все три, а потом три полученных дерева объединяем в одно консенсусное дерево, которое и рассматриваем как результат Этапа-1.

Однако нас не очень интересует топология, полученная из корневых когнаций, потому что мы понимаем, что наши входные лексические данные можно улучшить и таким образом усилить филогенетический сигнал.

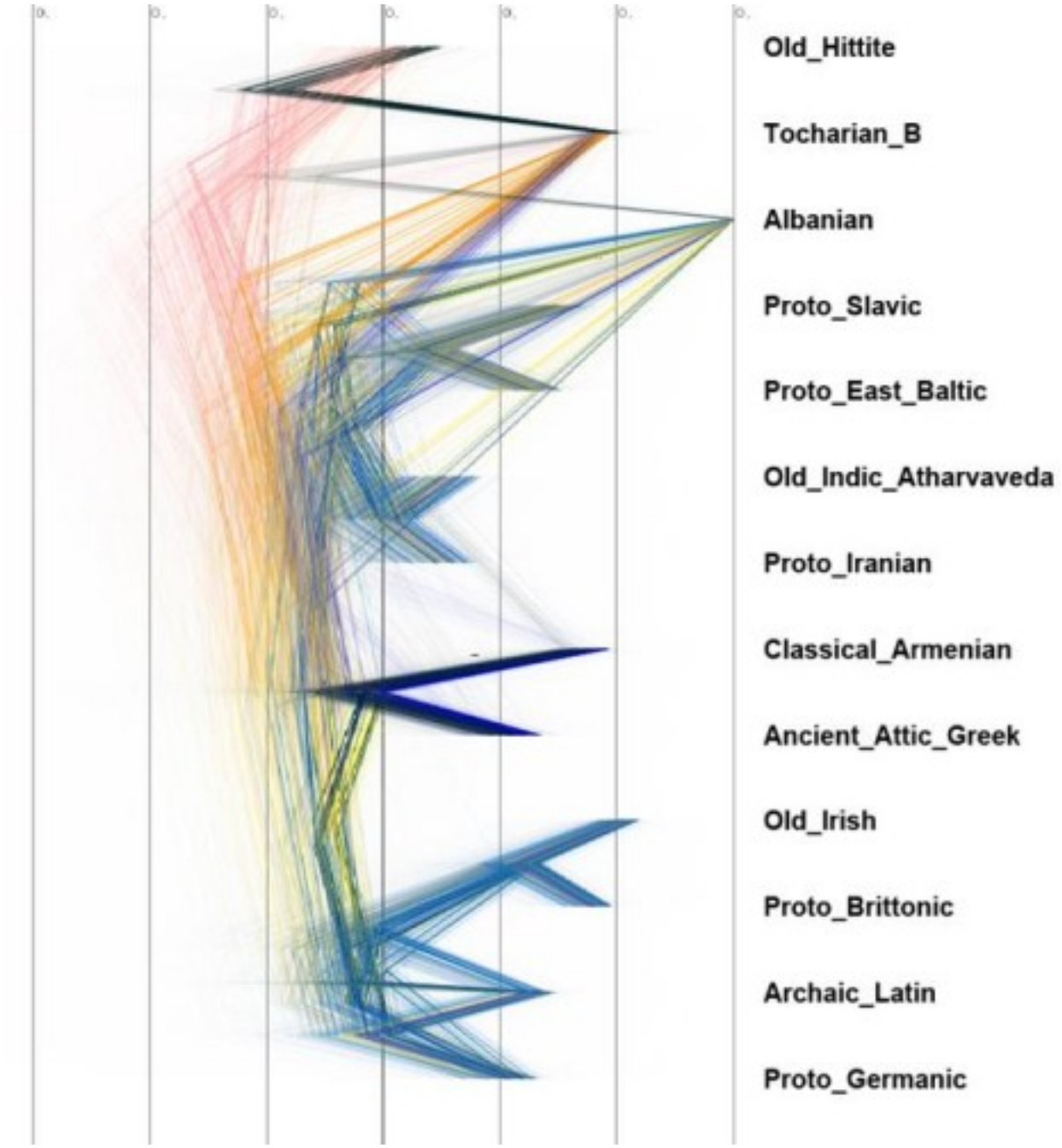

Рис. 3. График DensiTree, суммирующий байесовские деревья, полученные на Этапе-3.

На Этапе-2 мы убираем из лексической матрицы так называемый деривационный дрейф, т.е. случаи, когда мы подозреваем в разных языках параллельные морфологические образования. Напр., герм. **we-nd-a-* и слав. **vě-tr-ъ* имеют общий корень (и.-е. **xwe:-* ‘веять’), но это девербативы с разными суффиксами, и интуитивно хочется считать, что появление славянского

новообразования **vě-tr-ъ* — это отдельное эволюционное событие на лексикостатистическом дереве. С другой стороны, не всякая разница в морфологическом оформлении говорит о параллельных новообразованиях, напр., совсем не хочется считать др.-греч. *kard-í-a:* ‘сердце’ и герм. **xert-o:n* ‘сердце’ не связанными друг с другом производными (исторически это результат различной адаптации праиндоевропейской атематической парадигмы). Пока мы предложили два критерия деривационного дрейфа (критериев может быть больше, эта тема нуждается в дополнительной разработке):

1. Если две основы из сопоставляемых языков имеют общий корень, но различаются аффиксальной структурой, и есть свидетельства в пользу того, что хотя бы одна из основ подверглась частеречному изменению (например, существительное ↔ глагол), эти основы, скорее всего, демонстрируют гомопластичное развитие. Пример с герм. **we-nd-a-* и слав. **vě-tr-ъ* как раз иллюстрирует этот критерий (мы имеем тут разные аффиксы и частеречный переход глагол → существительное).
2. Если две основы из сравниваемых языков 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 гомопластичное развитие. Например, в латинском, балтийском и кельтском обозначения ‘человека’ образованы от индоевропейского термина ‘земля’ (т.е. ‘человек’ как ‘земной, землянин’), но с разными суффиксами: **-on* в латинском и в прабалтийском (*hom-in-*, **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 по частотной семантической модели.

Применяя эти два критерия к нашей лексической матрице, мы размечаем основы с деривационным дрейфом как неродственные друг другу. Таким образом мы получаем матрицу, где герм. *wenda- ≠ слав. **větrъ* (по-прежнему скр. *agni* = лат. *ignis*). Назовем эту матрицу Этап-2.

На Этапе-2 мы также строим три дерева и суммируем их в виде консенсусного дерева. В принципе, такое дерево уже является полноценным научным результатом, но всё же филогенетический сигнал можно усиливать и дальше.

Поэтому мы приступаем к Этапу-3, на котором производим гомопластичную оптимизацию матрицы. Эта процедура заключается в том, что, если в матрице есть когнаты, которые противоречат древесной структуре (полученной на Этапе-2), то во многих (далеко не всех) случаях мы можем с большой степенью вероятности предположить, что эти формы представляют собой параллельное (т.е. гомопластичное) развитие. Напр., и.-е. основа для значения ‘огонь’ вполне надежно восстанавливается как **pex-wr*, она сохраняется в обоих аутлайерах (анатолийском и тохарском) и в ряде узкоиндоевропейских групп: в греческой, армянской, германской, италийской. В балто-славяно-индо-иранской кладе эта основа вытесняется основой **ng-n-i-* (откуда русск. *огонь*, др.-инд. *agni-* и пр.). Загадочным образом *ignis* значит ‘огонь’ и в латыни, хотя италийская группа сохраняет старое обозначение ‘огня’: умбр. *pir*. Какова была изначальная семантика основы **ng-n-i-*, мы не знаем, но, поскольку ни одна из классификаций не объединяет латынь в отдельную кладу с балто-славяно-индо-иранским и тем более в обход умбрского, то с большой вероятностью лат. *ignis* ‘огонь’ — это параллельное семантическое развитие, независимое от **ng-n-i-* ‘огонь’ в балто-славяно-индо-иранской кладе. На основе этих рассуждений в датасете Этапа-3 мы помечаем лат. *ignis* ‘огонь’ как форму, не родственную балто-славяно-индо-иранскому **ng-n-i-* ‘огонь’, это и есть гомопластичная оптимизация. Таким образом, имея консенсусное дерево (Этап-2) и очистив матрицу от явных гомопластичных событий, мы получаем уже итоговую для нашего исследования матрицу, на основе которой опять строим консенсусное дерево. Это Этап-3 и финал нашей работы.

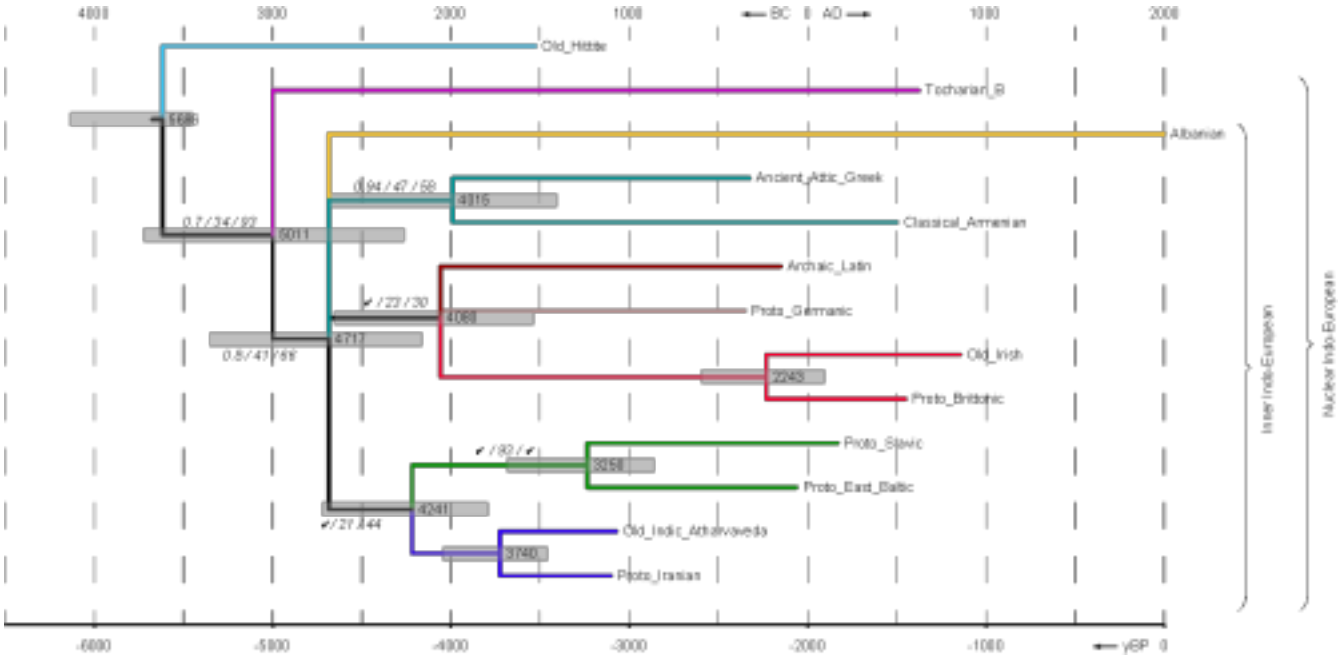

Рис. 4. Строгое консенсусное дерево индоевропейской семьи на основе набора данных Этапа-3 (wind ≠ veter, agni ≠ ignis). Дерево суммирует три дерева, полученных индивидуальными методами. Даты получены с помощью анализа Bayesian MCMC: серые линейки показывают временной интервал с 95% вероятностью времени расхождения; справа от каждого узла указывается среднее время расхождения. Нижняя шкала — 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, не указано для узлов с P ≥ 0,95 во всех методах). Традиционные группы обозначены цветом.

А что в итоге? Наше итоговое дерево не содержит никаких несуразиц с точки зрения классической индоевропеистики, как в плане топологии дерева, так и в плане датировок, например, первое разделение на анатолийскую и не-анатолийскую ветви мы помещаем в промежуток 4139–3450 до н.э. (средняя дата 3686 до н.э.) — точно так же и современные индоевропеисты предполагают, что распад произошел в первой половине 4-го тысячелетия до н.э.

По всей видимости, за всю историю индоевропейских штудий это первый раз, когда, используя формальные биологические классификационные методы без каких-либо предварительных ограничений на топологию и вообще без какого-либо насилия над материалом и матаппаратом, были получены дерево и датировки, не противоречащие традиционным взглядам индоевропеистов.

Наша главная находка — это одномоментное разделение узкоиндоевропейского узла на четыре ветви в промежутке 3400–2200 до н.э.: (1) греко-армянскую, (2) албанскую, (3) итало-германо-кельтскую, (4) балто-славяно-индо-иранскую. Такой быстрый распад очень хорошо согласуется с тем фактом, что за последние 150 лет (первое дерево и.-е. языков было опубликовано Шлейхером в 1861 г.) индоевропеисты пришли более-менее к консенсусу о первых аутлайерах (анатолийский, тохарский) и о молодых кладах вроде балто-славянской или индо-иранской. А вот какие ветвления происходят в середине дерева, нет понимания до такой степени, что подавляющее большинство традиционных индоевропеистов вообще отказывается мыслить и.-е. семью в древесном виде и предпочитает не касаться этой темы (если вы откроете современные учебники по индоевропеистике, то вряд ли вы там увидите филогенетические картинки). Это значит, что нет никаких мощных пучков изоглосс, способных помочь внутренней классификации, и в свете этого мультифуркация на четыре ветви представляется наиболее правдоподобным сценарием.

Полученные датировки промежуточных узлов удивительно хорошо соответствуют радиоуглеродным датам некоторых археологических культур, которые традиционно связываются с расселением индоевропейцев (поиск прародины или промежуточных прародин индоевропейцев ни в коем случае не входил в наши задачи, но я хочу отметить, что и тут наши результаты не вступают в противоречие с взглядами индоевропеистов):

1. Возникновение афанасьевской культуры: 2800 до н.э. // Отделение тохарского: 3727–2262 до н.э. (средняя дата 3011 до н.э.).
2. Закат синташтинской культуры: 1800 до н.э. // Распад индоиранской клады: 2044–1458 до н.э. (средняя дата 1740 до н.э.).
3. Закат культуры шнуровой керамики: 2300-2000 до н.э. // Бинарный распад балто-славяно-индо-иранской клады: 2723–1790 до н.э. (средняя дата 2241 до н.э.).
4. Закат культуры колоколовидных кубков: 2100 до н.э. // Троичный распад итало-германо-кельтской клады: 2655–1537

до н.э. (средняя дата 2080 до н.э.).

Рис. 5. Современная аэросъемка Аркаима (синташтинская культура).

Под конец упомяну такой частный аспект нашего исследования. Помимо 13 индоевропейских списков в датасет был опционально добавлен прасамодийский список. Самодийская группа — это одна из двух ветвей уральской семьи. Прауральский язык, видимо, является ближайшим родственником праиндоевропейского. Сегодня далеко не все ученые признают обоснованными вообще какие-либо внешние связи индоевропейского, но тем не менее индо-уральская гипотеза постепенно завоевывает признание и, наверное, в ближайшем будущем станет мейнстримом (например, не так давно наша группа обосновала индоевропейско-уральское родство со статистической точки зрения: Kassian, Starostin & Zhivlov 2015). Мы продублировали все подсчеты и деревья с самодийским и без самодийского, и оказалось, что различий в топологии и хронологии фактически нет: самодийский оказывается аутгрупом (первым отделившимся таксоном) и не меняет индоевропейское дерево. Это должно указывать на то, что лексические схождения между самодийским (уральским) и индоевропейским не стохастические случайные созвучия, а могут представлять собой древний индо-уральский лексический фонд, независимо изменяющийся в обеих ветвях по предполагаемой нами эволюционной модели. Так что в принципе это является еще одним косвенным свидетельством в пользу индоевропейско-уральского родства.

Литература

- Blanchard, Ph., F. Petroni, M. Serva & D. Volchenkov. 2011. Geometric representations of language taxonomies. *Computer Speech & Language* 25(3). 679–699. doi:10.1016/j.csl.2010.05.003.
- 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. doi:10.1371/journal.pone.0116950.
- Kassian, Alexei S. 2017. Linguistic homoplasy and phylogeny reconstruction. The cases of Lezgian and Tsezic languages (North Caucasus). *Folia Linguistica* 51(s38). 217–262. doi:10.1515/flih-2017-0008.
- Kassian, Alexei S., George Starostin, Anna Dybo & 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. 2015. Proto-Indo-European-Uralic comparison from the probabilistic point of view. *Journal of Indo-European Studies* 43(3–4). 301–347.
- Kushniarevich, Alena, Olga Utevska, Marina Chuhryaeva, Anastasia Agdzhoyan, Khadizhat Dibirova, Ingrida Uktveryte, Märt

Möls, et al. 2015. Genetic heritage of the Balto-Slavic speaking populations: a synthesis of autosomal, mitochondrial and Y-chromosomal data. *PLOS ONE* 10(9). 1–19. doi:10.1371/journal.pone.0135820.

- Ringe, Don, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1). 59–129. doi:10.1111/1467-968X.00091.
- Starostin, Sergei A. 2007. *Trudy po jazykoznaniju [Works on linguistics]*. Moscow: Jazyki slavjanskix kul'tur.