УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
Том 151, кн. 3 Физико-математические науки *2009*

УДК 81:004.9

ПРЕОБРАЗОВАНИЕ МЕТРИК, ИСПОЛЬЗУЕМЫХ В МЕТОДАХ КЛАСТЕРИЗАЦИИ ДЛЯ ПОСТРОЕНИЯ ФИЛОГЕНЕТИЧЕСКИХ ДЕРЕВЬЕВ ЯЗЫКОВ

В.Д. Соловьев, Р.Ф. Фасхутдинов

Аннотация

С появлением несколько лет назад больших типологических баз данных возникла проблема выбора математических средств извлечения из них знаний (в форме кластеризации языков). Обычно для этих целей используются филогенетические алгоритмы, основанные на метрике Хемминга. Однако в кластерном анализе было показано, что некоторые другие метрики дают лучшие результаты. В статье введены две новые метрики и на большом числе реальных лингвистических примерах продемонстрировано, что филогенетические алгоритмы, основанные на этих метриках, дают лучшие результаты.

Ключевые слова: лингвистические базы данных, метрики, филогенетические алгоритмы.

Введение

В последние годы для исследователей стали доступны две большие типологические базы данных: «Языки мира» и WALS (World Atlas of Linguistic Structures – Всемирный атлас языковой структуры). На сегодняшний день они являются крупнейшими из лингвистических баз данных, описывающих грамматические свойства языков.

В базе «Языки мира» [1] содержится 315 языков Евразии, каждый из которых описан 3821 признаком. Все признаки представлены в бинарном виде и относятся к одной из трех сфер описания языка: фонетика, морфология, синтаксис. Всего в базе данных представлено одиннадцать языковых семей.

В WALS [2] содержится 2559 языков из всех языковых семей мира. Описание языков в ней менее детализировано: количество признаков менее 1500 (при переводе их в бинарный формат), а для многих языков указаны значения лишь немногих признаков.

С появлением этих баз данных стало возможным применение новых для лингвистики методов, которые могут помочь в решении ряда проблем, оставшихся нерешенными ранее, таких, например, как построение дерева эволюции языков на глубину более 10 тысяч лет. Среди таких методов наиболее популярными являются методы кластеризации, а также разработанные в рамках эволюционной биологии специальные методы реконструкции эволюционных деревьев – максимальной бережливости [3], максимального правдоподобия [3] и байесовский анализ [4]. Каждый из этих методов реконструирует филогенетическое дерево – граф, являющийся деревом, листья которого помечены названиями существующих языков, а внутренние вершины соответствуют протоязыкам. Самыми широко используемыми методами последовательной кластеризации являются метод невзвешенного парного арифметического среднего (unweighted pair-group method using arithmetic

averages, сокращенно, upgma [3]) и метод ближнего соседа (neighbor-joining, nj [3]). Они, как и все методы последовательной кластеризации, основаны на вычислении расстояний между сравниваемыми объектами, то есть для их применения необходимо задать матрицу расстояний между объектами. Для этого чаще всего используют метрику Хемминга. В работе [5] введена λ-метрика (основанная на гипотезе λ-компактности), которая во многих случаях дает лучшие результаты. В целом следует отметить, что все используемые в настоящее время методы не дают достаточно точных и надежных результатов, что заставляет искать новые подходы. В настоящей статье приведены результаты кластеризации с использованием новых матриц расстояний, полученных трансформацией расстояния Хемминга и λ-расстояния.

Общая идея исследования состоит в том, чтобы взять некоторое множество языков (с одной стороны, достаточно представительное, а, с другой – для него имеется общепринятое дерево эволюции (эталонное)), построить для него деревья с помощью различных алгоритмов и метрик и, сравнивая их с эталонным, выявить наиболее перспективные методы. В дальнейшем эти методы могут быть применены в менее ясных ситуациях – с неустановленным родством языков.

1. Исследуемые языки

В работе изучается группа из 42 языков, принадлежащих различным языковым семьям: индоевропейской, северокавказской, чукотско-корякской, уральской, алтайской. Один язык (бирманский) принадлежит сино-тибетской семье. Нивхский язык является изолятом – не относится ни к одной семье. Языки выбирались из тех соображений, чтобы они были полно описаны и представляли все основные семьи, содержащиеся в базе данных. Кроме того, описания этих языков были подвергнуты дополнительной экспертизе.

Как принято в исторической лингвистике, семьи делятся на ветви, далее по мере дробления выделяют группы и подгруппы языков. Эволюционное дерево показано на рис. 1. Здесь использована классификация, приведенная в [6] и являющаяся практически общепринятой.

2. Алгоритмы и метрики

Тестировались два наиболее популярных филогенетических алгоритма – upgma и nj, и две метрики – по Хеммингу и λ-расстояние. Стоит отметить, что в работе [5] введено два различных определения λ-расстояния. В первом случае λ-расстояние определено как не метрическое (в нем не выполняется неравенство треугольника, то есть это псевдометрика), во втором случае λ-расстояние строится на основе первого с применением специального преобразования, обеспечивающего выполнение всех аксиом метрики. В дальнейшем в статье будет идти речь только о неметрическое λ-расстоянии, которое мы назовем λ-псевдометрикой (деревья, построенные с ее использованием, будут иметь маркер L), так как ее применение оказалось более эффективным, в том смысле, что позволяет строить деревья, более близкие к эталонному.

Определим новые псевдометрики, получаемые с помощью преобразования метрики Хемминга и λ-псевдометрики. Пусть $D(i,j)$ – расстояние между объектами i и j (по Хеммингу, либо λ-псевдометрике, такие расстояния в дальнейшем будут называться обычными, матрицы расстояний – обычными матрицами расстояний, а деревья, построенные при их использовании, – обычными деревьями).

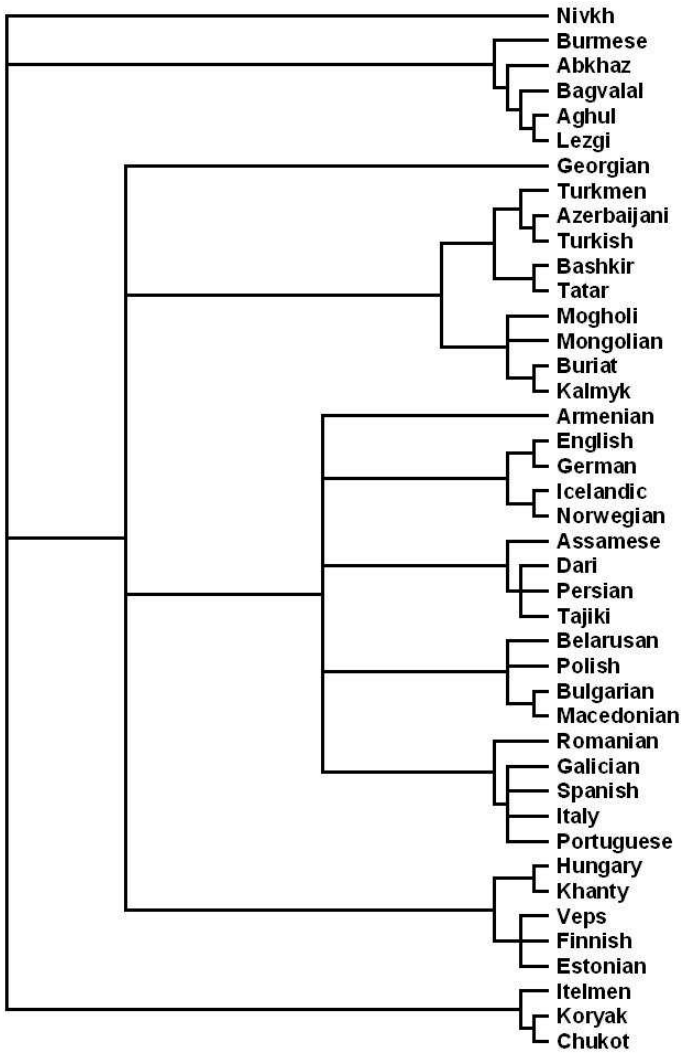

Рис. 1. Общепринятая генеалогическая классификация рассматриваемых языков

Определение 1. Для каждого объекта i найдем максимально удаленный от него объект. Пусть это будет объект k. Далее расстояние $D'(i,j)$ определим как $D(i,j)/D(i,k)$. Чтобы матрица расстояний была симметричной, элементы новой матрицы (назовем ее матрицей по максимуму) будем вычислять по формуле $M(i,j) = (D'(i,j) + D/(j,i))/2$. Полученная псевдометрика (в ней не выполняется аксиома треугольника) будет называться псевдометрикой по максимуму.

Определение 2. Для каждого объекта i найдем среднее расстояние до всех остальных объектов, пусть это будет число $s(i)$. Расстояние $D'(i,j)$ определим как $D(i,j)/s(i)$. Чтобы матрица расстояний была симметричной, элементы новой матрицы (матрицы по среднему) будем вычислять по формуле $M(i,j) = (D'(i,j) + D'(j,i))/2$. Псевдометрика, рассчитанная таким образом (в ней также не выполняется неравенство треугольника), будет называться псевдометрикой по среднему.

В дальнейшем метод с применением псевдометрики по максимуму, а также полученное по нему дерево будет иметь маркер m, а псевдометрики по среднему – s.

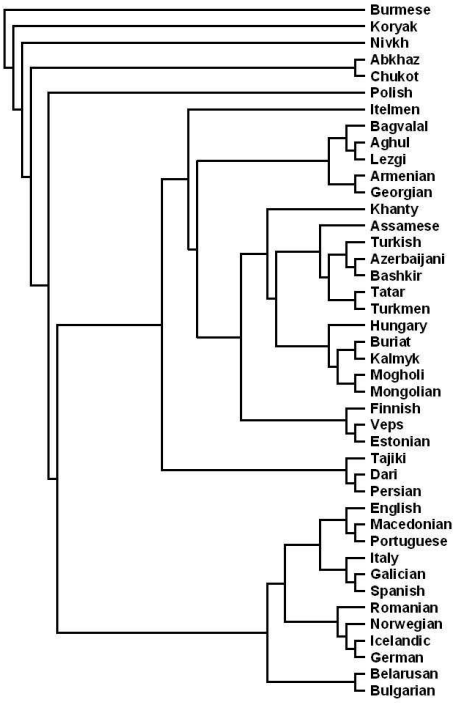

Рис. 2. Филогенетическое дерево, построенное по методу upgma с метрикой Хемминга

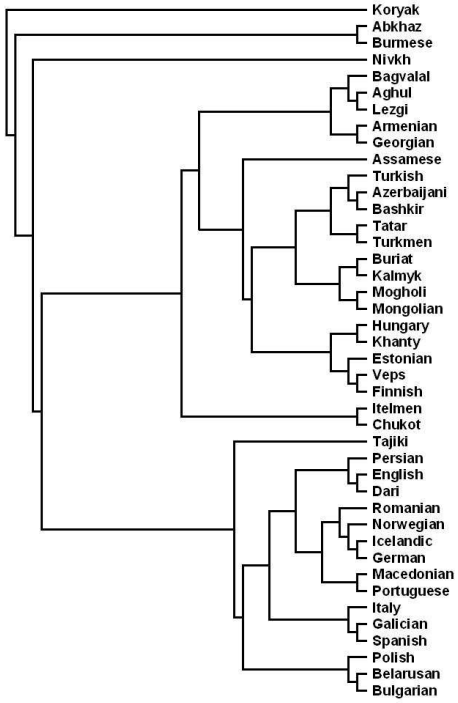

Рис. 3. Филогенетическое дерево, построенное по методу upgma с m-псевдометрикой Хемминга

Целью исследования является сравнение деревьев, построенных по обычным метрикам и псевдометрикам. Деревья сравнивались как на основе правильной классификации языков по семьям и ветвям, так и путем применения двух метрик на деревьях. Это следующие метрики:

1) метрика Робинсона – Фоулдса [7];

2) метрика, основанная на количестве квартетов, имеющих одинаковую топологию у двух деревьев – эталонного и исследуемого [8].

3. Филогенетические деревья по разным метрикам

Для каждого метода (upgma и nj) были построены деревья по двум расстояниям (Хемминга и λ-псевдометрики), а также путем преобразования этих расстояний двумя вышеописанными способами (m- и s-преобразования). Таким образом, всего было построено 12 деревьев. Рассмотрим результаты отдельно по каждому методу.

3.1. Метод upgma, метрика Хемминга. В дереве, построенном с использованием метрики Хемминга (рис. 2), уральская семья разделена на три части (в одной оказались финский, вепсский и эстонский языки, в двух других – хантыйский и венгерский соответственно). При применении m- и s-псевдометрик уральская семья была классифицирована в отдельное поддерево. Метод с метрикой Хемминга выделил в одно поддерево с алтайскими языками два языка из других семей: монгольская ветвь объединяется с венгерским языком, а тюркская – с ассамским. При использовании s-псевдометрики получилась аналогичная картина, но вместо

Табл. 1

Количество языков, верно классифицированных методом upgma по разным матрицам расстояний для семей

Семья	Общее число языков	upgma	upgma-m	upgma-s
Уральская	5	3	5	5
Алтайская	9	9 (с ассамским и венгерским)	9	9 (с ассамским и ительменским)
Индоевропейская	18	12	16	16
Чукотско-камчатская	3	–	2	–
Северокавказская	3	3	3	3

Табл. 2

Количество языков, верно классифицированных методом upgma по разным матрицам расстояний для ветвей

Ветвь	Общее число языков	upgma	upgma-m	upgma-s
Тюркская	5	5	5	5
Монгольская	4	4	4	4
Иранская	3	3	2 (с английским)	3
Германская	4	3	3	3
Романская	5	4 (с македонским и английским)	3	3
Славянская	4	2	3	3

венгерского языка монгольская ветвь объединена в общее поддерево с ительменским языком. В m-дереве (рис. 3) алтайская семья классифицируется верно.

В целом наилучшие результаты получились с применением m-псевдометрики. Однако в таком дереве лишь два из трех иранских языка оказались родственными. Дари и персидский языки образуют поддерево с английским языком, а третий (таджикский язык) расположен отдельно. В двух других деревьях иранские языки классифицированы верно. Дерево, построенное по метрике Хемминга, было наименее точное. Это демонстрируют табл. 1 и 2, в которых приведены результаты по всем трем способам отдельно по семьям и ветвям.

Сравнение деревьев с эталонным путем использования метрик на деревьях дало следующие результаты:

– расстояние Робинсона–Фоулдса между деревом по метрике Хемминга и эталонным получилось равным 53, по обоим псевдометрикам – 49;

– число одинаковых квартетов с эталонным деревом для дерева, построенного по метрике Хемминга, – 63713, m-псевдометрике – 70514, s-псевдометрике – 69506. Отметим, что в этом методе чем больше одинаковых квартетов, тем деревья ближе.

Таким образом, результаты при сравнении деревьев по точности классификации языковых семей и ветвей и значениям метрик на деревьях показали, что наиболее близким к истинному дереву языков получилось дерево, построенное по m-псев-

Табл. 3

Количество языков, верно классифицированных методом upgma-L по разным матрицам расстояний для семей

Семья	Общее число языков	upgma-L	upgma-L-m	upgma-L-s
Уральская	5	4	5	5
Алтайская	9	9 (с агульским и лезгинским)	9	9
Индоевропейская	18	13	13	13
Чукотско-камчатская	3	2 (с абхазским)	2 (с абхазским)	2 (с абхазским, бирманским и нивхским)
Северокавказская	3	2	3	3

Табл. 4

Количество языков, верно классифицированных методом upgma-L по разным матрицам расстояний для ветвей

Ветвь	Общее число языков	upgma-L	upgma-L-m	upgma-L-s
Тюркская	5	5	5	5
Монгольская	4	4	4	4
Иранская	3	3	3	3
Германская	4	3	3	3
Романская	5	3	3	3
Славянская	4	3	3	3

дометрике. В целом любое из двух деревьев, построенных по псевдометрикам, было более точно реконструировано, чем дерево, построенное по метрике Хемминга.

3.2. Метод upgma, λ-псевдометрика. В данном случае результаты по m- и s-псевдометрикам вновь оказались лучше (в сравнении с обычной λ-псевдометрикой). Алгоритм upgma с обычной λ-псевдометрикой объединяет в поддерево 4 из 5 уральских языков (отдельно стоял хантыйский язык вместе с армянским и грузинским). Вместе с алтайскими языками в одном поддереве оказались два северокавказских языка – агульский и лезгинский. m- и s-псевдометрики данные семьи выделили верно. Метод с применением λ-псевдометрики объединил лишь два из трех северокавказских языков (багвалинский язык оказался в одном поддереве с ассамским, а агульский и лезгинский – с монгольскими языками).

Расстояние Робинсона–Фоулдса между эталонным деревом и деревом, полученным с применением λ-псевдометрики, равно 55, число одинаковых квартетов – 63392. Для m- и s-деревьев расстояние Робинсона–Фоулда равно 49; число одинаковых квартетов – 68446 и 68797 соответственно. Таким образом, преобразованные m- и s-псевдометрики при применении алгоритма upgma дают примерно одинаковые результаты.

Полученные деревья, построенные по методу upgma с разными матрицами расстояний, свидетельствуют об эффективности применения псевдометрик. Деревья по ним получились более точными (см. табл. 3 и 4), что и подтвердили метрики на деревьях.

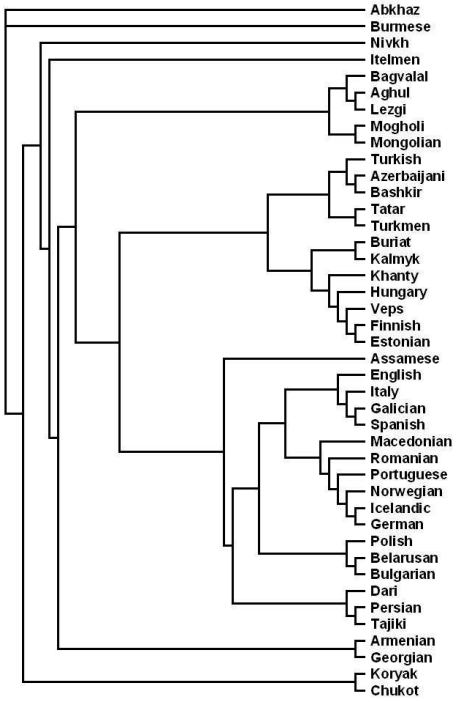

Рис. 4. Филогенетическое дерево, построенное по методу nj с метрикой Хемминга

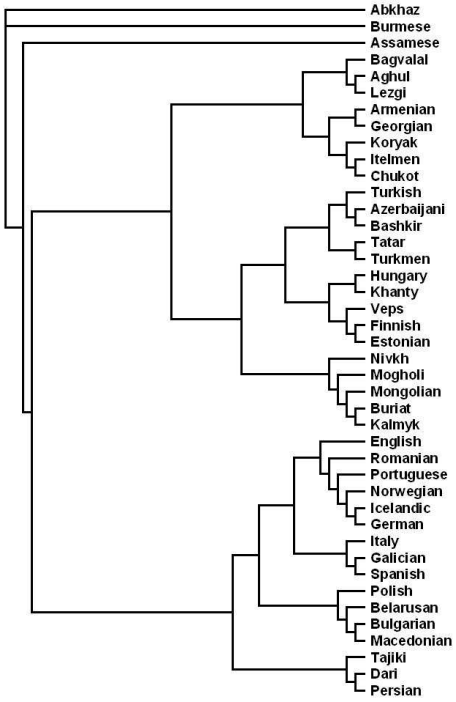

Рис. 5. Филогенетическое дерево, построенное по методу nj с s-псевдометрикой Хемминга

3.3. Метод nj, метрика Хемминга. В этом случае результаты с использованием метрики Хемминга (рис. 4) и m-псевдометрики были схожими, а дерево, построенное по s-матрице (рис. 5), было наиболее точным, что подтвердило как сравнение по правильности классификации языковых семей и ветвей (табл. 5, 6), так и метрики на деревьях.

Стоит отметить, что в методе nj с преобразованной метрикой Хемминга по среднему все четыре славянских языка были объединены в одно поддерево. В двух других деревьях был «потерян» македонский язык. В дереве, построенном по s-псевдометрике, все три чукотско-камчатских языка также образовали поддерево, что было только в данном дереве среди всех 12 построенных.

Метрики на деревьях показали, что, как и раньше, m- и s-матрицы расстояний дают наиболее точные результаты. Расстояние Робинсона–Фоулдса между эталонным и деревом с метрикой Хемминга получилось равным 50, число одинаковых квартетов – 72943. Для m-дерева эти показатели равняются 48 и 73412 соответственно. Для s-дерева расстояние Робинсона–Фоулдса равняется 42 (абсолютно лучший результат по данной метрике среди всех 12 деревьев), число одинаковых кластеров – 73253. Результаты двух метрик на деревьях для псевдометрик получились противоположными. Если по метрике Робинсона–Фоулдса наилучшим деревом считается s-дерево, то по количеству одинаковых кластеров – m-дерево (правда, разница здесь очень мала).

3.4. Метод nj, λ-псевдометрика. В данном случае дерево, построенное с применением λ-псевдометрики (рис. 6), было одним из самых точных среди

Табл. 5

Количество языков, верно классифицированных методом nj по разным матрицам расстояний для семей

Семья	Общее число языков	nj	nj-m	nj-s
Уральская	5	5	5	5
Алтайская	9	3 отдельные группы	3 отдельные группы	2 отдельные группы
Индоевропейская	18	17	17	16
Чукотско-камчатская	3	2	2	3
Северокавказская	3	3	3	3

Табл. 6

Количество языков, верно классифицированных методом nj по разным матрицам расстояний для ветвей

Ветвь	Общее число языков	nj	nj-m	nj-s
Тюркская	5	5	5	5
Монгольская	4	2	2	4
Иранская	3	3	3	3
Германская	4	3	3	3
Романская	5	3	4 (с английским)	3
Славянская	4	3	3	4

всех 12. m-дерево было схожим с обычным по топологии, а по количеству одинаковых квартетов с эталонным деревом немного лучше. Дерево, построенное по s-псевдометрике (рис. 7), было худшим (что подтвердили обе метрики на деревьях, а также классификация языков по семьям и ветвям (табл. 7,8)).

Расстояние Робинсона – Фоулдса между эталонным деревом и построенным с λ-псевдометрикой получилось равным 46, число одинаковых квартетов – 76398, для m-дерева 46 и 77803 соответственно, для s-дерева – 50 и 70601. Таким образом, это единственный случай из всех рассмотренных, когда применение псевдометрики привело к худшему результату.

В заключение приведем данные (табл. 9), полученные при применении к псевдометрикам алгоритма метризации, описанного в [5]. Практически во всех случаях использование метрик вместо псевдометрик приводит к ухудшению результатов.

Заключение

При использовании введенных m- и s-псевдометрик в трех из четырех случаях (методы upgma, upgma-L, nj) получались более точно реконструированные филогенетические деревья. Это подтвердили как содержательная лингвистическая классификация языков по семьям и ветвям, так и формальные метрики на деревьях. Лишь в методе nj-L использование s-псевдометрики привело к ухудшению результата.

По метрике Робинсона – Фоулдса наилучшее дерево получается при использовании преобразованной по среднему метрики Хемминга для метода nj. Далее следуют

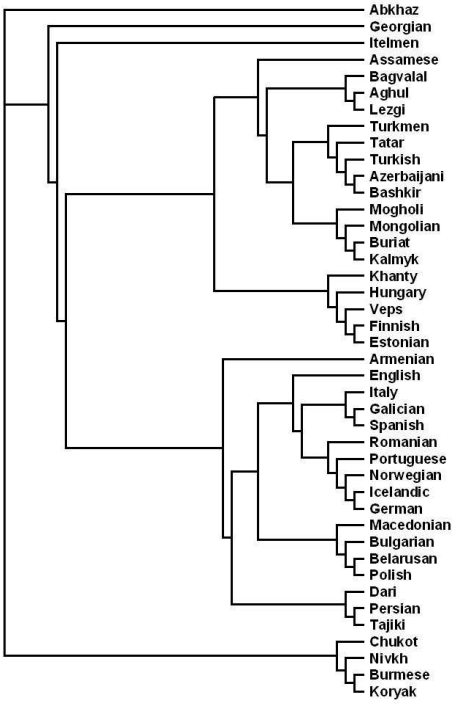

Рис. 6. Филогенетическое дерево, построенное по методу nj с λ-псевдометрикой

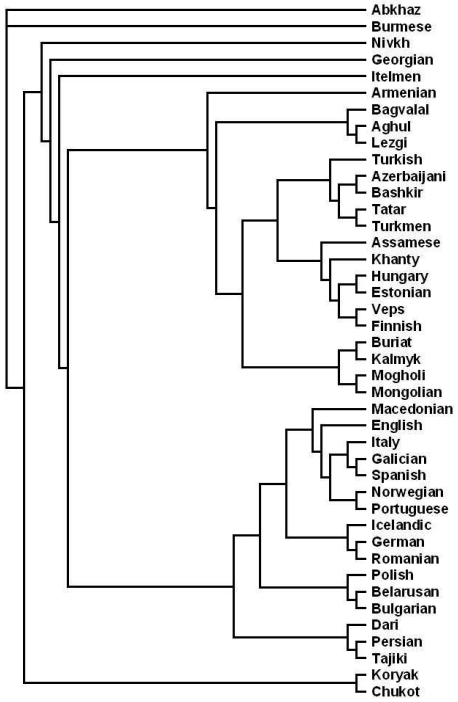

Рис. 7. Филогенетическое дерево, построенное по методу nj с λ-псевдометрикой, преобразованной по среднему

Табл. 7

Количество языков, верно классифицированных методом nj-L по разным матрицам расстояний для семей

Семья	Общее число языков	nj-L	nj-L-m	nj-L-s
Уральская	5	5	5	5
Алтайская	9	9	9	9 (с уральскими и ассамским)
Индоевропейская	18	17	17 (с грузинским)	16
Чукотско-камчатская	3	2 (с нивхским и бирманским)	2 (с нивхским и бирманским)	2
Северокавказская	3	3	3	3

деревья, полученные методом nj-L (с λ-псевдометрикой) и m-псевдометрикой расстояний, эти же деревья содержали наибольшее число одинаковых квартетов с квартетами из истинного дерева.

В целом при применении m- и s-псевдометрик деревья получались более точными. Среднее значение расстояния Робинсона – Фоулдса у метрики Хемминга и λ-псевдометрики равно 51, а количество одинаковых квартетов – 69112,

Табл. 8

Количество языков, верно классифицированных методом nj-l по разным матрицам расстояний для ветвей

Ветвь	Общее число языков	nj-L	nj-L-m	nj-L-s
Тюркская	5	5	5	5
Монгольская	4	4	4	4
Иранская	3	3	3	3
Германская	4	3	3	2 (с румынским)
Романская	5	3	3	4 (с норвежским)
Славянская	4	4	4	3

Табл. 9

Объединенные результаты по метрикам на деревьях

Метрика	Расстояние Робинсона – Фоулдса		Количество одинаковых кластеров	
	Без метризации	С метризацией	Без метризации	С метризацией
upgma	53*		63713*	
upgma-m	49	55	70514	60471
upgma-s	49	57	69506	66136
upgma-L	55	51	63392	63179
upgma-L-m	49	57	68446	64087
upgma-L-s	49	57	68797	62247
nj	50*		72943*	
nj-m	48	60	73412	62635
nj-s	42	64	73253	60954
nj-L	46	54	76398	72945
nj-L-m	46	62	77803	51337
nj-L-s	50	62	70601	52874

* В этих случаях использовалась метрика Хемминга, то есть метризация не требовалась.

у m-псевдометрики эти показатели составляют 48 и 72544 соответственно, у s-псевдометрики – 47.5 и 70539.

Можно заметить также, что во всех метриках использование метода nj является более эффективным, чем использование метода upgma.

Полученные результаты указывают, что имеет смысл производить поиск новых расстояний, а также модифицировать имеющиеся, для получения более точных филогенетических деревьев.

Работа выполнена при финансовой поддержке ФАО РФ (проект № 2.2.1.1/6944 «Развитие Российского научно-образовательного центра по лингвистике им. И.А. Бодуэна де Куртенэ»).

Summary

V.D. Solovyev, R.F. Fashutdinov. Transformation of Metrics Used in Clusterization Methods for Building the Phylogenetic Language Trees.

As large typological databases appeared a few years ago, the problem of data mining (as clusterization of languages) arose. Usually phylogenetic algorithms based on Hamming-distance are used for these purposes. But it was found out in cluster analysis that some other metrics give better results. In the paper two new metrics are proposed and it is shown on a great

number of linguistic examples that phylogenetic algorithms based on these metrics give better results.

Key words: linguistic database, metrics, phylogenetic algorithms.

Литература

1. *Поляков В.Н., Соловьев В.Д.* Компьютерные модели и методы в типологии и компаративистике. – Казань: Казан. гос. ун-т, 2006. – 208 с.
2. *Haspelmath M., Dryer M.S., Gil D., Comrie B.* (*eds.*). The World Atlas of Language Structures. – Oxford: Oxford Univ. Press, 2005. – 712 p.
3. *Semple Ch., Steel M.* Phylogenetics. – Oxford: Oxford Univ. Press, 2003. – 239 p.
4. *Holder M., Lewis P.O.* Phylogeny Estimation: Traditional and Bayesian Approaches // Nature Rev. Genet. – 2003. – No 4. –P. 275–284.
5. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. – Новосибирск: Ин-т матем. СО РАН, 1999. – 270 с.
6. *Бурлак С.А., Старостин С.А.* Введение в лингвистическую компаративистику. – М.: Эдиториал УРСС, 2001. – 272 с.
7. *Pattengale N.D., Gottlieb E.J., Moret B.M.E.* Efficiently Computing the Robinson-Foulds Metric // J. Comput. Biol. – 2007. – V. 14, No 6. – P. 724–735.
8. *Estabrook G.F., McMorris F.R., Meacham C.A.* Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units // System. Zool. – 1985. – V. 34, No 2. – P. 193–200.

Поступила в редакцию
12.05.09

Соловьев Валерий Дмитриевич – доктор физико-математических наук, профессор кафедры теоретической кибернетики Казанского государственного университета.
E-mail: *maki.solovyev@mail.ru*

Фасхутдинов Ренат Фархутдинович – аспирант Института проблем информатики АН Республики Татарстан, г. Казань.
E-mail: *jvenal@mail.ru*