# Как строить деревья? Проверка на лезгинских языках

Алексей Касьян

Для построения языкового филогенетического дерева самое главное – качественные исходные данные

**Впервые проведенный полноценный тест современных филогенетических методов на лексическом материале лезгинской языковой группы, показал, что залогом надежной реконструкции дерева является качественно подготовленные входные данные (лексические списки), а выбор конкретных математических алгоритмов менее важен.**

В современной биологии используется целый ряд математических методов для построения филогенетического дерева рассматриваемых биологических единиц. Методы можно разделить на два больших класса: дистантные (distance-based) и признаковые (character-based). Пожалуй, наиболее популярный среди дистантных — это метод ближайших соседей, а среди признаковых — Монте-Карло для марковских цепей в рамках байесовского подхода (MCMC) и метод максимальной бережливости. Дистантные методы требуют значительно меньших вычислительных мощностей и имеют более давнюю традицию использования, нежели признаковые. Однако у биологов постепенно возобладает мнение, что признаковые методы в целом реконструируют филогению более надежно по сравнению с дистантными.

Первые полноценные попытки применения биологических филогенетических алгоритмов к языковым данным относятся к середине XX в., это лексикостатистика Морриса Сводеша (Morris Swadesh). Несмотря на накопленный лингвистами за последние десятки лет опыт остаются открытыми два принципиальных вопроса:

- Какого рода лингвистические данные лучше использовать для формальной генеалогической классификации языков?
- Какие из известных методов ближе соответствуют естественной языковой эволюции?

**Выбор данных**

По первому вопросу среди лингвистов нет консенсуса. Лингвистические признаки, пригодные для генеалогической классификации, можно разделить на два класса:

- Признаки, относящиеся к словарю, иначе говоря, лексика. В подавляющем большинстве случаев исследователи используют лексикостатистику. В ее основе лежит выяснение того, родственными или неродственными словами в двух сравниваемых языках выражается данное значение (напр., для значения ‘кора’ в украинском используется слово *кора*, и в болгарском — *кора*, это этимологически родственные лексические единицы; а для значения ‘облако’ — *хмара* и *облак* соответственно, это этимологически неродственные слова).
- Признаки из области грамматического описания, а именно, относящиеся к фонетике (в том числе исторической), морфологии и/или синтаксису.

На практике матрицы, состоящие исключительно из грамматических признаков, используются редко, причем известны случаи очевидно неудовлетворительных классификаций, когда полученное формальное дерево резко отличается от бесспорной традиционной филогении (как, например, было с языковой семьей на-дене). Достаточного опыта использования смешанных лексико-грамматических матриц пока не накоплено. Наиболее обиходный сегодня метод — это лексикостатистика. В частности она является очевидным выбором, когда мы классифицируем языки с редуцированной грамматикой (вроде китайского или английского).

> Дополнительно можно отметить такие недостатки грамматических признаков. Во-первых, они зачастую не универсальны, а это делает полученные классификации несовместимыми на более высоком таксономическом уровне. Напр., признак *базового порядка слов* универсален и применим ко всем языкам мира (в русском базовый порядок — SVO, как во фразе *Петя читает книгу*). С другой стороны, для классификации славянских языков может быть уместно в качестве одного из признаков взять окончания первого лица единственного числа настоящего времени: выглядит ли оно как гласный типа *-у* или как согласный *-м* (ср. русское *я зна-ю* и его польский эквивалент *ja zna-m*), но этот признак теряет свой смысл за пределами славянской группы.
> Во-вторых, грамматические признаки легко могут образовывать вторичные ареальные изоглоссы (особенно если речь идет о языках, чье родство еще ощущается носителями), причем выявить источник инновации часто оказывается затруднительно.
> В-третьих, грамматические признаки образуют систему, т. е. изменение одного признака с высокой вероятностью влечет за собой изменение других признаков.
> Для лексических же признаков, используемых в лексикостатистике, все перечисленные недостатки характерны в значительно меньшей степени.

**Выбор методов**

Теперь следует вернуться ко второму обозначенному вопросу, а именно, какие из известных математических методов ближе к естественной языковой эволюции. Тут у современной науки значительно меньше понимания, чем в проблеме выбора классификационных признаков.

Какими бы теоретическими рассуждениями мы ни обосновывали выбор того или иного математического алгоритма, подобные рассуждения немногого стоят, если не подкреплены тестами на конкретном языковом материале. Имеются вполне естественные ограничения, накладываемые на проверочный материал. Во-первых, лингвистические данные (напр., списки слов для лексикостатистики) должны быть подготовлены максимально качественно, чтобы исключить возмущающий фактор шума. Во-вторых, для рассматриваемого набора языков уже должна иметься общепризнанная и не подвергаемая сомнению классификация — «золотой стандарт», с которым мы и будем сравнивать получаемые формальными методами деревья. Как ни странно — и это, конечно, упрек современной исторической лингвистике — такие тесты если и проводились, то до сих не публиковались.

Известна объемная статья [Nakhleh L, Warnow T, Ringe D, Evans SN (2005) A Comparison of phylogenetic reconstruction methods on an IE dataset. The Transactions of the Philological Society 103: 171–192. doi: 10.1111/j.1467-968x.2005.00149.x ], где филогенетические методы последовательно применяются к языкам индоевропейской семьи (использовался смешанный набор лексических и грамматических признаков). Различные методы дали различающиеся по своей топологии деревья. Однако возникает неопределенность с оценкой достоверности этих деревьев, поскольку среди языковедов нет никакого консенсуса касательно филогении индоевропейских языков: нам понятно, кто отделился первым (это хетто-лувийская и тохарская ветви) и понятно объединение языков в неглубокие группы (славянская, германская, кельтская и т.д.), но промежуточное членение остается предметом споров (напр., кто ближе к славянским языкам — индийские или германские?). Следовательно в деревьях из статьи [Nakhleh et al. 2005] мы, конечно, можем выискивать отдельные ошибки (заведомо неприемлемые ветвления), но отсутствие «золотого стандарта» не позволяет нам дать общую оценку адекватности деревьев, реконструированных тем или иным методом.

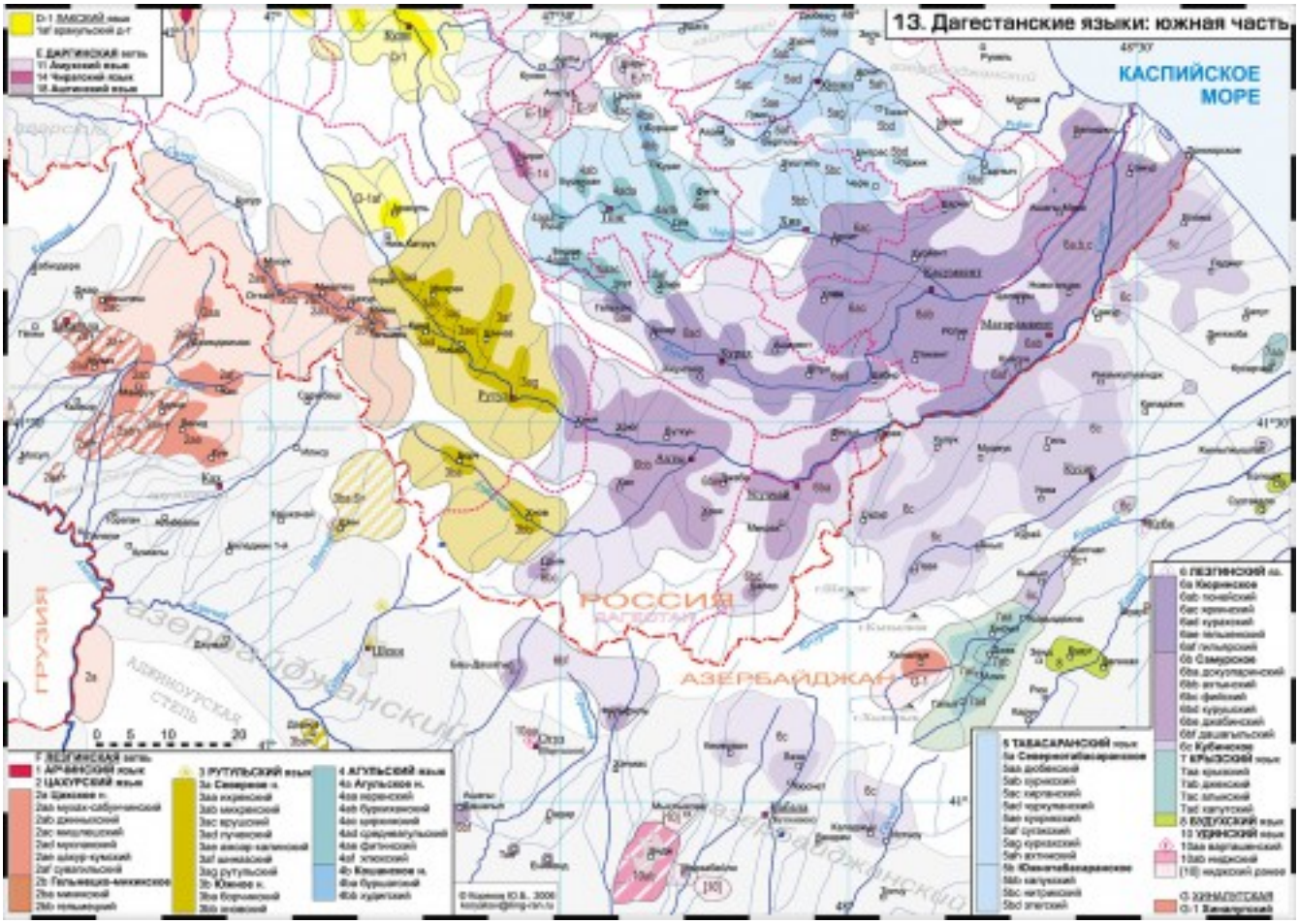
Второй попыткой тестирования филогенетических методов для реконструкции языковых деревьев была статья [Barbançon F, Evans SN, Nakhleh L, Ringe D, Warnow T (2013) An experimental study comparing linguistic phylogenetic reconstruction methods. Diachronica 30/2: 143–170. doi: 10.1075/dia.30.2.01bar , однако здесь мы имеем дело не с полевой проверкой, а с искусственно моделируемыми языковыми ситуациями. Интересно, что авторы в результате приходят к следующей иерархии методов: самый адекватный — это метод максимальной бережливости, несколько менее точный — байесовский MCMC, затем дистантный метод ближайших соседей и аутсайдер — метод невзвешенного попарного среднего.

**Тестирование на лезгинских языках**

И, наконец, первая полноценная проверка филогенетических маталгоритмов на языковом материале была опубликована в феврале 2015 г.: [Kassian A (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]

Русский лингвист Алексей Касьян из Института языкознания РАН исследовал вопрос на сводешевских списках лезгинской языковой группы, собранных им для международного проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database» (руководитель проекта Георгий Старостин).

Лезгинские языки распространены на юго-востоке Дагестана (Россия) и отчасти в прилегающих районах Азербайджана. Насчитывается 9 современных лезгинских языков, причем большинство обладает довольно развитой диалектной структурой: удинский, арчинский, крызский, будухский, цахурский, рутульский, агульский, табасаранский, собственно лезгинский. Они изображены на карте, рис. 1 (автор карты Юрий Коряков).



Карта современных лезгинских языков. Автор Юрий Коряков.

Лезгинская группа довольно глубока хронологически, глубже, скажем, славянской или германской. Она входит в состав нахско-дагестанской клады северокавказской языковой семьи. Интересно, что первый зафиксированный язык северокавказской семьи относится к лезгинским: это язык богослужебных христианских рукописей государства Кавказская Албания. Два пергаменных палимпсеста (т.е. рукописей со смытым старым текстом, поверх которого записан новый текст, однако нижний слой частично читается, особенно при определенных технических ухищрениях), обнаруженных совсем недавно в кладке одного из синайских монастырей, относятся к концу 1-го тысячелетия н.э. Можно было бы ожидать и большего количества письменных памятников Кавказской Албании (если сравнить с обилием древних грузинских и древних армянских христианских рукописей), но довольно быстро эта часть Кавказа была исламизирована, и о полноценной письменной культуре на местных языках пришлось забыть.

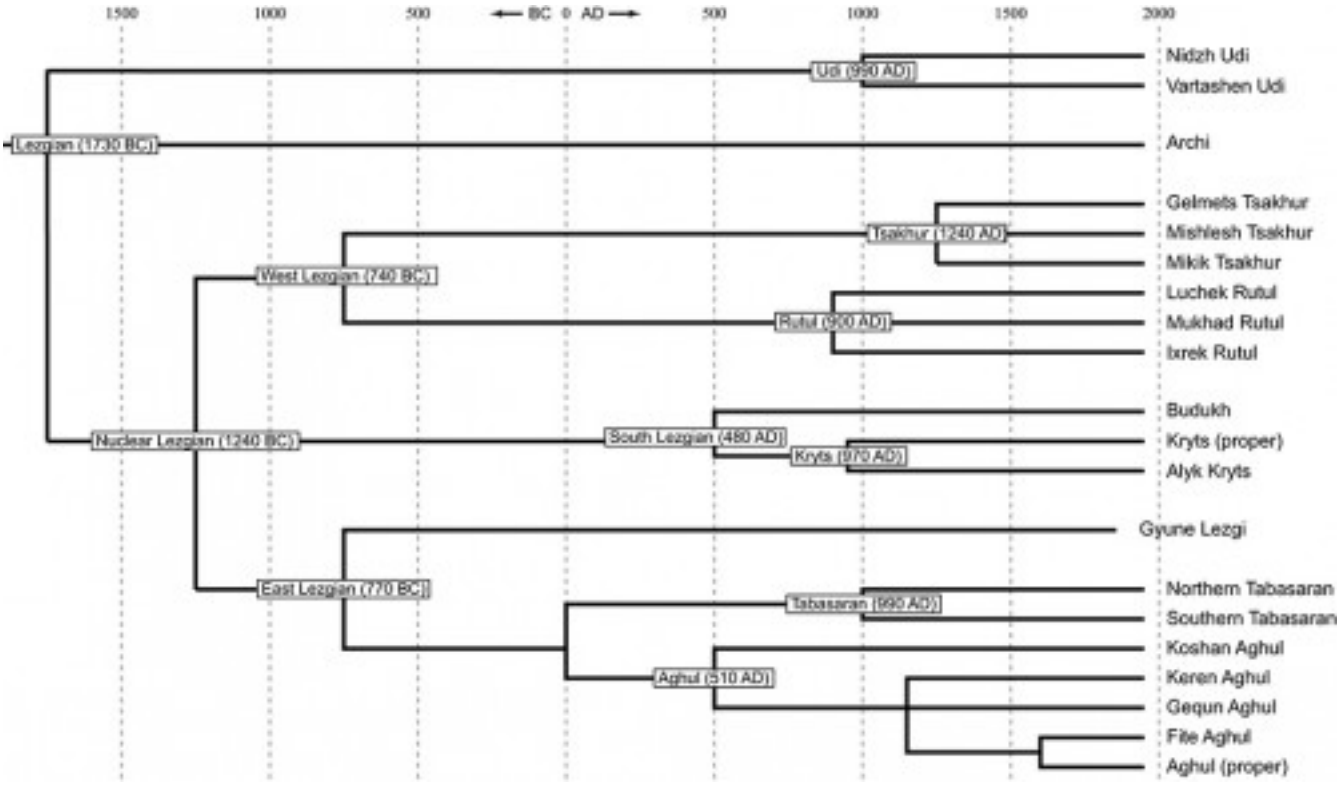
Использованная в тестировании база данных с 110-словными лексическими списками лезгинских языков и диалектов (см. сайт «Глобальная лексикостатистическая база данных») обладает рядом важных свойств (см Врез).

1) база включает в себя достаточно большое число идиомов: 20 единиц, причем среди них есть как языки, длительное время развивающиеся в изоляции, например арчинский, так и языки, активно контактирующие с другими языками данной группы (что потенциально дает подскок лексикостатистических совпадений благодаря контактно обусловленной гомоплазии), например, агульский;
2) среди кавказоведов имеется обоснованный консенсус относительно членения лезгинской группы (аутлайеры удинский и арчинский, отделившиеся первыми, плюс узколезгинская, или самурская, подгруппа, состоящая из трех кластеров: западного, южного и восточного с дальнейшим диалектным членением) — это дает тот «золотой стандарт», с которым мы будем сравнивать получаемые деревья;
3) лезгинскую группу, наверное, можно охарактеризовать как среднюю или чуть выше среднего по параметру надежности и подробности имеющихся лексикографических и грамматических описаний среди языковых групп мира — это отличает лезгинские языки, например, от чрезвычайно подробно описанных индоевропейских языков и делает лезгинскую группу модельным случаем с точки зрения типологии;
4) общее лексикографическое качество, так сказать, степень «очистки» лезгинских лексикостатистических списков (равно как и списков других языков в проекте «Глобальная лексикостатистическая база данных») беспрецедентно высока для мировой лингвистики.

**Разные методы – одинаковые деревья**

А. Касьяном были протестированы основные филогенетические методы: байесовский MCMC, метод максимальной бережливости, ближайших соседей (три разновидности) и невзвешенного попарного среднего.

Результаты получились интересными: все методы дали весьма схожие деревья, которые при этом совпадают с традиционной классификацией. Видимо, наиболее отличающееся дерево выдал метод максимальной бережливости, но и у него расхождения с традиционной классификацией локальны и не носят фатального характера. Итоговое консенсусное дерево лезгинских см. на рис. 2.



Консенсусное генеалогическое дерево лезгинских языков с глоттохронологическими датами.

Лезгинский тест подтверждает некоторые положения, составляющие идеологическую основу проекта «Глобальная лексикостатистическая база данных»:

1) Лексические признаки хорошо подходят для генеалогической классификации языков, а с грамматическими признаками возникают различные проблемы (о чем см. выше).

2) Точность филогенетического дерева зависит в первую очередь не от математического метода, а от степени очистки входных данных, иными словами, не от труда компьютера, а от труда лингвиста, кропотливо анкетирующего индивидуальные диалекты по принятому списку признаков.

С практической точки зрения важным оказывается и такой вопрос: с ухудшением лингвистических данных при каком из математических методов качество реконструируемого дерева деградирует медленней? Предположим, у нас имеются плохо изученные и некачественно описанные языки (довольно обычное дела для регионов вроде Океании или Африки), для которых, тем не менее, мы хотим построить генеалогическую классификацию, пусть и вчерне, — какому из маталгоритмов следует доверять в большей степени (если они порождают отличающиеся друга от друга деревья)?

Для выяснения этого в статье Kassian 2015 был проведен дополнительный тест (см. врез)

> Если в основном тесте, который описан выше, тождество слов (поясните плз это слово или замените другим, если можно) между словами размечалось исходя из этимологического родства или же отсутствия такового, то для дополнительной проверки тождество определялось по автоматизированному алгоритму, замеряющему фонетическое сходство форм (расстояния Левенштейна). Такой машинный подход дает зашумление входной матрицы, что должно исказить филогенетический сигнал.
> Тут были получены более неожиданные результаты. При фонетической расстановке когнаций ближе всего в золотому

стандарту оказались деревья, построенные дистантными (т.е. архаичными) методами: ближайших соседей и невзвешенного попарного среднего. А современные признаковые методы — байесовский MCMC и Maximum parsimony — показали себя более нежными и чувствительными в шуму.

Так или иначе, для более определенного решения поставленных вопросов требуются дополнительные проверки на реальном языковом материале. С дальнейшим развитием проекта «Глобальная лексикостатистическая база данных» в наше распоряжение должно поступать всё больше высококачественных списков Сводеша, пригодных как для лексикостатистических тестов, так и для калибровки глоттохронологического аппарата, датирующего узлы на дереве.

**Источник:**

Kassian A (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
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0116950

Kassian Alexei

PLoS ONE**,** February 26, 2015, DOI: 10.1371/journal.pone.0116950