Научная статья
УДК 81.32
EDN PTMYCH
DOI 10.17150/2713-1734.2023.5(3).303-318

А.В. Боровский
Байкальский государственный университет, г. Иркутск, Российская Федерация
В.В. Братищенко
Байкальский государственный университет, г. Иркутск, Российская Федерация
Е.Е. Раковская
Байкальский государственный университет, г. Иркутск, Российская Федерация

Лексикостатистика взаимосвязи русского и бурятского языков

Аннотация. Представлены результаты статистического анализа данных матрицы мер близости слов, составленных из консонантных классов на основе частотного русско-бурятского словаря с применением меры сходства Рэтклиффа-Обершелпа. Рассчитано число полных совпадений слов по всему полю матрицы методом независимых вероятностей появления букв и классов в словах, и методом расчета вероятностей с применением цепей Маркова. Наилучший результат получен для расчета вероятностей в предположении марковской зависимости классов.

Исследовано влияние ранга матрицы мер близости слов на степень совпадения списков. Средние меры сходства мало изменяются при уменьшении ранга матрицы до величины r = 64. Дальнейшее уменьшение ранга матрицы является нецелесообразным. Определены количественные характеристики языков - вероятности появления консонантных классов для слов словаря, вероятности слов разной длины в русском и бурятском языках. Для проверки нулевой гипотезы (наблюдаемые сходства метрик случайны) использован перестановочный тест, после проведения которого сделан вывод о неслучайном характере совпадения слов в русском и бурятском языках и о наличии заимствований в языках.

Ключевые слова. Лексикостатистика русского и бурятского языков, математическая и историческая лингвистика, консонантные классы, матрица мер сходства слов русского и бурятского языков, марковские цепи в лексикостатистике.

Информация о статье. Дата поступления: 4 апреля 2023 г.; дата принятия к публикации: 19 июня 2023 г.; дата онлайн-размещения: 28 сентября 2023 г.

Original article

A.V. Borovsky
Baikal State University, Irkutsk, Russian Federation
V.V. Bratishchenko
Baikal State University, Irkutsk, Russian Federation

ISSN 2713-1734

E.E. Rakovskaya
Baikal State University,
Irkutsk, Russian Federation

Lexicostatistics of the Relationship Between the Russian and Buryat Languages

Abstract. The article presents the results of the statistical analysis of indicators of the matrix of word meanings, composed of consonant classes based on the frequencies of the Russian-Buryat dictionary using the Ratcliffe-Obershelp similarity measures. We calculated a number of complete matches of words over the entire field of the matrix by the method of known probabilities for detecting letters and classes in words, and by calculating the probabilities using Markov chains. The best result was obtained for calculating probabilities under the assumption of Markovian dependence of the classes.

The influence of the rank of the matrix of measures of similarity of words on the degree of coincidence of lists has been studied. The average similarity measures change little as the rank of the matrix decreases to r = 64. Further reduction of the rank of the matrix is inexpedient. The quantitative characteristics of languages are determined - the probabilities of the appearance of consonant classes for dictionary words, the probabilities of words of different lengths in the Russian and Buryat languages.

To test the null hypothesis (the observed similarities of the metrics are random), a permutation test was used, after which a conclusion was made about the non-random nature of the coincidence of words in the Russian and Buryat languages and the presence of borrowings in the languages.

Keywords. Lexicostatistics of the Russian and Buryat languages, mathematical and historical linguistics, consonant classes, matrix of similarity measures for Russian and Buryat words, Markov chains in lexicostatistics.

Article info. Received 4 April, 2023; Accepted 19 June, 2023; Available online 28 September, 2023.

Введение

Вопросы генетического анализа возникают в биологии, лингвистике, в некоторых сложных комбинаторных задачах [1].

Ранее авторы предлагаемой статьи занимались изучением топонимики Иркутской области. Оказалось, что значительная часть бурятско-эвенкийской топонимики являются искажениями слов санскрита. В процентном отношении количество не русских топонимов, коррелирующих с санскритом составило 63–65% [2]. По мнению авторов, санскрит является старым языком славянских племён, что подтверждается ареалом распространения некоторых характерных старых санскрит-топонимов на территории России.

Для более глубокого изучения установленной закономерности авторы провели исследование связи бурятского и русского языков методом матрицы мер близости русских и бурятских слов из двух языков [3]. Исследование показало степень заимствования русского языка в бурятском на уровне 83 %. Такое заимствование шло через обучение бурятского населения санскриту буддийскими ламами, начиная с конца XVI в., старорусскому языку казачеством

и русскими переселенцами с XVII в., и современному русскому языку с XIX в.

В настоящей статье, которая является продолжением исследования [3], авторы обращают своё внимание на математические вопросы статистической устойчивости метода матрицы мер близости слов двух языков при количестве метрик 1000 x 1000 ≧ 10^6. Авторы исследуют вопросы, насколько результаты расчёта средних величин чувствительны к изменению ранга матрицы мер. Зависят они или нет от случайного формирования матрицы мер. Могут ли влиять на результаты расчётов случайные совпадения слов в двух языках. Ответы на эти вопросы лучше обосновывают возможность применения метода матрицы мер для изучения объёмов заимствования при сравнении двух языков.

Методы, позволяющие установить сходство языков

Все языки меняются во времени. Самым распространенным приемом, позволяющим установить сходство и генетическое родство языков, является применение методов на основе лексикостатистики, использующих лексические признаки языков, которые более предпочтительны, т.к. могут быть характерны для большого числа языков, и не требуется выбор специальных признаков для конкретных языков.

Все слова языка можно представить в виде последовательностей звуков и обнаружение родственных слов определяется при сравнении таких последовательностей. Для выявления звуковых соответствий последовательностей звуков в словах разработаны многочисленные меры подобия, основанные на форме сравниваемых строк, без учета семантики. По метрике «расстояния» Левенштейна рассчитывается минимальное количество удалений, вставок или замен, необходимых для преобразования одной строки в другую. Было предложено множество вариантов улучшения расстояния Левенштейна, позволяющих учитывать длину слова или скорректированные веса для различных операций редактирования [4].

Распознавание образов Рэтклиффа-Обершелпа [5] определяет сходство двух строк в виде отношения удвоенного количества совпадающих символов к общему количеству символов в обеих строках.

Распространенным подходом для оценки сходства строк является преобразование последовательностей звуков в векторные представления, над которыми в дальнейшем выполняются вычисления сходства [6].

Для установления степени совпадения списков слов с учетом фонетических изменений в процессе эволюции языков применяется метод консонантных классов А.Р. Долгопольского [7]. В этом методе фонетический алфавит разделен на несколько непересе-

кающихся подмножеств (классов). Основная идея использования консонантных классов состоит в том, что в процессе эволюции взаимные переходы для звуков одного класса более вероятны, чем для звуков разных классов. Каждое слово кодируется в соответствии с составляющими его классами согласных. Распространенным представлением с применением классов согласных является кодировка в виде двухконсонантной формы, т.е. для анализа используются первые два согласных звука в слове, остальные игнорируются.

Преимуществом представления в виде классов является учет специфических модификаций звуков в исторической перспективе.

Важным этапом после определения доли совпадений или расчета количественных мер сходства между двумя списками разных языков является оценка статистической значимости результата. Для этого используется перестановочный тест, основанный на многократной генерации выборок методом Монте-Карло на базе имеющихся списков [8].

Случайным образом выбирается слово из первого списка и сопоставляется со случайно выбранным словом из второго списка. Таким образом генерируется случайная парная выборка. Повторяя этот шаг многократно, получается набор случайных парных выборок, по которому можно оценить вероятность нулевой гипотезы (любые совпадения в списках слов являются случайными). Появляется возможность рассчитать среднее количество случайных совпадений в списках, вероятность совпадений слов в сгенерированных выборках меньшую, большую или равную, чем в исходной выборке (P-значение). Наиболее часто принимается уровень статистической значимости 5 %. Это означает, что нулевая гипотеза должна быть отвергнута, если P-значение меньше 0,05.

В традиционном, перестановочном тесте два списка слов сравниваются между собой, и определяется количество совпадений в списках или количественная мера сходства для каждого слова списка. Далее один из списков перемешивается случайным образом и определяется количество совпадений в списке для каждой новой конфигурации.

Для определения генетического родства языков применяется также взвешенный перестановочный тест [9] для списков Сводеша, где каждому понятию списка присваивается весовой коэффициент в соответствии с его устойчивостью. Общепризнано, что понятия Сводеша обладают разной средней степенью устойчивости, некоторые понятия дольше сохраняются в языке, другие понятия менее устойчивы и соответствующие им слова чаще меняют свое значение в ходе языковой эволюции. Для расчета весовых коэффициентов применяется индекс стабильности элемента Сводеша, равный M/L, где L – количество языков в языковой семье, M – мак-

симальное количество языков внутри языковой семьи, использующий один и тот же корень для соответствующего значения Сводеша. На следующем этапе для каждого концепта Сводеша берется среднее арифметическое его индексов устойчивости в отдельных языковых семьях. Полученное значение является весовым коэффициентом для слова в списке Сводеша.

1. Основная часть

1.1. Статистические характеристики бурятского и русского языков

Новым подходом к изучению взаимного влияния русского и бурятского языков и выявлению закономерностей для каждого языка является определение статистических характеристик – среднего по всей матрице мер, среднего от максимальных мер по строкам матрицы, среднего от максимальных мер по столбцам матрицы, расчёт дисперсии при вычислении средних [3].

Для проведения исследования был составлен рабочий частотный словарь русских слов размером 1025 слов на основе словаря «Новый частотный словарь русской лексики» [10]. Все слова рабочего словаря были преобразованы в кодированную форму с применением консонантных классов А.Р. Долгопольского. В табл. 1 приведены классы согласных букв для русского и бурятского языков.

Таблица 1

Классы согласных для русского и бурятского языков

	Класс согласных	Буквы русского и бурятского языка
1	P-класс	П, Б, В, Ф
2	T-класс	Т, Д
3	S-класс	С, З, Ц, Ч, Ш, Щ, Ч
4	M-класс	М
5	N-класс	Н
6	R-класс	Р, Л
7	K-класс	К, Г, Х, h
8	H-класс (нулевой класс)	Все гласные, включая Е, Ё, Ю, Я, Ѳ, Ү

Для сравнения языков на основе рабочего словаря были рассчитаны количественные характеристики русского и бурятского языка, включающие в себя вероятности появления консонантных классов для слов словаря (рис.1), вероятности появления слов из 2-х, 3-х, 4-х … букв или классов (рис. 2, 3).

Рисунки 1-3 позволяют увидеть характерные отличия между русским и бурятским языками:

1. В русском языке преобладают консонантные классы P, S, T. Частотность появления классов M и N одинаковая с бурятским

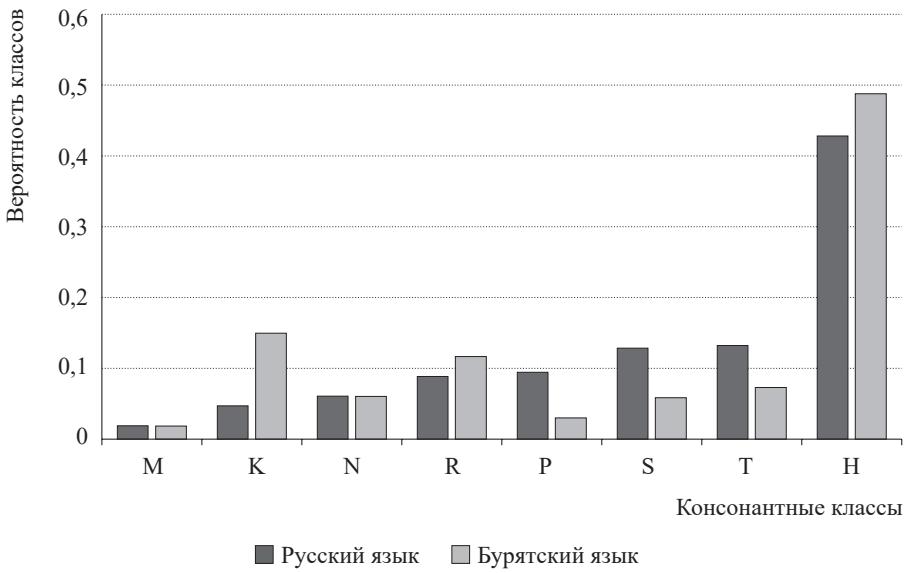

Рис. 1. Вероятности консонантных классов для слов рабочего словаря

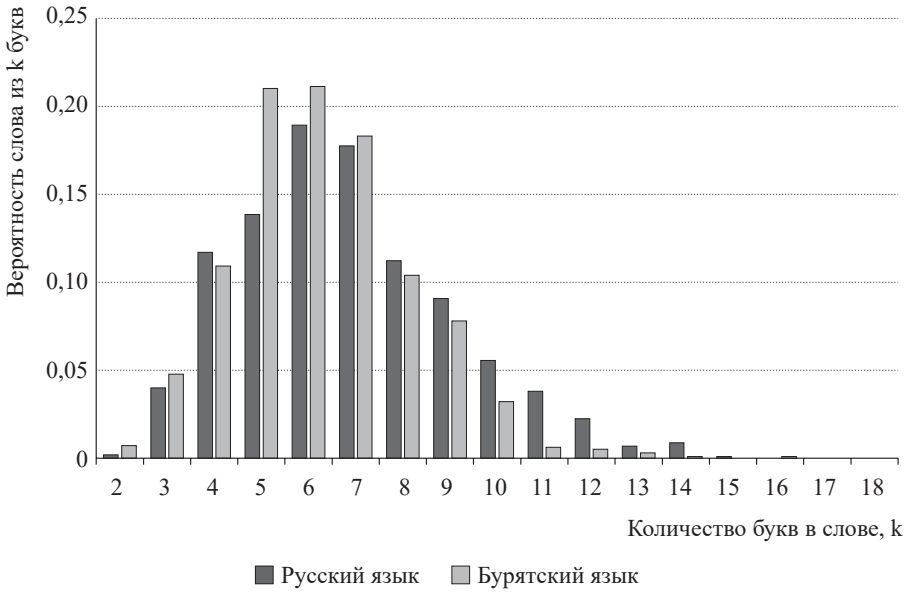

Рис. 2. Вероятности появления слов из 2-х, 3-х, 4-х … букв

языком, частотность появления классов K и R меньше, чем в бурятском языке. Частотность появления гласных звуков меньше, чем в бурятском.

2. В русском языке имеется больше многобуквенных слов по сравнению с бурятским языком, начиная с 7 букв в слове.

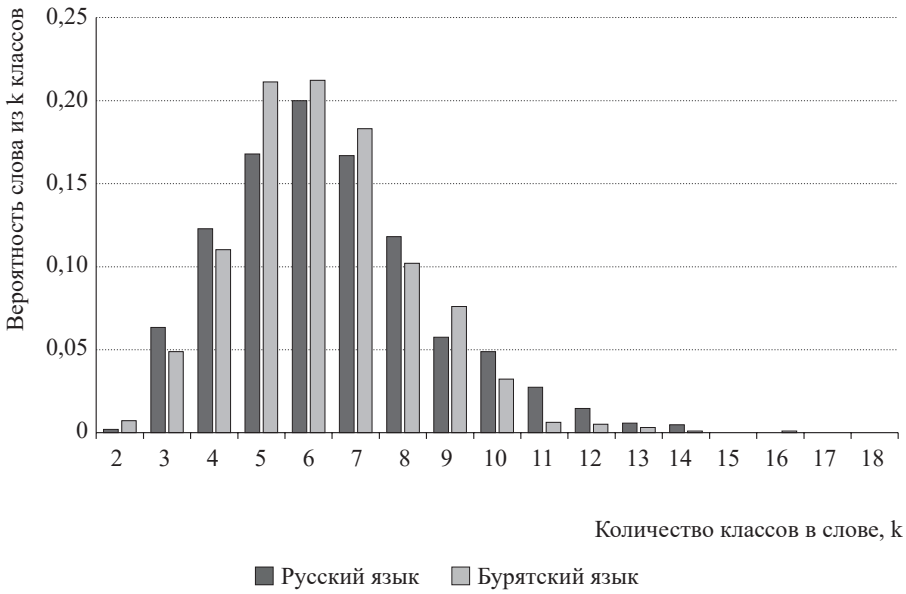

Рис. 3. Вероятности появления слов из 2-х, 3-х, 4-х … классов

3. В русском языке больше слов из четырёх букв.
4. В бурятском языке преобладают слова из 2-3 и 5-7 букв.

1.2. Метод независимых вероятностей появления букв и классов в словах

Для расчета независимых вероятностей появления букв или классов в словах словаря берем строку с номером *i*. Пусть в начале строки стоит слово $n_i = A_1A_2 \ldots A_{k(i)}$ на языке «а». Здесь *k(i)* – число букв (классов) в слове, а $A_1A_2 \ldots A_{k(i)}$ — конкретные буквы или классы для слова n_i. Тогда вероятность появления аналогичного слова в словаре на языке «б» рассчитывается по формуле

$$P_i^{(1)} = P(ni) = P(A_1)P(A_2) \ldots P(A_{k(i)}) \qquad (1)$$

Здесь $P(A_1) \ldots P(A_{k(i)})$ — вероятности появления букв $A_1A_2 \ldots A_{k(i)}$ при формирования слова *ni*. Для расчета вероятностей $P(A_1) \ldots P(A_{k(i)})$ применяется соотношение $P\left(A_{k(i)}\right) = \frac{N_{k(i)}}{N}$, где $N_{k(i)}$–количество букв (классов) *k(i)* в словаре, N – общее количество букв (классов) в словаре.

1.3. Метод цепей Маркова

Предполагается, что последовательность классов генерируется конечной цепью Маркова. В цепи выделяется начальное состояние (маркируется как «0») и конечное состояние (маркируется как «1»), остальные состояния соответствуют классам букв. Пе-

реход из состояния в состояние зависит только от предыдущего состояния (марковское свойство). Вероятность последовательности классов $A_1A_2 \ldots A_k$ слова будет вычисляться как произведение вероятностей переходов:

$$P_i^{(2)} = P\{A_1A_2 \ldots A_k\} = \prod_{i=1}^{k+1} P\{A_i|A_{i-1}\}, \qquad (2)$$

где $P\{A_i|A_{i-1}\}$ вероятность перехода в A_i из состояния A_{i+1}, A_0 = «0» начальное состояние, A_{k+1} = «1» конечное состояние.

1.4. Метод строгих условных вероятностей

Для расчета строгих условных вероятностей появления букв или классов в словах словаря берем строку с номером *i*. В начале строки стоит слово *ni*. Здесь $A_1A_2 \ldots A_{k(i)}$ — конкретные буквы или классы для слова *ni*. Тогда вероятность появления слова в словаре рассчитывается по формуле:

$$P_i^{(3)} = \prod_{i=1}^{k(i)+1} P(A_i|A_1A_2 \ldots A_{i-1}). \qquad (3)$$

Здесь $P(A_i|A_1A_2 \ldots A_{i-1})$ — вероятность появления буквы A_i, при условии, что перед ней уже стоят буквы $A_1A_2 \ldots A_{i-1}$. Аналогично (2) $P(A_1|A_0)$ — условная вероятность появления буквы A_1 в начале слова. $P(A_{k+1}|A_1A_2 \ldots A_k)$ — вероятность наступления конца слова.

В статье рассматривался русско-бурятский частотный словарь. Если расположить список русских слов по вертикали, а список бурятских слов по горизонтали и раскрыть (3) для каждой строки лингвистической матрицы, подсчитывая условные вероятности для бурятского списка слов, то получим число единиц в каждой строке, делённое на число столбцов матрицы. Суммируя по всем строкам, получим число единиц на поле матрицы, делённое на число столбцов (слов в словаре). Таким образом (3) даёт вероятность полных звуковых совпадений слов (или классов звуков) в словаре. Отметим, что расчёт с использованием полных условных вероятностей совпадает с расчётом по методу матрицы мер близости слов (классов) в списках двух языков.

2. Статистическая проверка числа совпадения классов слов для частотного русско-бурятского словаря

В рассмотренном фрагменте русско-бурятского словаря по причинам многозначности перевода было выделено N = 960 уни-

кальных бурятских слов и $M = 1025$ уникальных русских слов. Попарное сравнение всех со всеми слов, записанных в виде соответствующих консонантных классов, выявило 149 полных совпадений. Для 960 слов бурятского словаря было выявлено 82 слова, классы которых совпали с классами ровно одного русского слова, 20 бурятских слов, классы каждого из которых совпали с классами двух русских слов, 9 бурятских слов, классы каждого из которых совпали с классами 3 русских слов. Предложенные модели позволяют оценить вероятности таких совпадений. Для этого можно использовать оценку $P_i^{(1)}$ вероятности классов *i*-го слова по формуле (1), в предположении независимости классов в слове, и оценку $P_i^{(2)}$ вероятности классов *i*-го слова по формуле (2), в предположении, что классы слова образуют марковскую последовательность.

Для каждого *i*-го слова бурятского словаря количество совпадений его классов с классами *M* русских слов можно рассматривать как проведение *M* независимых испытаний на совпадение с вероятностью $P_i^{(1)}$ или $P_i^{(2)}$. Количество $X_i^{(1)}$ или $X_i^{(2)}$ совпадений будет биномиальной случайной величиной с соответствующими параметрами. Для вычисления $P_i^{(1)}$ и $P_i^{(2)}$ применяются вероятностные характеристики (1) и (2) классов русского языка.

Хорошей аппроксимацией случайных величин $X_i^{(1)}$ и $X_i^{(2)}$ можно считать пуассоновское распределение с параметрами $\lambda_i^{(1)} = MP_i^{(1)}$ и $\lambda_i^{(2)} = MP_i^{(2)}$, так как выполняются условия такой аппроксимации: большое *M*, а также вероятности $P_i^{(1)}$ и $P_i^{(2)}$ близкие к нулю. Пуассоновское распределение является устойчивым, поэтому суммы $\sum_{i=1}^{N} P_i^{(1)}$ и $\sum_{i=1}^{N} P_i^{(2)}$ также будут описываться распределением Пуассона. У пуассоновской случайной величины математическое ожидание и дисперсии одинаковы и совпадают с параметром распределения. Для рассматриваемых моделей эти значения равны

$$L_1 = M\sum_{i=1}^{N} P_i^{(1)} = 64{,}71 \text{ и } L_2 = M\sum_{i=1}^{N} P_i^{(2)} = 147{,}65.$$

Модель (1) независимых вероятностей в последовательности классов демонстрирует значимые отклонения. Наблюдаемое количество совпадений 149 не попадает в доверительный интервал (38.24, 91.18) с доверительной вероятностью 0.9. Поэтому данная модель неправильно описывает частотные характеристики последовательности консонантных классов.

Отклонение наблюдаемого количества 149 совпадений классов незначительно отличается от математического ожидания L_2 и

попадает в доверительный интервал (127.66, 167.63) с доверительной вероятностью 0.9. Это свидетельствует о хорошей точности описания совпадения слов, записанных в виде консонантных классов, марковской моделью (2).

Аналогичные статистические выводы получены для частот совпадения консонантных классов русских слов словаря с вероятностными моделями следования классов в бурятском языке.

Таблица 2

Участок матрицы мер сходства для слов русско-бурятского частотного словаря, вычисленных на основе формулы Рэтклиффа-Обершелпа [5]

Слово:Класс	байха: PHHKH	жэл: SHR	шадаха: SHTHKH	бираха: PHRHKH
быть:PHT	0.5000	0.3333	0.4444	0.4444
год:KHT	0.5000	0.3333	0.4444	0.4444
мочь:MHS	0.2500	0.3333	0.2222	0.2222
человек:SHRHPHK	0.5000	0.6000	0.6154	0.6154
сказать:SKHSHT	0.3636	0.4444	0.5000	0.3333
один:HTHN	0.4444	0.2857	0.6000	0.4000
время:PRHMH	0.6000	0.2500	0.3636	0.7273
другой:TRHKHH	0.5455	0.2222	0.6667	0.6667
говорить:KHPHRHT	0.3333	0.4000	0.3077	0.6154
знать:SNHT	0.2222	0.5714	0.6000	0.2000
стать:STHT	0.2222	0.5714	0.6000	0.2000
дело:THRH	0.4444	0.5714	0.6000	0.6000
жизнь:SHSN	0.2222	0.5714	0.4000	0.2000
первый:PHRPHH	0.5455	0.4444	0.5000	0.8333
два:TPH	0.5000	0.3333	0.4444	0.4444
день:THN	0.2500	0.3333	0.4444	0.2222
новый:NHPHH	0.6000	0.2500	0.5455	0.5455
рука:RHKH	0.6667	0.2857	0.6000	0.8000
работа:RHPHTH	0.5455	0.2222	0.5000	0.5000

Проведённое исследование показывает, что метод независимых вероятностей появления букв в словах для лингвистических задач не работает. Буквы (звуки) в словах появляются с учётом предыстории. Расчёт вероятностей совпадения слов, составленных из классов, в русском и бурятском списках частотного словаря, выполненный на основе марковских цепей совпадает со строгим расчётом с точностью 0,9 %.

Вероятностные расчёты не позволяют отделить заимствования от случайных совпадений. Конечный результат учитывает и то и другое.

В табл. 3 приведены примеры вероятностей переходов для бурятского языка.

Таблица 3

Вероятности переходов для бурятского языка

Символ 1	Символ 2	Сочетания символов 1 и 2	Вероятности переходов
0	N	0N	0.0614583
0	S	0S	0.1427083
0	K	0K	0.2250000
0	M	0M	0.0281250
0	H	0H	0.2989583
0	P	0P	0.0989583
0	T	0T	0.1447917
N	1	N1	0.5509642
N	S	NS	0.0220386
N	K	NK	0.0688705
N	H	NH	0.3030303
N	T	NT	0.0550964
R	1	R1	0.1554922
R	R	RR	0.0099857
R	S	RS	0.0499287
R	K	RK	0.2667618
R	M	RM	0.0114123
R	H	RH	0.4065621
R	P	RP	0.0185449

3. Исследование влияния на степень совпадения списков ранга матрицы мер близости слов.

По метрике сходства Рэтклиффа-Обершелпа было определено сходство классов слов для частотного русского-бурятского словаря размером 1025 русских слов и 960 бурятских.

Среднее значение метрик по частотному словарю: $\langle P \rangle = \frac{1}{N}\sum_{i=1}^{N} P_i = 0.46601$.

Для исследования влияния ранга матрицы мер близости на степень совпадения списков, количество строк в матрице было уменьшено, и для расчета применялись каждая 2-я, 4-я, 8-я, 16-я, 32-я строки матрицы.

Получены результаты:

Относительная ошибка при расчете средних, кроме последнего, не превышала 1,34 %. Для $j = 32i$ (размер выборки уменьшили до 1025/32 ≈ 32 строк) относительная ошибка равна 4,58 %.

На рис. 4 и 5 показаны средние меры сходства слов и их относительные погрешности для матриц с j строками, где j = 2, 4, 8, 16, 32 i.

Данные расчёты показывают применимость метода матрицы мер близости слов с уменьшением ранга матрицы вплоть до $j = 16i$. В последнем случае ранг матрицы равен r = 64. При даль-

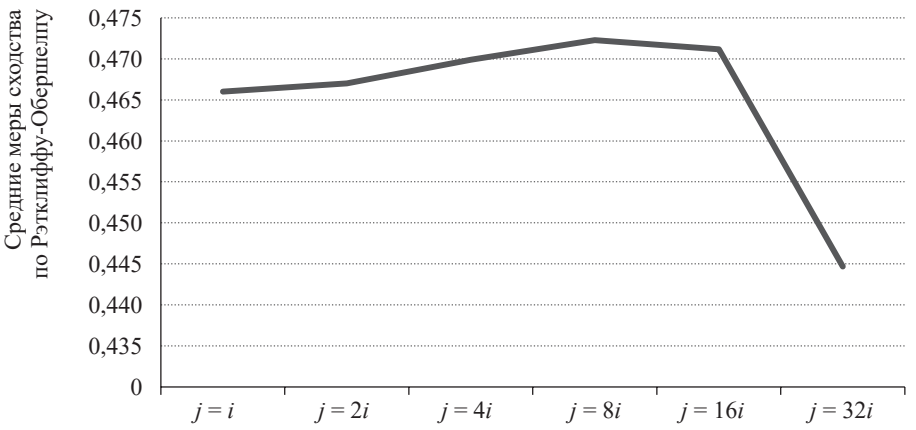

Рис.4. Средние меры сходства слов по Рэтклиффу-Обершелпу при понижении ранга матрицы

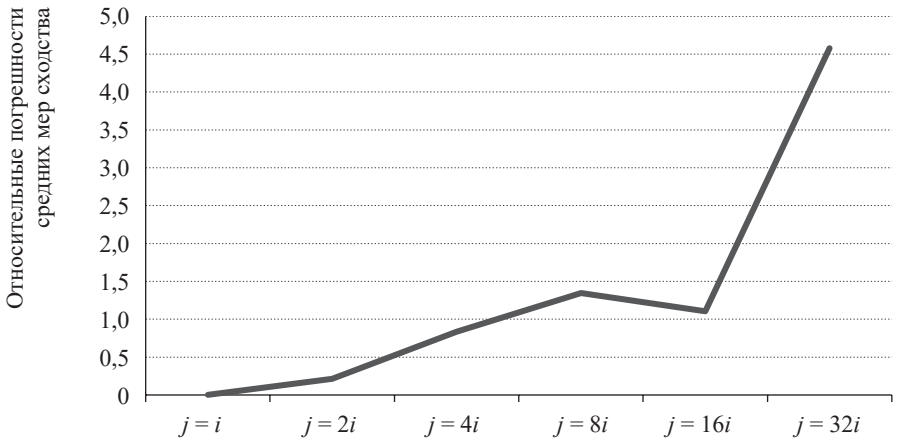

Рис.5. Относительные погрешности средних мер сходства по Рэтклиффу-Обершелпу при понижении ранга матрицы

нейшем уменьшении ранга матрицы результаты расчета средних величин становятся неприемлемыми.

4. Перестановочный тест

Для оценки выборки на репрезентативность и проверки нулевой гипотезы (наблюдаемые сходства метрик случайны) был применен перестановочный тест. При этом для каждого русского слова из словаря (1025 слов) случайным образом были подобраны слова из списка бурятского языка (960 слов) и рассчитаны метрики сходства Рэтклиффа-Обершелпа. Случайные выборки были сформированы 500 раз и для всех этих выборок были рассчитаны средние меры сходства, стандартные отклонения, количества слов, уровень сходства для которых превышает 0,5 (по Рэтклиффу-Обершелпу), средние меры сходства для слов, имеющих меру

сходства более 0,5. По всем показателям исходная выборка лучше, чем сгенерированные, а это значит, что сходство русского и бурятского языка не случайное. Из 500 сгенерированных выборок только три имели количество слов с мерой сходства > 0,5 большее, чем в исходном словаре. Р-значение равно 0,006.

Таблица 4

Некоторые результаты перестановочного теста

	Исходная выборка	Случайные выборки
Среднее ρ_{ik}	0.46826	0.45309
Количество слов, имеющих $\rho_{ik} > 0.5$	466	423
Среднее ρ_{ik} для слов, имеющих $\rho_{ik} > 0.5$	0.58630	0.57775
Среднее отклонение ρ_{ik} для слов, имеющих $\rho_{ik} > 0.5$	0.06852	0.06094

Заключение

В статье исследованы статистические особенности матрицы мер близости слов для частотного русско-бурятского словаря. Так как имеет место многозначность русских слов в бурятском языке, то список русских слов насчитывал 1025 наименований, а список бурятских слов 960 наименований.

Слова русского и бурятского списков были переведены в классы по А.Б.Долгопольскому, что позволяет учитывать изменчивость звуков в словах в исторической перспективе.

Меры близости для слов, составленных из классов, рассчитывались на основе метрики Рэтклиффа - Обершелпа [5].

Для матрицы мер близости рассчитывались средние меры, средние от максимальных по строкам и столбцам и средне-квадратичные отклонения (дисперсии).

Искусственное уменьшение ранга матрицы показало, что средние меры мало чувствительны к этому процессу и находятся в пределах 1,5 % отклонения вплоть до ранга $r = 64$. Матрицы с рангом $r < 64$ использоваться не должны.

На поле исходной матрицы обнаружены 149 единиц. Это говорит о наличии 149 полных совпадений русских и бурятских слов, составленных из классов А.Б. Долгопольского. Кроме полных совпадений есть, конечно, и частичные совпадения слов. В статье исследованы полные совпадения слов методами теории вероятностей. Рассмотрены три комбинаторных метода лексикостатистики:

– метод вычисления вероятностей совпадения слов на основе независимого появления букв (классов) в слове;

– метод вычисления вероятностей совпадения слов на основе марковских цепей;

– метод вычисления вероятностей совпадения слов на основе формул строгих условных вероятностей.

Вычисления в рамках третьего метода совпадают с вычислениями методом матрицы мер близости слов. Вычисления с использованием марковских цепей близки к результатам третьего метода (отклонение не превышает 1 %). Первый метод авторам пришлось забраковать. Он дает большие отклонения результатов, см. п. 1.3.

Кроме того, в данной статье авторы применили «перестановочный тест» к расчётам средних мер близости слов. На основе частотного русско-бурятского словаря сделаны 500 случайных выборок, что соответствует перемешиванию слов в списках. Для каждой выборки рассчитывались средние меры. В 94 % случаев средние меры уменьшались, по сравнению с исходным случаем. Перестановочный тест говорит о неслучайном характере совпадения слов в русско-бурятском словаре, т.е. о наличии заимствований.

С точки зрения авторов средние меры по полю матрицы устанавливают степень похожести языков, на которую влияют заимствования и случайные совпадения. Средние от максимальных мер по строкам (или столбцам) матрицы устанавливают в большей степени заимствование языков. Высказанные утверждения являются гипотезами. Математические методы, отделяющие заимствования от случайных совпадений нуждаются в дальнейшей разработке.

Список использованной литературы

1. Мартьянов В.И. Адаптация логико-эвристических методов для некоторых задач филогенетического анализа / В.И. Мартьянов, М.Л. Скуматов. — EDN OEFVVT // Современные технологии. Системный анализ. Моделирование. — 2011. — № 2. — С. 27–31.

2. Боровский А.В. Изучение корреляции между топонимами Иркутской области и словами санскрита / А.В. Боровский, Е.Е. Раковская, И.А. Картеева. — DOI 10.17150/2500-2759.2022.32(3).610-622. — EDN SUWJUC // Известия Байкальского государственного университета. — 2022. — Т. 32, № 3. — С. 610–622.

3. Боровский А.В. Изучение связи между русским и бурятским языками методом матрицы мер близости между консонантными классами слов / А.В. Боровский, В.В. Братищенко, Е.Е. Раковская. — DOI 10.17150/2713-1734.2023.5(1).19-33 // System Analysis&Mathematical Modeling. — 2023. — Т. 5, № 1. — С. 19–33.

4. Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals / V.I. Levenshtein // Soviet physics doklady. — 1966. — Vol. 10, No. 8. — P. 707–710.

5. Ratcliff J.W. Pattern-matching-the Gestalt Approach / J.W. Ratcliff, D.E. Metzener // Dr. Dobb`s Journal. — 1988. — Vol. 13, № 7. — P. 46.

6. Cohen W.W. A Comparison of String Distance Metrics for NameMatching Tasks / W.W. Cohen, P. Ravikumar, S.F. Fienberg // II Web. — 2003. — Vol. 3. — P. 73–78.

7. Долгопольский А.Б. Гипотеза древнейшего родства языковых семей Северной Евразии с вероятностной точки зрения / А.Б. Долгопольский // Вопросы языкознания. — 1964. — № 2. — С. 64–69.

8. Peter T. Analyzing genetic connections between languages by matching consonant classes / T. Peter, P. Ilia, G.M. Murray // Вопросы языкового родства. — 2010. — No. 5. — P. 117–126.

9. Circumpolar peoples and their languages: lexical and genomic data suggest ancient Chukotko-Kamchatkan —Nivkh and Yukaghir-Samoyedic connections / G. Starostin, N.E. Altınışık, M. Zhivlov [et al] // bioRxiv. — 2021. — Available at: https://www.biorxiv.org/content/10.1101/2021.02.27.433193v1.

10. Ляшевская О.Н. Новый частотный словарь русской лексики / О.Н. Ляшевская, С.А. Шаров. — Москва : Словари.ру, 2012. — 1087 с.

References

1. Mart'yanov V.I., Skumatov M.L. Adaptation of the Logical-Heuristic Approaches for Some Problems of the Phylogenetic Analysis. *Sovremennye tekhnologii. Sistemnyi analiz. Modelirovanie = Modern Technologies. System Analysis. Modeling*, 2011, no. 2, pp. 27–31. (In Russian). EDN: OEFVVT.

2. Borovskii A.V., Rakovskaya E.E., Karteeva I.A. Study of the Correlation between Toponyms of the Irkutsk Region and Sanskrit Words. *Izvestiya Baikal'skogo gosudarstvennogo universiteta = Bulletin of Baikal State University*, 2022, vol. 32, no. 3, pp. 610–622. (In Russian). EDN: SUWJUC. DOI: 10.17150/2500-2759.2022.32(3).610-622.

3. Borovskii A.V., Bratishchenko V.V., Rakovskaya E.E. The Study of the Relationship Between the Russian and Buryat Languages Using the Matrix of Proximity Measures Between Consonant Word Classes. *System Analysis&Mathematical Modeling*, 2023, vol. 5, no. 1, pp. 19–33. (In Russian). DOI: 10.17150/2713-1734.2023.5(1).19-33.

4. Levenshtein V.I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet physics doklady*, 1966, vol. 10, no. 8, pp. 707–710.

5. Ratcliff J.W., Metzener D.E. Pattern-Matching-the Gestalt Approach. *Dr. Dobb`s Journal*, 1988, vol. 13, no. 7, pp. 46.

6. Cohen W.W., Ravikumar P., Fienberg S.F. A Comparison of String Distance Metrics for NameMatching Tasks. *II Web*, 2003, vol. 3, pp. 73–78.

7. Dolgopol'skii A.B. Hypothesis of ancient relationship of language families of Northern Eurasia from a probabilistic point of view. *Voprosy jazykoznanija = Topics in the Study of Language*, 1964, no. 2, pp. 64–69. (In Russian).

8. Peter T., Ilia P., Murray G.M. Analyzing Genetic Connections between Languages by Matching Consonant Classes. *Voprosy yazykovogo rodstva = Journal of Language Relationship*, 2010, no. 5, pp. 117–126.

9. Starostin G., Altınışık N.E., Zhivlov M., Changmai P., Flegontova O. [et al] Circumpolar peoples and their languages: lexical and genomic data suggest ancient Chukotko-Kamchatkan —Nivkh and Yukaghir-Samoyedic connections. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2021.02.27.433193v1.

10. Lyashevskaya O.N., Sharov S.A. *A New Frequency Dictionary of Russian Vocabulary*. Moscow, Slovari.ru Publ., 2012. 1087 p.

Информация об авторах

Боровский Андрей Викторович — доктор физико-математических наук, профессор кафедры математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: andrei-borovskii@mail.ru, https://orcid.org/0000-0003-2119-1072, SPIN-код: 7243-8706, AuthorID РИНЦ: 22229.

Братищенко Владимир Владимирович — кандидат физико-математических наук, доцент, кафедра математических методов и цифровых технологий,

Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: vbrat56@mail.ru, https://orcid.org/0000-0002-7755-4170, SPIN-код: 4131-7181, AuthorID РИНЦ: 280037.

Раковская Елена Евгеньевна — аспирант, кафедра математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: rakovskaya19@mail.ru, https://orcid.org/0000-0002-2493-8699, SPIN-код: 1945-0473, AuthorID РИНЦ: 1097855.

Information about the Authors

Andrei V. Borovsky — D.Sc. in Physical and Mathematical Sciences, Professor, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, Russian Federation, e-mail: andrei-borovskii@mail.ru, https://orcid.org/0000-0003-2119-1072, SPIN-Code: 7243-8706, AuthorID RSCI: 22229.

Vladimir V. Bratishchenko — PhD in Physical and Mathematical Sciences, Associate Professor, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, Russian Federation, e-mail: vbrat56@mail.ru, https://orcid.org/0000-0002-7755-4170, SPIN-Code: 4131-7181, AuthorID RSCI: 280037.

Elena E. Rakovskaya — PhD Student, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, Russian Federation, e-mail: rakovskaya19@mail.ru, https://orcid.org/0000-0002-2493-8699, SPIN-Code: 1945-0473, AuthorID RSCI: 1097855.

Вклад авторов

Все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the Authors

The authors contributed equally to this article. The authors declare no conflicts of interests.

Для цитирования

Боровский А.В. Лексикостатистика взаимосвязи русского и бурятского языков / А.В. Боровский, В.В. Братищенко, Е.Е. Раковская. — DOI 10.17150/2713-1734.2023.5(3).303-318. — EDN PTMYCH // System Analysis & Mathematical Modeling. — 2023. — Т. 5, № 3. — С. 303–318.

For Citation

Borovsky A.V., Bratishchenko V.V., Rakovskaya E.E. Lexicostatistics of the Relationship Between the Russian and Buryat Languages. *System Analysis & Mathematical Modeling*, 2023, vol. 5, no. 3, pp. 303–318. (In Russian). EDN: PTMYCH. DOI: 10.17150/2713-1734.2023.5(3).303-318.