А. С. Касьян
Институт языкознания РАН (Москва)

К формальной генеалогической классификации лезгинских языков (Северный Кавказ)*

В статье предлагается лексикостатистическая классификация 20 языков и диалектов лезгинской группы северокавказской семьи, выполненная на основе высококачественных 110-словных списков проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database». К лексическому материалу последовательно применяются основные филогенетические методы, как дистантные, так и дискретные: метод ближайших соседей, реализованный в программе Starling (Starling neighbor joining), стандартный метод ближайших соседей (Neighbor joining), метод попарного внутригруппового невзвешенного среднего (Unweighted pair group method with arithmetic mean), метод Монте-Карло с цепями Маркова (Markov chain Monte Carlo), метод невзвешенной максимальной бережливости (Unweighted maximum parsimony). Все методы, кроме последнего, породили деревья, достаточно совместимые между собой, чтобы на их основе можно было составить сводное филогенетическое дерево лезгинских языков. Полученное сводное дерево согласуется с традиционной и некоторыми предшествующими формальными классификациями этой языковой группы. Вопреки теоретическим ожиданиям, метод максимальной бережливости предложил наименее правдоподобное дерево из всех.

Ключевые слова: генеалогическая классификация языков, лексикостатистика, филогения, лезгинские языки.

Данные

В рамках международного научного проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database»[1] были составлены 110-словные списки базисной лексики для 20 языков и диалектов лезгинской группы (северокавказская языковая семья): удинский (2 диалекта), арчинский, крызский (2 диалекта), будухский, цахурский (3 диалекта), рутульский (3 диалекта), агульский (5 диалектов), табасаранский (2 диалекта), лезгинский, см. [Kassian 2011—2012]. Это максимальное количество лезгинских идиомов, для которых можно составить сводешевский список без полевой работы. Двадцать рассматриваемых списков полностью соответствуют лексикографическим стандартам проекта «Глобальная лексикостатистическая база данных».

* Я выражаю искреннюю благодарность Валерию Запорожченко (Москва) и Йоханну-Маттису Листу (Johann-Mattis List; Марбург) за консультации по компьютерным программам филогенетического анализа, а также Дмитрию Лещинеру (Москва) за консультации по ряду математических вопросов. Кавказоведческие аспекты статьи обсуждались с Тимуром Майсаком (Москва). Кроме того, данная работа вряд ли была бы возможна без бесед на смежные или более широкие темы с Георгием Старостиным, Михаилом Живловым, Анной Дыбо, Филиппом Минлосом и другими моими коллегами по московской школе компаративистики.

[1] http://starling.rinet.ru/new100/main.htm

1) Отбор слов производился в соответствии с семантическими спецификациями сводешевского списка, предложенными в [Kassian et al. 2010].

2) Были использованы фактически все релевантные источники по соответствующим языкам — словари, грамматики, собрания текстов, — причем не только современные публикации, но и материалы П. К. Услара, А. Дирра, А. Старчевского и других кавказоведов кон. XIX — нач. XX в.

3) Все языковые формы были единообразно транскрибированы фонетическим алфавитом, базирующимся на системе IPA; формы в традиционных кириллических орфографиях приводятся в скобках.

4) Лексические списки аннотированы. При языковых формах дается ссылка на источник, а в примечаниях эксплицитно обсуждаются существенные фонетические, морфологические и семантические особенности используемых форм и их синонимов. Также в примечаниях цитируются лексические данные из языков и диалектов, опубликованные материалы по которым недостаточны для составления полноценных 110-словных сводешевских списков (например, удинский список сопровожден формами из кавказско-албанских палимпсестов).

В полученных таким образом синхронных списках когнации размечались по этимологическому принципу. Я основывался на пралезгинской реконструкции С. А. Старостина [Starostin & Nikolayev 1994: 122 ff.; S. Starostin 1994; С. Старостин б. д.], по необходимости внося в нее определенные уточнения и изменения, см. [Kassian 2011—2012]. Публикации С. А. Старостина — это единственная на сегодняшний день обнародованная полноценная фонетическая и лексическая реконструкция лезгинского праязыка. Не так давно немецкий кавказовед В. Шульце [Schulze 1988; 2001; Gippert et al. 2008] объявил о разработке своей версии пралезгинской реконструкции. Совокупность лексических этимологий, уже опубликованных в работах В. Шульце, пока недостаточна для окончательных выводов, но я вынужден отметить, что многие диахронические идеи и решения В. Шульце не кажутся мне удачными или приемлемыми.

Для укоренения деревьев в сравнение был введен 110-словный список чеченского литературного языка [G. Starostin 2011]. Чеченский язык был выбран, с одной стороны, как родственный лезгинским в рамках северокавказской семьи, а с другой — как заведомо не входящий в лезгинскую группу. Этимологическое сопоставление чеченского списка с лезгинскими проводилось по [Starostin & Nikolayev 1994] с уточнениями из [G. Starostin 2011].

Метод

Генеалогические деревья строились несколькими методами.

1. Модифицированный метод ближайших соседей, разработанный С. А. Старостиным для лексикостатистического анализа и реализованный в программе Starling (метод Starling neighbor joining, далее — StarlingNJ). См. описание StarlingNJ в [Бурлак & Старостин 2005: 163 сл.] (в настоящей статье используется порог, после которого начинается усреднение значений, не в 70%, а в 75%, что является значением по умолчанию в последних версиях Starling). Дерево строилось в программе Starling (v. 2.5.3; см. [С. Старостин 1993/2007; Бурлак & Старостин 2005: 270 ff.]) из лексикостатистической базы данных, представляющей собой многозначную матрицу с возможностью синонимии (узлы датировались так называемым «экспериментальный методом», при котором сводешевским словам присваиваются индивидуальные индексы стабильности, [С. Старостин 2007a;

G. Starostin 2010]). Дерево укоренялось методом иерархической кластеризации, что стандартно для алгоритма StarlingNJ. Для данных, обработанных в программе Starling, приводятся деревья двух типов: дерево, допускающее только бифуркацию (как того требует метод ближайших соседей), и это же дерево, но в котором соседние узлы сведены в один, если временно́й промежуток между ними составляет 300 или менее лет[2] (300 лет соответствуют замене приблизительно в полтора слова в каждом из двух идиомов). Датировки узлов определялись по принципу строгих молекулярных часов, подробнее об этом и о калибровке шкалы см. [С. Старостин 1989/2007; S. Starostin 1999/2000]. Деревья визуализировались в программе Starling.

2. Стандартный метод ближайших соседей (Neighbor joining, далее NJ), см. [Saitou & Nei 1987; Makarenkov et al. 2006: 65—66]. Дерево строилось в программе SplitsTree4 (v. 4.13.1, см. [Huson & Bryant 2006]) из лексикостатистической бинарной матрицы формата NEXUS, которая была получена из многозначной путем кодирования каждого задействованного в списке пракорня (всего 481 пракорень) как присутствующего («1») или отсутствующего («0») в данном сводешевском значении в соответствующем языке; «?» значит, что в данном языке данное сводешевское значение выражается через иноязычное заимствование или же выражение для него не известно (не найдено в источниках или отсутствует в языке). Для оценки устойчивости топологии дерева использовался непараметрический бутстреп-тест: 10 000 реплик. Дерево укоренялось через внешнюю группу, т. е. через чеченский список. Дерево не датированное. Дерево визуализировалось в программе FigTree (v. 1.4.0). Также в SplitsTree было построено дополнительное дерево методом BioNJ [Gascuel 1997], которое оказалось идентичным дереву NJ.

3. Метод попарного внутригруппового невзвешенного среднего (Unweighted pair group method with arithmetic mean, далее — UPGMA), см. [Sneath & Sokal 1973: 230—234; Makarenkov et al. 2006: 65—66]. Дерево строилось в программе SplitsTree4 (v. 4.13.1) из бинарной матрицы, описанной выше. Для оценки устойчивости топологии дерева использовался непараметрический бутстреп-тест: 10 000 реплик. Дерево укоренялось через внешнюю группу, т. е. через чеченский список. Дерево не датированное. Дерево визуализировалось в программе FigTree (v. 1.4.0).

4. Метод Монте-Карло с цепями Маркова (Markov chain Monte Carlo, далее — MCMC; см. [Makarenkov et al. 2006: 68—69]), впервые примененный в рамках байесовского подхода к лингвистическим данным в [Gray & Atkinson 2003]. Дерево строилось в программе MrBayes (v. 3.2.1, см. [Huelsenbeck & Ronquist 2001]) из бинарной матрицы, описанной выше. Программа запускалась 4 раза, каждый раз с 4 цепочками; чеченский список был указан как внешняя группа. При каждом запуске порождались 5 000 000 случайных конфигураций дерева (поколений), из них каждое 500-е дерево сэмплировалось; при каждом запуске первые 25% деревьев использовались для отжига и исключались из дальнейшего анализа. Дерево укоренялось через внешнюю группу, т. е. через чеченский список. Дерево не датированное. Дерево визуализировалось в программе FigTree (v. 1.4.0).

5. Метод невзвешенной максимальной бережливости (Unweighted maximum parsimony, далее — UMP), см. [Makarenkov et al. 2006: 66—67]. Деревья строились в программе TNT (Willi Hennig Society edition of TNT, v.1.1, 08 May 2013, см. [Goloboff et al. 2008]) из бинарной матрицы, описанной выше, по строгому принципу ветвей и границ (Branch &

[2] На нынешнем этапе разработки метода затруднительно строго определить погрешность датирования, но представляется, что интервал погрешности вряд ли составляет менее чем несколько сотен лет. Также отмечу, что дистантные методы типа StarlingNJ, NJ, UPGMA в принципе способны порождать только бинарные деревья, и это само по себе предполагает, что близкие узлы логично объединять в общий узел.

bound / Implicit enumeration). Принудительная бинаризация узлов была запрещена (Collapse trees after the search); чеченский список был указан как внешняя группа. Было получено 4 оптимальных дерева одинаковой стоимости, на их основе было построено строгое консенсусное дерево. Для оценки устойчивости топологии консенсусного дерева использовался непараметрический бутстреп-тест: 1000 реплик. Деревья укоренялись через внешнюю группу, т. е. через чеченский список. Деревья не датированные. Деревья визуализировались в программе FigTree (v. 1.4.0).

Результаты

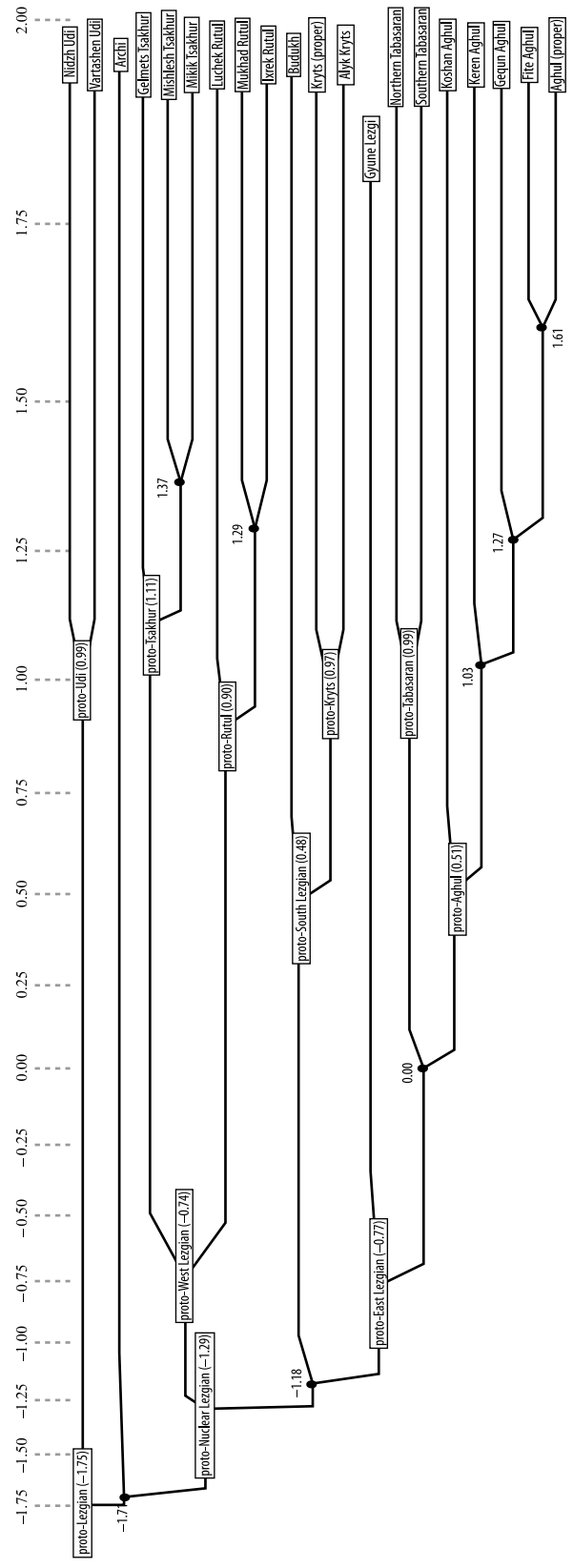
Были получены следующие деревья:

- рис. 1a, метод StarlingNJ без объединения узлов;
- рис. 1b, метод StarlingNJ с объединением близких узлов;
- рис. 2, метод NJ;
- рис. 3, метод UPGMA;
- рис. 4, метод MCMC;
- рис. 5, метод UMP;
- рис. 6, сводное дерево, составленное вручную.

Если исключить дерево UMP (о чем см. ниже), разница между деревьями по большей части не представляется принципиальной. Прокомментируем расхождения.

1) Все дистантные методы, т. е. StarlingNJ, NJ, UPGMA (рис. 1a, 2, 3), предполагают последовательные бифуркации с отделением сначала удинской ветви, затем арчинской и соответствующим обособлением узколезгинского (самурского) праязыка. Дистанция между двумя узлами (отделение удинского и отделение арчинского), однако, минимальна на всех деревьях, что хорошо видно из графического представления деревьев и вероятностной оценки узлов, и при введении временно́й погрешности в 300 лет в методе StarlingNJ (рис. 1b) первичное разделение пралезгинского языка оказывается тернарным: удинский, арчинский и узколезгинский. Напротив, используемый дискретный метод (MCMC, рис. 4) сразу предлагает тернарное разделение на удинский, арчинский и узколезгинский. Следует помнить, что дистантные методы StarlingNJ, NJ, UPGMA в принципе способны порождать только бинарные деревья. Дерево UMP здесь существенно расходится с остальными деревьями, см. ниже.

2) Все методы дают принципиальное членение узколезгинской подгруппы на три ветви: (1) западнолезгинскую (цахурский и рутульский языки); (2) южнолезгинскую (крызский и будухский языки); (3) восточнолезгинскую (агульский, табасаранский и лезгинский языки). Разница обнаруживается в иерархии членения. Методы StarlingNJ и NJ (рис. 1a, 2), а также UMP (рис. 5) указывают на первое отделение западнолезгинской ветви и последующую бифуркацию на южнолезгинскую и восточнолезгинскую ветви. Метод UPGMA (рис. 3) указывает на первое отделение южнолезгинской ветви. Наконец, метод MCMC (рис. 4) указывает на первое отделение восточнолезгинской ветви. Дистанция между двумя узлами (последовательные бифуркации между западнолезгинским, южнолезгинским и восточнолезгинским праязыками), однако, минимальна на всех деревьях, что хорошо видно из графического представления деревьев и вероятностной оценки узлов, и при введении временно́й погрешности в 300 лет в методе StarlingNJ (рис. 1b) разделение узколезгинского праязыка оказывается тернарным: западнолезгинский, южнолезгинский и восточнолезгинский.



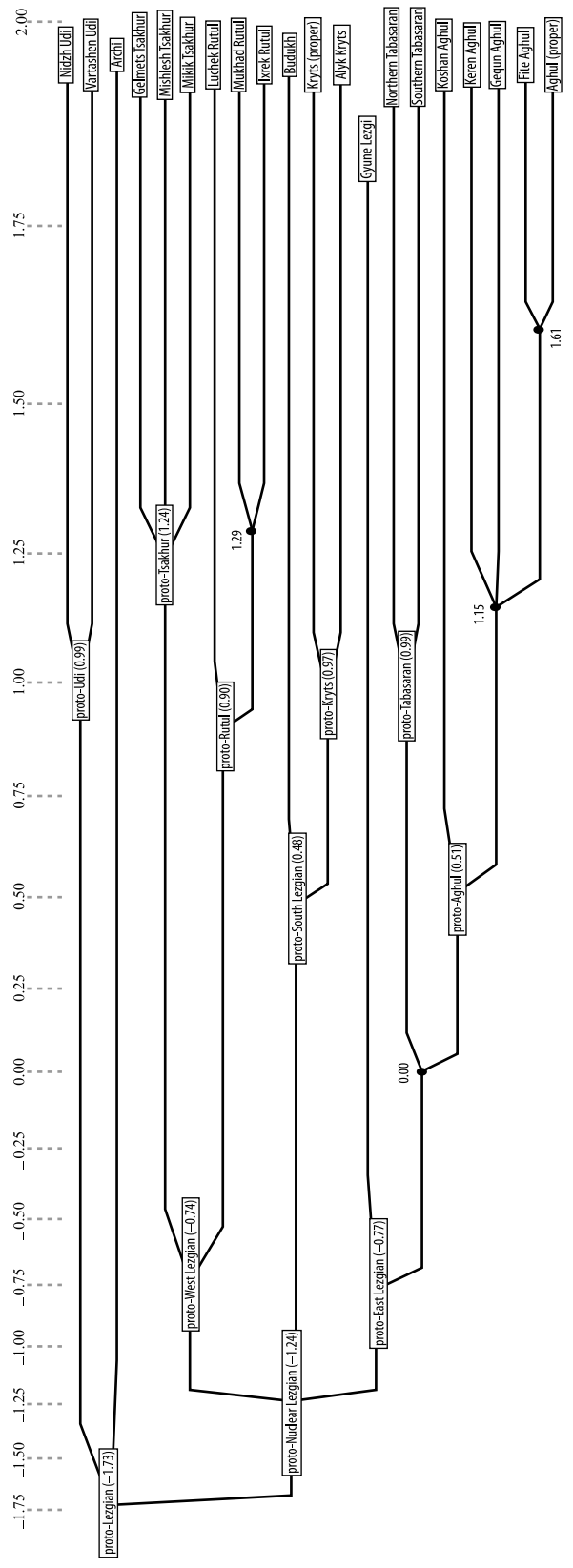


Рис. 1a (слева). Филогенетическое дерево лезгинских идиомов, полученное методом StarlingNJ в Starling без объединения близких узлов из многозначной матрицы. Даты даны в тысячах лет (например, –1.72 означает 1720 г. до н. э.).

Рис. 1b (справа). Филогенетическое дерево лезгинских идиомов, полученное методом StarlingNJ в Starling с объединением близких узлов (дистанция 300 или менее лет) из многозначной матрицы. Даты даны в тысячах лет (например, –1.72 означает 1720 г. до н. э.).

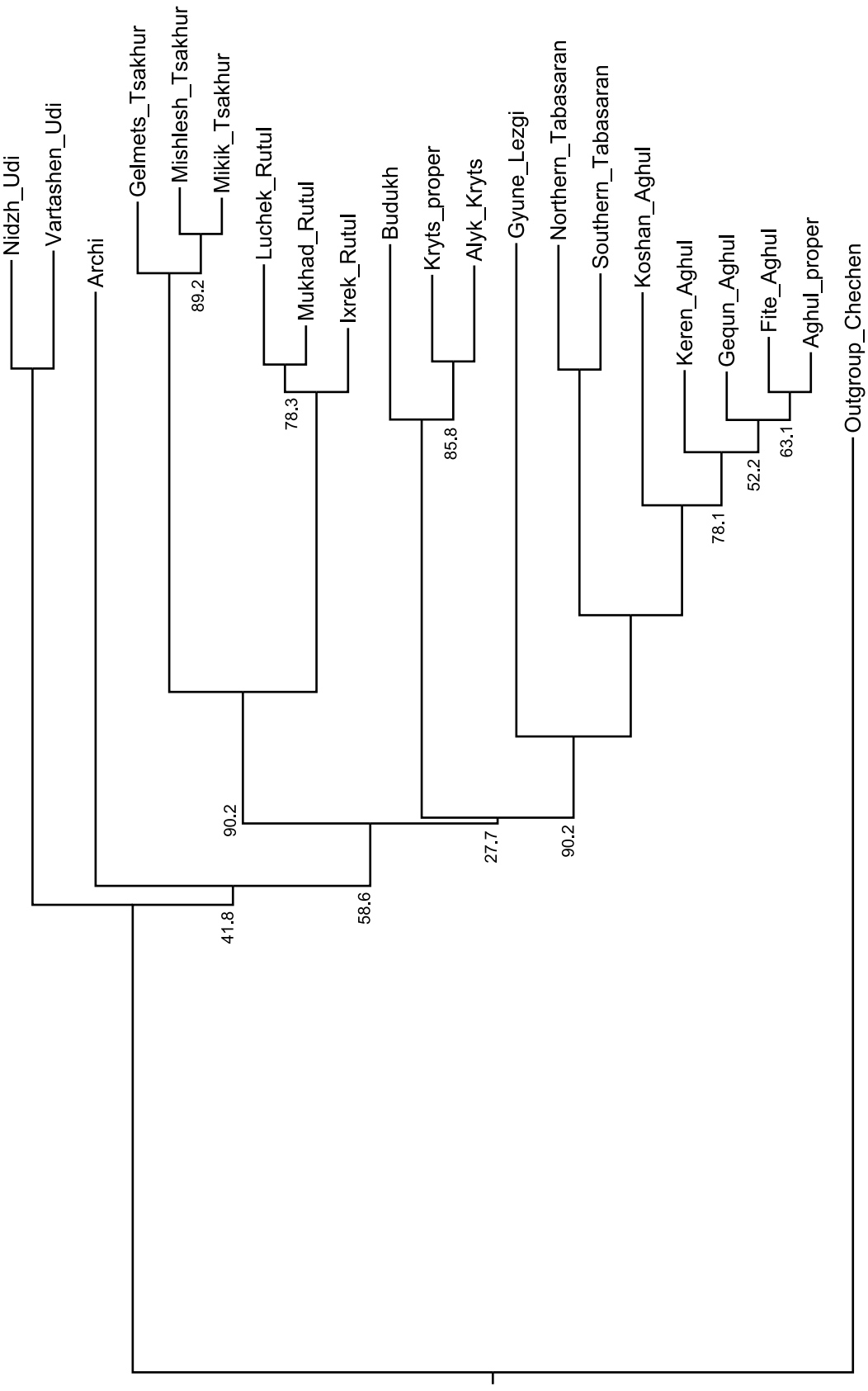


Рис. 2. Филогенетическое дерево лезгинских идиомов, полученное методом NJ в SplitsTree4 из бинарной матрицы. Статистическая поддержка той или иной ветви в процентах реплик бутстрепа обозначена числом при соответствующей ветви (для сверхустойчивых ветвей с поддержкой ≥ 95% этот параметр не указывается). Длина ветвей выражает относительное количество предполагаемых машиной лексических замен. Метод BioNJ дает идентичную топологию.

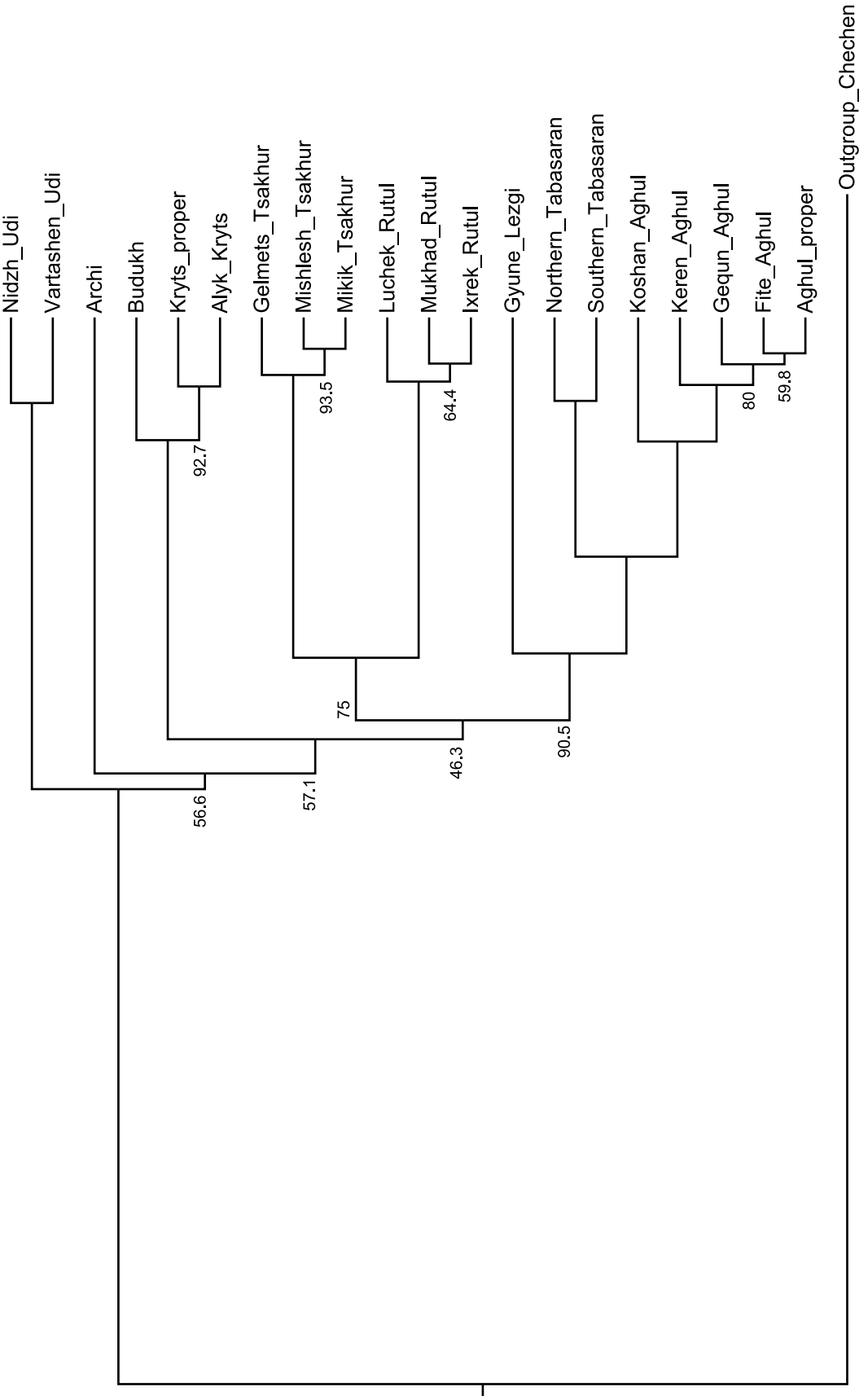


Рис. 3. Филогенетическое дерево лезгинских идиомов, полученное методом UPGMA в SplitsTree4 из бинарной матрицы. Статистическая поддержка той или иной ветви в процентах реплик бутстрепа обозначена числом при соответствующей ветви (для сверхустойчивых ветвей с поддержкой ≥ 95% этот параметр не указывается). Длина ветвей выражает относительное количество предполагаемых машиной лексических замен.

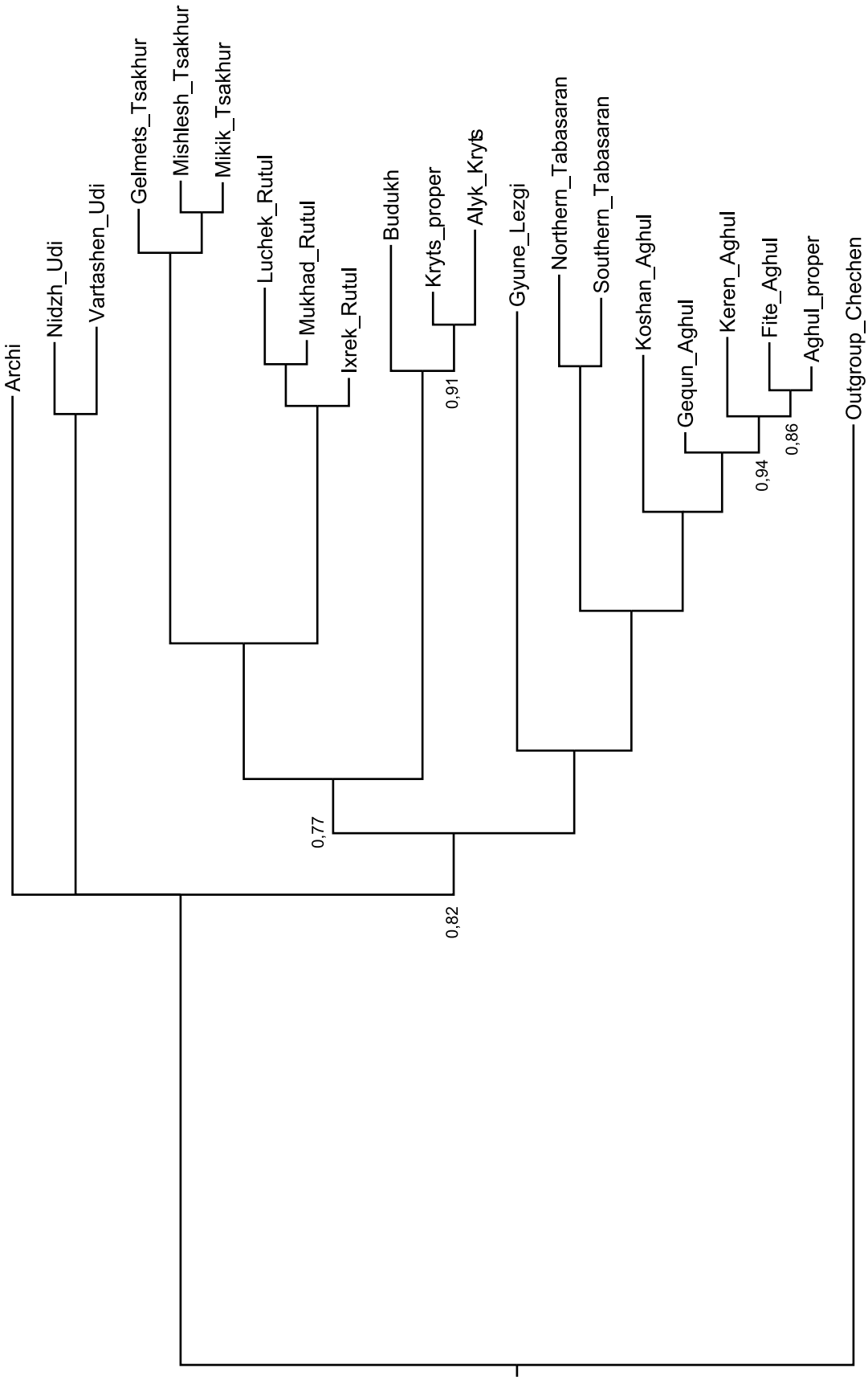


Рис. 4. Консенсусное филогенетическое дерево лезгинских идиомов, полученное методом MCMC в MrBayes из бинарной матрицы. Апостериорная вероятность той или иной ветви обозначена числом над соответствующей ветвью (вероятность для сверхустойчивых ветвей с $P \geq 0{,}95$ не указывается). Длина ветвей выражает относительное количество предполагаемых машиной лексических замен.

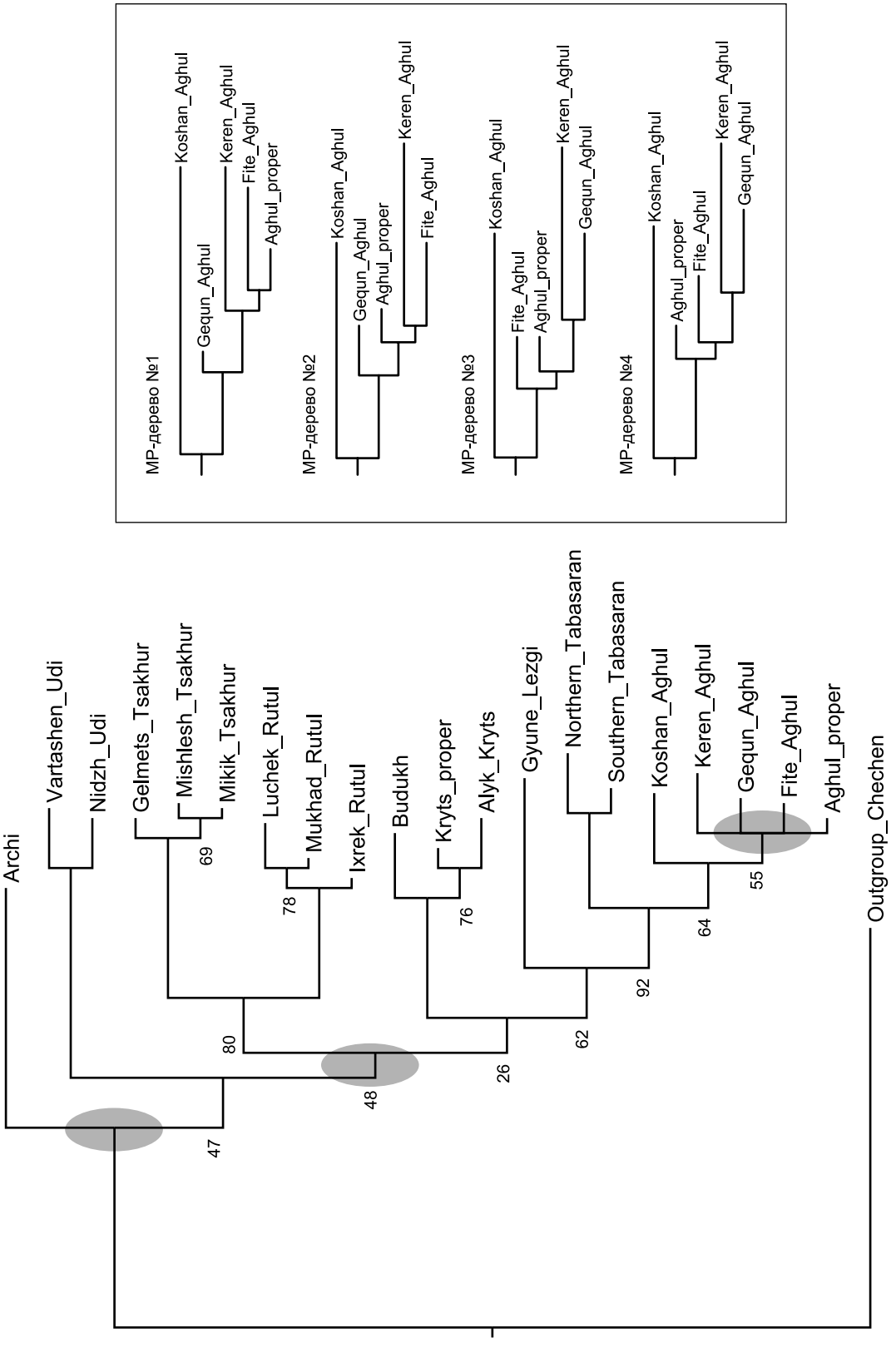


Рис. 5. Строгое консенсусное филогенетическое дерево лезгинских идиомов, полученное методом UMP в TNT из бинарной матрицы. Статистическая поддержка той или иной ветви в процентах реплик бутстрепа обозначена числом при соответствующей ветви (для сверхустойчивых ветвей с поддержкой ≥ 95% этот параметр не указывается). Длина ветвей выражает относительное количество предполагаемых машиной лексических замен. Четыре оптимальных дерева имеют различия только в агульском узле, что продемонстрировано на врезке. Серым выделены фрагменты дерева, проблемные при сопоставлении с другими методами.

3) Агульские диалекты. Все методы реконструируют первичное отделение кошанского диалекта (что соответствует интуитивным ожиданиям), но далее начинают противоречить друг другу. Дистантные методы, т. е. StarlingNJ, NJ, UPGMA (рис. 1a, 2, 3), предполагают последующее отделение керенского диалекта и далее гехюнского диалекта, а используемый дискретный метод (MCMC, рис. 4) наоборот — сначала отделение гехюнского диалекта и затем керенского. Дистанция между двумя узлами (последовательные бифуркации между керенским, гехюнским и собственно агульским/фитинским), однако, минимальна на всех деревьях, что хорошо видно из графического представления деревьев, и при введении временно́й погрешности в 300 лет в методе StarlingNJ (рис. 1b) разделение праагульского после отделения кошанского диалекта оказывается тернарным: керенский, гехюнский и собственно агульский/фитинский. Дерево UMP здесь существенно расходится с остальными деревьями, см. ниже.

4) Пожалуй, наиболее серьезное расхождение между имеющимися деревьями (исключая дерево UMP) касается членения трех рутульских диалектов. Методы StarlingNJ и UPGMA (рис. 1a, 3) предполагают, что первым отделился лучекский диалект. Напротив, NJ и MCMC (рис. 2, 4) предполагают, что первым отделился ихрекский диалект. При этом на рис. 1a (StarlingNJ) узлы находятся хронологически достаточно далеко друг от друга и не объединяются в один при введении временно́й погрешности в 300 лет (рис. 1b). Как видно из таблиц дистанций, и для многозначной, и для бинарной матрицы в рутульской части дерева получаемые лексикостатистические дистанции не удовлетворяют условию аддитивности: табл. 1, 2. При постулате постоянной скорости изменения сводешевского списка рутульская ситуация является аномальной, и разные методы классификации предлагают в таком случае разные решения. Лингвистически рутульские данные могут объясняться двумя возмущающими факторами: (1) междиалектными заимствованиями и контактно обусловленной гомоплазией (выявить такие случаи пока не представляется возможным); (2) несовершенством имеющихся лексикографических описаний, не позволяющих составить сводешевские списки более аккуратно. Сложно сказать, какая из двух топологий рутульских диалектов адекватнее соответствует исторической реальности (ср. [Ибрагимов 1978: 15]), однако в любом случае математические методы классификации вряд ли в полной мере применимы к ситуации взаимопонятных контактирующих идиомов (диалектного континуума), как то наблюдается на рутульской территории.

Табл. 1. Таблица обратных дистанций для трех рутульских диалектов при многозначной матрице.

	Ихрекский	Лучекский
Мухадский	0,96	0,94
Ихрекский	—	0,91

Табл. 2. Таблица обратных дистанций для трех рутульских диалектов при бинарной матрице.

	Ихрекский	Лучекский
Мухадский	0,92	0,91
Ихрекский	—	0,87

5) Наконец, в изолированном положении оказывается метод UMP, который порождает дерево, слабо совместимое как с деревьями, полученными остальными методами, так и с нашими неформальными интуитивными представлениями о членении лезгинской

языковой группы (см. рис. 5, где проблемные узлы выделены серым). Во-первых, арчинский язык оказывается первым отделившимся кластером, а удинский, напротив, хоть формально и отделяется вторым, стремится объединиться со следующим узколезгинским (самурским) узлом. Во-вторых, не удается выстроить осмысленную консенсусную топологию агульских диалектов. В-третьих, бросается в глаза неустойчивость консенсусного дерева в некоторых принципиальных узлах (вроде отделения арчинского и удинского), где бутстреп-тест демонстрирует поддержку < 50%. По этим причинам я был вынужден исключить дерево UMP из рассмотрения при составлении сводного лезгинского дерева (рис. 6).

С учетом вышеперечисленных расхождений, отбросив результаты анализа методом UPM, можно предложить сводное филогенетическое дерево лезгинских идиомов: рис. 6 (составлено вручную). На этом дереве объединены соседние узлы, (1) хронологическое расстояние между которыми ≤ 300 лет по подсчетам методом StarlingNJ (см. рис. 1a, 1b) или (2) топология которых зависит от используемого метода классификации. Серым цветом отмечены 4 объединенных тернарных узла, за которыми скрываются бинарные ветвления, различающиеся в зависимости от метода: три из этих узлов автоматически получаются при введении указанной временно́й погрешности, а четвертый узел — это рутульские диалекты, обсуждаемые выше. Как можно видеть, сводное дерево (рис. 6) идентично дереву StarlingNJ (см. рис. 1b), за исключением дополнительного объединения в тернарный узел трех рутульских диалектов.

Предыдущие классификации

Полученное сводное дерево лезгинских языков и диалектов (рис. 6) с двумя аутлайерами (удинский и арчинский) и многочисленной узколезгинской или самурской подгруппой, делящейся на три кластера (западный, южный, восточный), согласуется с такими предлагаемыми ранее филогенетическими реконструкциями лезгинской языковой группы:

1) Традиционная неформальная классификация, см., напр., [Талибов 1980: 11—16] с дальнейшей литературой.

2) Предшествующие более грубые лексикостатистические подсчеты, упомянутые в [Алексеев 1984: 91 сл.], на основе 100-словных списков, проэтимологизированных и обработанных методом типа UPGMA; списки не удовлетворяют современным критериям проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database».

3) Формальная классификация в проекте «The Automated Similarity Judgment Program», см. [Müller et al. 2010], где используются непроэтимологизированные 40-словные списки, суммарное измерение расстояний Левенштейна между которыми дает матрицу дистанций между языками, из которой строится дерево методом NJ в программе MEGA 4[3].

Напротив, предшествующие лексикостатистические классификации, согласно которым арчинский оказывается четвертым кластером внутри узколезгинской подгруппы, не подтверждаются и, видимо, должны быть отвергнуты. Речь идет о [Алексеев 1985: 17—23] (100-словные списки, проэтимологизированные и обработанные методом типа UPGMA) и [Коряков 2006: 21] (100 или 110-словные списки, проэтимологизированные и обработанные методом StarlingNJ в программе Starling); в обеих публикациях списки не удовлетворяют современным критериям проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database».

[3] Отметим, что в [Müller et al. 2010] в лезгинскую группу попал и хиналугский язык, что, видимо, неверно.

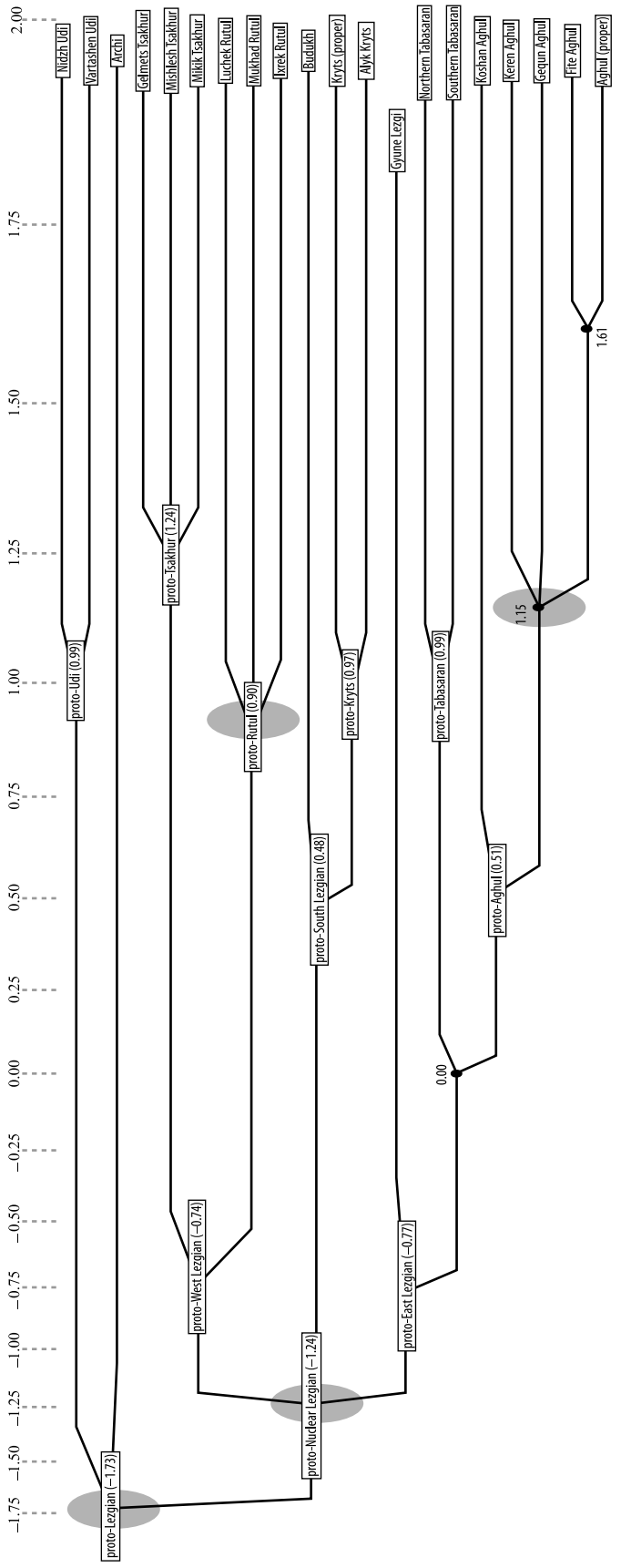


Рис. 6. Сводное дерево лезгинских идиомов, отражающее классификацию методами StarlingNJ, NJ, BioNJ, UPGMA, MCMC (но не UMP). Серым отмечены 4 объединенных тернарных узла, за которыми скрываются бинарные ветвления, различающиеся в зависимости от метода. Датировки даются по методу StarlingNJ в тысячах лет (например, –1.72 означает 1720 г. до н. э.).

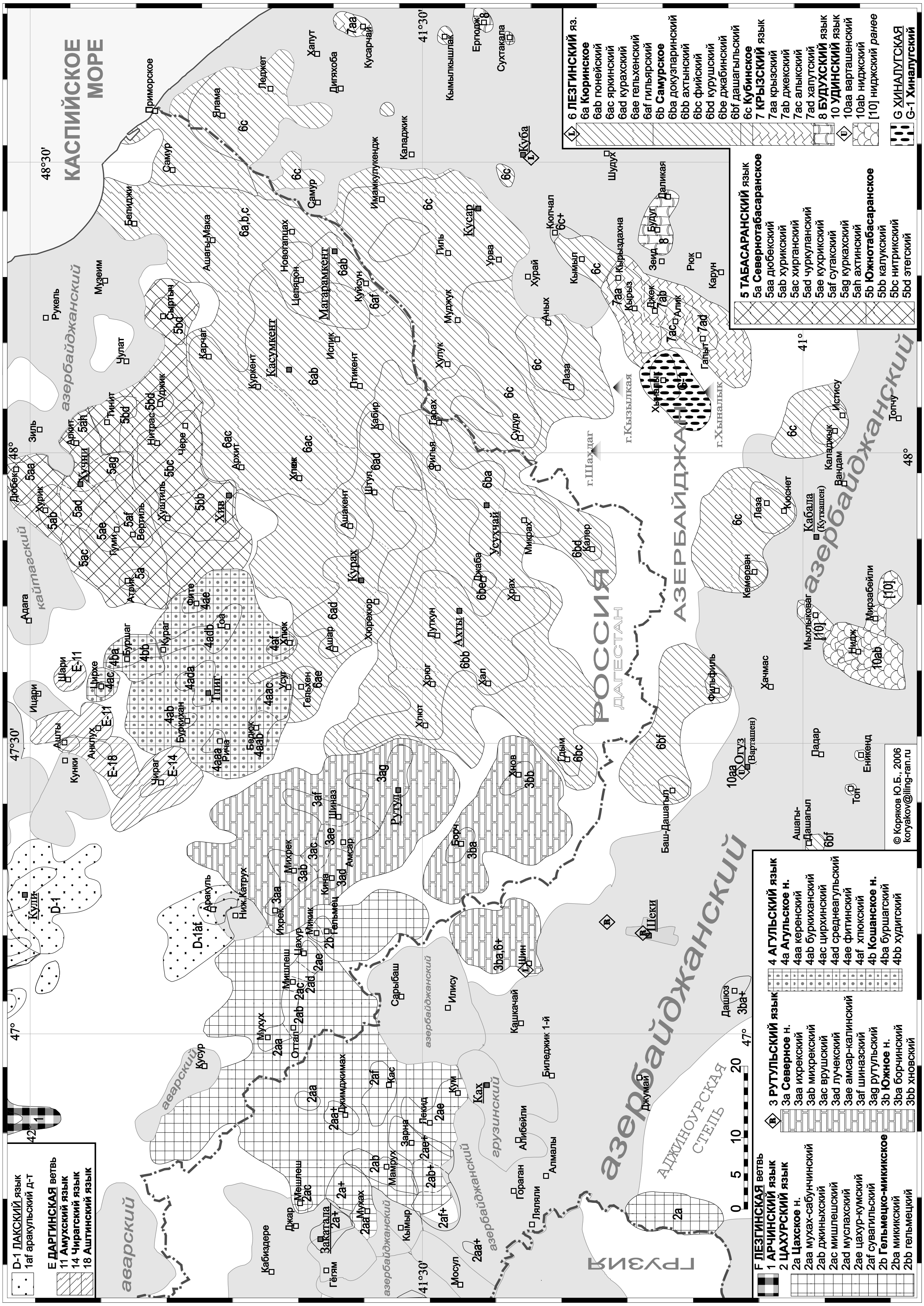


Рис. 7. Карта современного распределения лезгинских языков (по кн.: [Коряков 2006: карта № 13]).

Заведомо не находит поддержки идея В. Шульце [Schulze 2005; Gippert et al. 2008: II-65—75; против Schulze-Fürhoff 1994: 450] о том, что удинско-кавказско-албанская ветвь входит в восточнолезгинский кластер вместе с агульским, табасаранским и собственно лезгинским. В. Шульце [Gippert et al. 2008: II-65—75] опубликовал составленные им сводешевские списки для удинского и кавказско-албанского в сравнении с данными других лезгинских языков. К сожалению, В. Шульце не приводит никаких объяснений своей особой версии лексикостатистической процедуры, а в то же время лексикографическое качество его сводешевских списков весьма низко. Всё это позволяет заключить, что В. Шульце не смог представить какие-либо формальные аргументы в пользу своей филогенетической классификации. С неформальной, т. е. интуитивной точки зрения идея В. Шульце о месте удинско-кавказско-албанского внутри лезгинской группы также не представляется приемлемой.

Выводы

Как было частично указано выше, база данных с 110-словными лексическими списками лезгинских языков и диалектов [Kassian 2011—2012] обладает рядом важных свойств:

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) не будет большим преувеличением сказать, что среди кавказоведов имеется консенсус относительно членения лезгинской группы (аутлайеры удинский и арчинский, отделившиеся первыми, плюс узколезгинская, или самурская, подгруппа, состоящая из трех кластеров: западного, южного и восточного);

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;

4) общее качество, так сказать, степень «очистки» лезгинских лексикостатистических списков (равно как и списков других языков в проекте «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database») беспрецедентно высока для мировой лингвистики.

Всё это делает лезгинскую базу данных прекрасным полигоном для лингвистической апробации различных филогенетических методов.

В теоретической статье [Barbançon et al. 2013] симулированием различных лингвистических ситуаций сравнивается адекватность нескольких филогенетических методов. Авторы приходят к выводу, что по всем параметрам наиболее надежным является метод максимальной бережливости (MP), далее следует метод Монте-Карло с цепями Маркова (MCMC), затем метод ближайших соседей (NJ) и существенно менее точным методом оказывается метод попарного внутригруппового невзвешенного среднего (UPGMA). Оставляя в стороне некоторые спорные аспекты статьи[4], можно видеть, что авторы явным

[4] Одной из причин трудностей в применении метода MP с бинарными признаками к лингвистическим данным может быть неадекватность соответствующей модели представлениям об эволюции естественных языков. Метод MP сильнее других зависит от гомоплазии (т. е. от обратного или же параллельного развития), соответственно, для минимализации эффекта гомопластических возмущений [Barbançon et al.

образом отдают предпочтение дискретным методам (MP, MCMC) перед дистантными (NJ, UPGMA), и это, в принципе, является главным выводом публикации [Barbançon et al. 2013][5]. В качестве количественной оценки в [Barbançon et al. 2013: 166] предлагается считать, что все протестированные методы, кроме UPGMA, реконструируют порядка 90% ребер истинного с исторической точки зрения дерева.

Эксперименты с лезгинской лексикостатистической базой данных показывают, однако, еще более отрадную картину, если счесть, что каждое из ребер «истинного» дерева реконструировано хотя бы одним из методов, исключая UMP (т. е. отражено хотя бы на одном из деревьев на рис. 1—4). При введении небольшого доверительного интервала (и объединении соседних узлов, попавших в него, см. рис. 6 и комментарии к нему) видно, что протестированные методы StarlingNJ, NJ, UPGMA, MCMC противоречат друг другу только в иерархии трех рутульских диалектов. На сводном дереве на рис. 6 всего 33 ребра плюс, если мы не будем объединять рутульские диалекты в тернарный узел, дополнительное 1 ребро, таким образом 34 ребра. Расхождение между методами в членении рутульских диалектов дает ошибку в 1 ребре из 34, и это предполагает, что все методы (за исключением UMP) правильно реконструировали от 97% до 100% ребер. Неожиданным результатом лезгинского теста оказалась невысокая правдоподобность дерева, полученного методом максимальной бережливости (UMP, рис. 5), что прямо противоречит выкладкам [Barbançon et al. 2013].

Рассмотренные лезгинские данные подтверждают некоторые положения, составляющие идеологическую основу проекта «Глобальная лексикостатистическая база данных / The Global Lexicostatistical Database»:

1) При классификации языков лучше воздерживаться от использования грамматических (фонетических, морфологических, синтаксических) признаков[6], т. к., во-первых, эти признаки не универсальны; во-вторых, они 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

2013] предлагают вводить индивидуальные веса признаков (Weighted maximum parsimony), присваивая больший внутренний вес признакам, не демонстрирующим гомоплазию на данном лингвистическом материале. На бинарных данных с одинаковым весом перехода между состояниями наличие гомоплазии (параллельное или же обратное развитие какого-либо признака) математически эквивалентно присутствию в исходной матрице двух несовместимых (incompatible) признаков, т. е. принимающих все четыре возможные пары состояний «00», «01», «10» и «11» (см., напр., [Semple & Steel 2003: 69 сл.]). Использованный в [Barbançon et al. 2013] метод MP трактует переходы между состояниями «0» и «1» как равнозначные. Но если мы работаем, как в данной статье, с бинарной матрицей, где «1» значит маркированное значение признака, а «0» — немаркированное (например, «1» = наличие, а «0» = отсутствие данного пракорня в данном сводешевском значении в данном языке), то переход 1 > 0 (утрата корня) не является значимым событием, он может происходить параллельно и независимо в разных языках, и такую параллельную утрату лингвист вряд ли должен считать гомопластичной. Таким образом, чтобы обнаружить факт *лингвистической* гомоплазии, надо произвести реконструкцию значений используемых признаков для праязыка исследуемой языковой группы, что совсем не тривиальная теоретическая и практическая задача, в частности реконструкция невозможна без знания филогенетического дерева [Kassian 2013], — мы попадаем в порочный круг.

[5] Схожий взгляд на иерархию точности классификационных методов постепенно возобладает и в молекулярной биологии.

[6] Ср., например, [Nakhleh et al. 2005], где различные методы филогенетического анализа тестируются на материале индоевропейской семьи, при этом на вход подаются как лексические, так и грамматические признаки. Аналогично смешанный набор лексических и грамматических признаков используется в моделировании [Barbançon et al. 2013].

с высокой вероятностью влечет за собой изменение других признаков. Для лексических же признаков эти недостатки характерны в значительно меньшей степени.

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 (хотя отдельные филогенетические методы, вроде максимальной бережливости / maximum parsimony, вызывают сомнения).

Дополнительные материалы по исследованию доступны по адресам:

- http://jolr.ru/article.php?id=XXXX
- http://starling.rinet.ru/~kass/Lezgian_classification_RUS_2013.zip

Архив включает в себя:

- lez.xls, многозначная матрица в формате MS Excel;
- lez.nex, бинарная матрица в формате NEXUS;
- lez.tnt, бинарная матрица в формате NEXUS для программы TNT;
- lez-reverse-distances-multistate.xls, таблица обратных расстояний, полученная из многозначной матрицы в программе Starling;
- lez-distances-binary.txt, таблица расстояний, полученная из бинарной матрицы в программе SplitsTree4;
- *.tre, некоторые обсуждаемые в статье деревья в формате NEWICK.

Литература

Алексеев 1984 — М. Е. Алексеев. К вопросу о классификации лезгинских языков // *Вопросы языкознания*, 1984, № 5. С. 88—94. [M. E. Alekseev. K voprosu o klassifikatsii lezginskikh yazykov // *Voprosy yazyknaniya*, 1984, № 5. S. 88—94.]

Алексеев 1985 — М. Е. Алексеев. *Вопросы сравнительно-исторической грамматики лезгинских языков. Морфология, Синтаксис*. Москва, 1985. [M. E. Alekseev. *Voprosy sravnitel'no-istoricheskoj grammatiki lezginskikh yazykov. Morfologiya, Sintaksis*. Moskva, 1985.]

Бурлак & Старостин 2005 — С. А. Бурлак, С. А. Старостин. *Сравнительно-историческое языкознание*. 2-е изд. Москва, 2005. [S. A. Burlak, S. A. Starostin. *Sravnitel'no-istoricheskoe yazykoznanie*. 2-e izd. Moskva, 2005.]

Ибрагимов 1978 — Г. Х. Ибрагимов. *Рутульский язык*. М., 1978. [G. Kh. Ibragimov. *Rutul'skij yazyk*. M., 1978.]

Коряков 2006 — Ю. Б. Коряков. *Атлас кавказских языков. С приложением полного реестра языков*. Москва, 2006. [Yu. B. Koryakov. *Atlas kavkazskikh yazykov. S prilozheniem polnogo reestra yazykov*. Moskva, 2006.]

С. Старостин 1989/2007 — С. А. Старостин. Сравнительно-историческое языкознание и лексикостатистика // Старостин 2007. С. 407—447. [S. A. Starostin. Sravnitel'no-istoricheskoe yazykoznanie i leksikostatistika // Starostin 2007. S. 407—447.] [Впервые опубл. в: *Лингвистическая реконструкция и древнейшая история Востока*. М., Наука, 1989: 3—39. Статья опубликована также по-английски, см. S. Starostin 1999/2000.]

С. Старостин 1993/2007 — С. А. Старостин. Рабочая среда для лингвиста // Старостин 2007. С. 481—496. [S. A. Starostin. Rabochaya sreda dlya lingvista // Starostin 2007. S. 481—496.] [Впервые опубл. в: Базы данных по истории Евразии в средние века, вып. 2. М., Институт востоковедения РАН, 1993: 50—64. Перепечатано в: Гуманитарные науки и новые информационные технологии. М., РГГУ, 1994: 7—23.]

С. Старостин 2007 — С. А. Старостин. *Труды по языкознанию*. Москва, 2007. [S. A. Starostin. *Trudy po yazykoznaniyu*. Moskva, 2007.]

С. Старостин 2007а — С. А. Старостин. Определение устойчивости базисной лексики // Старостин 2007. С. 827—839. [S. A. Starostin. Opredelenie ustojchivosti bazisnoj leksiki // Starostin 2007. S. 827—839.]

С. Старостин б. д. — С. А. Старостин. *Историческая фонетика лезгинских языков*. Машинопись, 1980-е гг. [S. A. Starostin. *Istoricheskaya fonetika lezginskikh yazykov*. Mashinopis', 1980-e gg.]

Талибов 1980 — Б. Б. Талибов. *Сравнительная фонетика лезгинских языков*. Москва, 1980. [B. B. Talibov. *Sravnitel'naya fonetika lezginskikh yazykov*. Moskva, 1980.]

Barbançon et al. 2013 — François Barbançon, Steven N. Evans, Luay Nakhleh, Don Ringe, Tandy Warnow. An experimental study comparing linguistic phylogenetic reconstruction methods // *Diachronica* 30/2 (2013). P. 143—170.

Gascuel 1997 — O. Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data // *Molecular Biology and Evolution* 14 (1997). P. 685—695.

Gippert et al. 2008 — J. Gippert, W. Schulze, Z. Aleksidze, J.-P. Mahé. *The Caucasian Albanian Palimpsests of Mt. Sinai*. 2 vols. Brepols, 2008.

Goloboff et al. 2008 — P. A. Goloboff, J. S. Farris, K. C. Nixon. TNT, a free program for phylogenetic analysis // *Cladistics* 24/5 (2008). P. 774—786.

Gray & Atkinson 2003 — Russell D. Gray & Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin // *Nature* 426 (27 November 2003). P. 435—439.

Huelsenbeck & Ronquist 2001 — J. P. Huelsenbeck, F. Ronquist. MrBayes: Bayesian inference of phylogenetic trees // *Bioinformatics* 17/8 (2001). P. 754—755.

Huson & Bryant 2006 — D. H. Huson & D. Bryant. Application of phylogenetic networks in evolutionary studies // *Molecular Biology and Evolution* 23/2 (2006). P. 254—267.

Kassian 2011—2012 — *Annotated Swadesh wordlists for the Lezgian group (North Caucasian family)*. Database compiled and annotated by A. Kassian (November 2011 — October 2012). The Global Lexicostatistical Database project: http://starling.rinet.ru/cgi-bin/response.cgi?root=new100&morpho=0&basename=new100\ncc\lez&limit=-1

Kassian 2013 — A. Kassian. The Lezgian linguistic group within the framework of the Global Lexicostatistical Database. Talk at the conference *Comparative-Historical Linguistics of the 21st Century: Issues and Perspectives*, Moscow, March 20—22, 2013: http://www.academia.edu/3040336/

Kassian et al. 2010 — A. Kassian, G. Starostin, A. Dybo, V. Chernov. The Swadesh wordlist. An attempt at semantic specification // *Journal of Language Relationship*, No. 4 (2010). P. 46—89.

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

Müller et al. 2010 — André Müller, Søren Wichmann, Viveka Velupillai, Cecil H. Brown, Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias Urban, Helen Geyer, Anthony Grant. *ASJP World Language Tree of Lexical Similarity*. Version 3 (July 2010). http://email.eva.mpg.de/~wichmann/language_tree.htm

Nakhleh et al. 2005 — Luay Nakhleh, Tandy Warnow, Donald Ringe & Steven N. Evans. A Comparison of Phylogenetic Reconstruction Methods on an IE Dataset // *The Transactions of the Philological Society* 103 (2005). P. 171—192.

Saitou & Nei 1987 — N. Saitou & M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees // *Molecular Biology and Evolution* 4 (1987). P. 406—425.

Schulze 1988 — W. Schulze. *Studien zur Rekonstruktion des Lautstandes der südostkaukasischen (lezgischen) Grundsprache*. Habilitationsschrift Universität Bonn. Unpubl. ms.

Schulze 2001 — W. Schulze. *The Udi Gospels. Annotated text, etymological index, lemmatized concordance*. München/ Newcastle: Lincom, 2001.

Schulze 2005 — W. Schulze. *A Functional Grammar of Udi*. Unpubl. ms., 2005. Available on demand at author's site: http://www.lrz.de/~wschulze/FGU.htm

Schulze-Fürhoff 1994 — W. Schulze-Fürhoff. Udi // R. Smeets (ed.). *The Indigenous Languages of the Caucasus*, vol. 4. Caravan Books, 1994. P. 447—514.

Semple & Steel 2003 — Ch. Semple, M. Steel. *Phylogenetics*. Oxford University Press, 2003.

Sneath & Sokal 1973 — P. H. A. Sneath & R. R. Sokal. *Numerical Taxonomy*. San Francisco: W.H. Freeman and Company, 1973.

G. Starostin 2010 — G. S. Starostin. Preliminary lexicostatistics as a basis for language classification: A new approach // *Journal of Language Relationship*, No. 3 (2010). P. 79—116.

G. Starostin 2011 — *Annotated Swadesh wordlists for the Nakh group (North Caucasian family)*. Database compiled and annotated by G. Starostin (last revision: October 2011). The Global Lexicostatistical Database project: http://starling.rinet.ru/cgi-bin/response.cgi?root=new100&morpho=0&basename=new100\ncc\nah&limit=-1

S. Starostin 1994 — S. A. Starostin. *Lezgian Etymological Database*. Published in [Starostin & Nikolayev 1994]; available at: http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl

S. Starostin 1999/2000 — S. A. STAROSTIN. Comparative-historical linguistics and lexicostatistics // *Historical Linguistics and Lexicostatistics*. Melbourne, 1999. P. 3—50 [republ. in: *Time Depth in Historical Linguistics*. McDonald Institute for Archaeological Research, Oxford, 2000, p. 223—259.]

Starostin & Nikolayev 1994 — S. A. STAROSTIN, S. L. NIKOLAYEV. *A North Caucasian Etymological Dictionary*. Moscow, 1994 [reprinted: 3 vols. Ann Arbor: Caravan Books, 2007]. Available online at the Tower of Babel project as Caucet.dbf: http://starling.rinet.ru/cgi-bin/main.cgi?flags=eygtnnl

Alexei KASSIAN. Towards a formal genealogical classification of Lezgian languages of the North Caucasus.

A lexicostatistical classification is proposed for 20 languages and dialects of the Lezgian group of the North Caucasian family, based on meticulously compiled 110-item wordlists, published as part of the *Global Lexicostatistical Database* project. The lexical data have been subsequently analyzed with the aid of the principal phylogenetic methods, both distance-based and character-based: Starling neighbor joining (StarlingNJ), Neighbor joining (NJ), Unweighted pair group method with arithmetic mean (UPGMA), Markov chain Monte Carlo (MCMC), Unweighted maximum parsimony (UMP). All these methods, with the exception of UMP, have yielded trees that are sufficiently compatible with each other to generate a summary phylogenetic tree of the Lezgian lects. The obtained summary tree agrees with the traditional classification as well as some of the previously proposed formal classifications of this linguistic group. Contrary to theoretical expectations, the UMP method has suggested the least plausible tree of all.

Keywords: language classification, lexicostatistics, phylogeny, Lezgian languages.