М. Е. Васильев
Институт востоковедения РАН (Россия, Москва); mvhumanity@gmail.com

Оценка достоверности исходных данных при лексикостатистических расчётах

В статье рассматривается проблема надёжности исходных данных, используемых в лексикостатистике для получения генеалогических классификаций и лингвистических датировок. Такие данные, представленные таблицей совпадений между основными списками языков, могут содержать ошибки или неточности, которые обусловлены сложностью и субъективностью процедуры этимологического анализа и существенно снижают достоверность лексикостатистических расчётов. Для решения этой проблемы предложена формальная методика, основанная на критерии взаимной согласованности (транзитивности) долей совпадений в лексикостатистической таблице. Использование этого критерия позволяет выявить недостоверные значения в исходных данных, а также численно оценить величину ошибки в каждом случае. Преимуществами предложенного подхода являются его простота и универсальность, объективность результатов, а также удобство реализации в виде компьютерного приложения. Апробация методики на материале романских и тюркских языков подтверждает её применимость и практическую эффективность при рассмотрении как небольших, так и крупных языковых групп.

Ключевые слова: лексикостатистика; глоттохронология; матрица расстояний; критерий согласованности.

Одним из главных факторов, влияющих на точность и достоверность генеалогических классификаций, а также лингвистических датировок, полученных с помощью лексикостатистики, является качество исходных данных. Первичными данными для большинства лексикостатистических исследований служат доли совпадений, которые рассчитываются путём сопоставления основных списков (ОС) рассматриваемых языков и установления количества этимологических соответствий (когнатов) между ними. Найденные значения долей совпадений, представленные в табличном виде, называют исходной лексикостатистической таблицей (или матрицей)[1].

Наиболее критичными и ответственными при сборе данных являются подготовительные этапы, а именно: составление списков базисной лексики и проведение сравнительного этимологического анализа. Поскольку обе эти процедуры плохо поддаются формализации, в большинстве случаев они проводятся вручную, что неизбежно приводит к субъективному («экспертному») характеру решений, как при отборе лексики, так и при определении этимологии. Как неоднократно было показано на разнообразном языковом материале, именно от качества собранных списков и их этимологического анализа в первую очередь зависит как точность лексикостатистических расчётов, так и в целом адекватность полученных результатов[2].

[1] По сути данная таблица является аналогом матрицы расстояний, используемой для кластеризации и анализа данных во многих научных областях (математической статистике, информатике, экономике, социологии, биологии и т. д.).

[2] В частности, работы Starostin 2000: 228; Милитарёв 2000; Бурлак & Старостин 2005: 148; Vydrin 2009: 112–114; дискуссия Соловьёв, Дыбо & Старостин 2010: 194; Starostin 2010: 84 убедительно демонстрируют,

Проблема неточности экспертных оценок и возможные пути её решения (или смягчения) широко обсуждаются в работах по компаративистике[3], однако большинство авторов сходятся во мнении, что процесс составления и этимологизации основных списков не может быть полностью формализован и в любом случае подразумевает некоторую субъективность. В связи с этим наиболее очевидным способом нивелировать возникающие ошибки представляется тщательная и всесторонняя проверка полученных данных — как самим исследователем, так и с привлечением других лингвистов.

Данный содержательный подход, при всех неоспоримых достоинствах, обладает также рядом очевидных недостатков. Во-первых, он не гарантирует выявление всех ошибок в исходных данных, а лишь повышает надёжность экспертных оценок за счёт увеличения их количества[4]. Во-вторых, результат проверки по-прежнему остаётся субъективным и в значительной степени зависит от возможностей компаративиста и/или его владения анализируемым материалом. В-третьих, ручное сравнение основных списков позволяет определить количество когнатов в базисной лексике каждой пары языков, однако не обеспечивает согласованность[5] полученных этимологических оценок. Наконец, в-четвёртых, процесс верификации списков и этимологий по своей трудоёмкости часто не уступает первоначальному сбору и анализу данных.

Перечисленные недостатки существенно ограничивают применимость такого подхода, а также значительно снижают его эффективность во многих случаях. При этом большинство указанных ограничений обусловлено качественным и экспертным характером процедуры верификации.

Возможной альтернативой этой процедуре могла бы стать некая формальная методика, позволяющая быстро и объективно оценить достоверность исходных данных, а также по возможности выявить заведомо недостоверные значения. Такую методику можно предложить на основе совокупного количественного анализа долей совпадений в исходной лексикостатистической таблице.

Представим два языка A и B, которые в равной степени близкородственны с некоторым третьим языком C. Очевидно, что в этом случае A и B не могут оказаться очень далёкими родственниками, так как это будет противоречить их близкому родству с идиомом C. Из этого следует, что доли совпадений, полученные в ходе попарного сравнения ОС этих языков, обладают взаимозависимостью, а значит — в какой-то степени определяют друг друга. Учитывая это обстоятельство, естественно предположить, что мы можем оценить достоверность долей совпадений для любой выбранной пары языков, опираясь на доли совпадений этих идиомов с остальными языками. При этом критерием достоверности будет согласованность (или непротиворечивость) этой оценки по отношению к остальным значениям лексикостатистической таблицы.

Поясним смысл согласованности оценок для группы из трёх языков A, B и C, доли совпадений между ОС которых представлены в табл. 1.

как систематические экспертные ошибки, допущенные на этом этапе, не только существенно сказываются на точности топологии генетических деревьев или глоттохронологических датировках, но и могут приводить к совершенно абсурдным результатам.

[3] См., например, Старостин 2007: 416–417; Militarev 2005: 345; Starostin 2010: 79–116.

[4] В скобках отметим, что и это не всегда осуществимо. Например, в случае с малоизученными языками (для которых применение лексикостатистики особенно актуально) верификация собранного материала другими специалистами зачастую невозможна из-за их отсутствия.

[5] Смысл согласованности будет подробно обсуждаться ниже в тексте статьи.

Согласно таблице, доля общей лексики в списках языков A и B составляет 98%, что соответствует разнице («расстоянию») в 100–98=2 слова[6] между ними. Аналогично для языков B и C доля совпадений равна 96% (расстояние в 4 слова), а для языков A и C — 97% (расстояние в 3 слова). Для наглядности обозначим доли совпадений и расстояния между языками на диаграмме (рис. 1).

Язык	Доли совпадений		
	A	B	C
A	1	0,98	0,97
B	0,98	1	0,96
C	0,97	0,96	1

Таблица 1. Доли совпадений ОС языков A, B и C

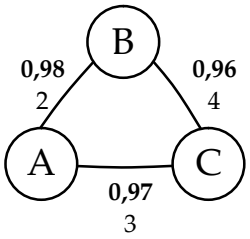

Рис. 1. Доли совпадений и расстояния между языками A, B и C

Если мы предположим, что 2 значения, различающие языки A и B, не совпадают с 4 значениями, отличающимися в языках B и C, то количество несовпадающих слов (расстояние) между языками A и C будет равно сумме расстояний A–B и B–C: 2+4=6.

Если же допустить, что оба эти значения входят в число четырёх различающихся значений идиомов B и C, то основные списки языков A и C будут отличаться всего 2 словами, что соответствует разнице тех же расстояний: 4–2=2.

Следовательно, расстояние между языками A и C может составлять от 2 до 6 слов, а доля совпадений между ними, соответственно, должна находиться в диапазоне 0,98…0,94 (рис. 2а). Очевидно, что значение N_{AC}=0,97, заданное в табл. 1, входит в этот интервал, а значит — согласуется с известными долями совпадений каждого из языков A и C с языком B (N_{AB} и N_{BC}).

Аналогичные рассуждения и расчёты, проведённые для остальных пар языков (AB и BC — рис. 2б, в), показывают, что значения N для любой пары идиомов согласованы относительно долей совпадений с оставшимся третьим языком.

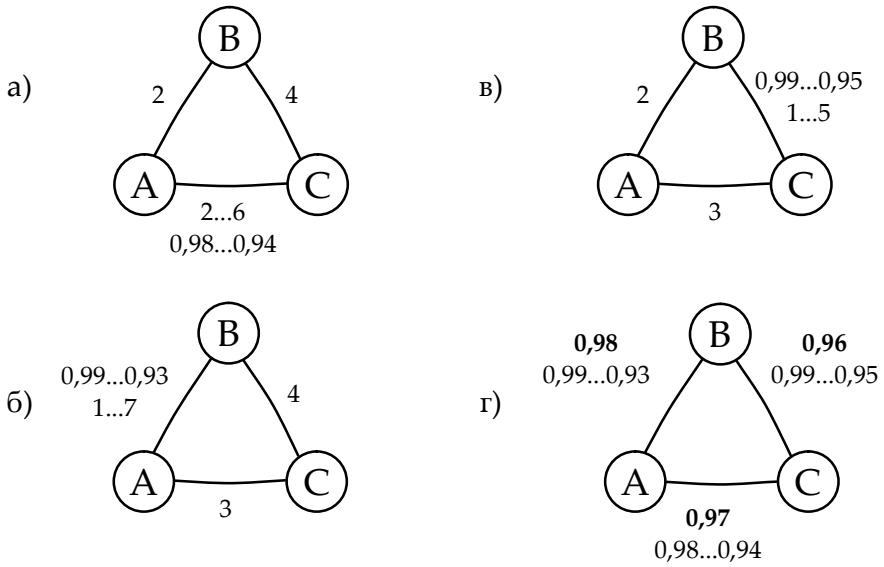

Рис. 2. Проверка взаимной согласованности значений N_{AB}, N_{BC}, N_{AC}

[6] При использовании полных 100-словных списков.

Сопоставляя найденные допустимые диапазоны и известные значения N_{AB}, N_{BC}, N_{AC} (рис. 2г), нетрудно убедиться, что все доли совпадения для триады языков A, B и C являются *взаимно согласованными*.

Используя тот же пример, смоделируем теперь ситуацию несогласованности исходных данных. Для этого изменим долю совпадений для языков B и C в исходной матрице (табл. 1) на 0,94. Полученные значения представлены в табл. 2 и на рис. 3.

Язык	Доли совпадений		
	A	B	C
A	1	0,98	0,97
B	0,98	1	0,94
C	0,97	0,94	1

Таблица 2. Доли совпадений ОС языков A, B и C (несогласованные значения)

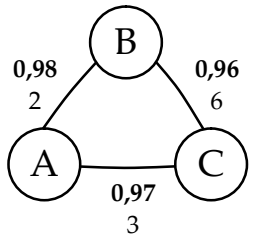

Рис. 3. Доли совпадений и расстояния между языками A, B и C (несогласованные значения)

В данном случае расстояние между языками A и C может принимать значения от 8 (6+2) до 4 (6–2) слов, а соответствующая доля значений N_{AC} должна находиться в диапазоне 0,92...0,96 (рис. 4а). При этом табличное значение N_{AC}=0,97 выходит за пределы этого интервала и не согласуется с двумя остальными долями совпадений. Более того, анализируя другие пары языков (рис. 4б, в), мы видим, что теперь ни одно значение N_{AB}, N_{BC} и N_{AC} не укладывается в диапазоны согласованности (рис. 4г). Таким образом, изменение всего одного значения N_{BC} в исходной таблице привело к нарушению согласованности одновременно для всех пар внутри рассматриваемой триады языков.

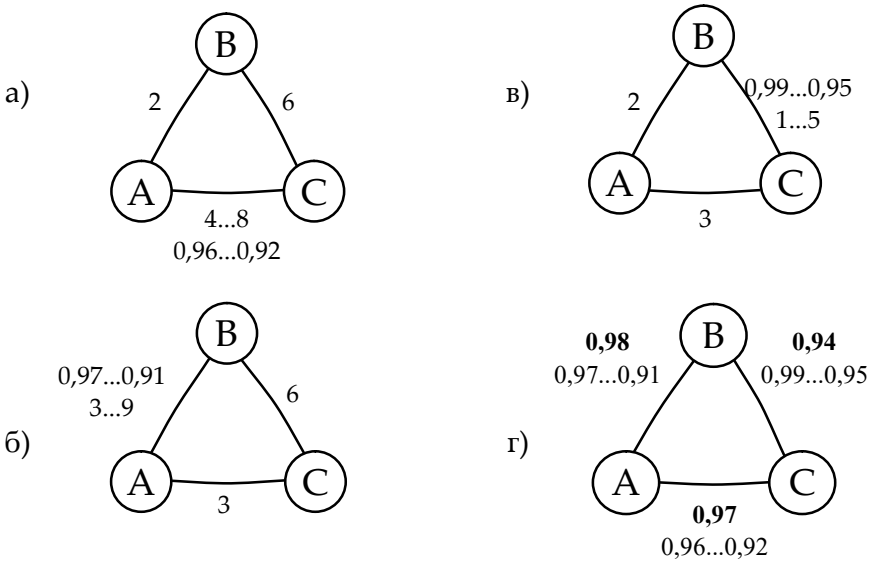

Рис. 4. Иллюстрация несогласованности значений N_{AB}, N_{BC}, N_{AC}

Обратим внимание, что при этом ошибка согласованности (E_c) для любой пары языков одинакова и составляет одно слово (см. рис. 4г). Так, для пары языков AB величина ошибки равна разнице между их долей совпадений (N_{AB}=0,98) и ближайшим значением из допустимого диапазона (0,97…0,91): E_{AB}=|0,98–0,97|=0,01[7].

[7] Аналогично для пары BC получаем E_{BC}=|0,94–0,95|=0,01, а для пары AC – E_{AC}=|0,97–0,96|=0,01.

Из рассмотренных примеров можно сделать вывод, что для проверки согласованности (или наоборот — выявления несогласованности) долей совпадений между тремя языками достаточно проверить согласованность любой пары из этих языков относительно третьего. Иначе говоря, согласованность значений N для любых двух идиомов в триаде языков является необходимым и достаточным условием согласованности всей рассматриваемой триады[8].

Как было показано на примерах выше, это условие выполняется в случае, когда доли совпадений для любых родственных языков A, B и C удовлетворяют двум следующим критериям:

1) доля несовпадающих значений (расстояние) между языками A и C не может быть больше суммы расстояний между языками A–B и B–C:

$$1 - N_{AC} \leq (1 - N_{AB}) + (1 - N_{BC})$$

или, после упрощения,

$$N_{AC} \geq N_{AB} + N_{BC} - 1; \quad (1)$$

2) расстояние между идиомами A и C не может быть меньше разницы расстояний между идиомами A–B и B–C, взятой по модулю:

$$1 - N_{AC} \geq |(1 - N_{AB}) - (1 - N_{BC})|,$$
$$N_{AC} \leq 1 - |N_{AB} - N_{BC}|. \quad (2)$$

Объединив неравенства (1) и (2), получим общее выражение условия согласованности для языков A, B и C:

$$N_{AB} + N_{BC} - 1 \leq N_{AC} \leq 1 - |N_{AB} - N_{BC}|. \quad (3)$$

Данное условие, вслед за С. А. Бурлак и С. А. Старостиным, можно также назвать *критерием транзитивности*[9].

Проверим справедливость критерия транзитивности в его формальном выражении на уже известных примерах. Для этого поочерёдно подставим в полученное неравенство (3) сначала согласованные (табл. 1), а затем — несогласованные данные (табл. 2):

$0{,}98 + 0{,}96 - 1 = 0{,}94 \leq 0{,}97 \leq 1 - |0{,}98 - 0{,}96| = 0{,}98$ (по данным табл. 1),
$0{,}98 + 0{,}94 - 1 = 0{,}92 \leq 0{,}97 \leq 1 - |0{,}98 - 0{,}94| = 0{,}96$ (по данным табл. 2).

Очевидно, что при подстановке согласованных долей совпадений (табл. 1) мы получаем верное неравенство ($0{,}94 \leq 0{,}97 \leq 0{,}98$) — т. е. критерий транзитивности соблюдается. Тогда как при расчёте по несогласованным значениям (табл. 2) правая часть неравенства ($0{,}92 \leq \mathbf{0{,}97} \leq \mathbf{0{,}96}$) не выполняется и, следовательно, исходные данные не удовлетворяют условию согласованности.

Перейдём к обсуждению применения критерия транзитивности для анализа реальных лексикостатистических данных. Очевидно, что при рассмотрении группы языков, число которых превышает три, мы должны проанализировать на согласованность все возможные триады идиомов в этой группе. Например, в группе, состоящей из 4-х языков A, B, C и D, нам потребуется проверить всего 4 триады: A–B–C, A–B–D, A–C–D и B–C–D.

[8] Данное утверждение можно доказать аналитически для общего случая. См. полное доказательство в Приложении 1.

[9] Это название, хорошо отражающее смысл условия согласованности, предложено в работе Бурлак & Старостин 2005: 103, в которой аналогичный ход рассуждений применяется для выявления скрытых заимствований в основных списках сравниваемых языков.

В общем случае, для группы из n языков, число таких триад определяется как количество сочетаний из n по три:

$$C_n^3 = \frac{n!}{(n-3)!3!},$$

например, для 5 языков (n=5) получим уже 10 триад, а для 9 языков (n=9) — 84 триады:

$$C_5^3 = \frac{5!}{(5-3)!3!} = 10,$$

$$C_9^3 = \frac{9!}{(9-3)!3!} = 84.$$

Используя критерий транзитивности, описанный выше, проанализируем согласованность лексикостатистических данных для небольшой группы романских языков (табл. 3).

№	Название языка	Доли совпадений основных список языков								
		LAT	ITA	FRA	PRT	ESP	GAL	VAL	PRV	ROM
LAT	классическая латынь	1	0,88	0,83	0,87	0,85	0,89	0,87	0,83	0,82
ITA	итальянский	0,88	1	0,93	0,90	0,89	0,92	0,93	0,90	0,88
FRA	французский	0,83	0,93	1	0,83	0,85	0,85	0,90	0,97	0,86
PRT	португальский	0,87	0,9	0,83	1	0,96	0,97	0,95	0,86	0,84
ESP	испанский	0,85	0,89	0,85	0,96	1	0,95	0,96	0,87	0,84
GAL	галисийский	0,89	0,92	0,85	0,97	0,95	1	0,95	0,87	0,87
VAL	вален. наречие каталанского	0,87	0,93	0,90	0,95	0,96	0,95	1	0,93	0,87
PRV	провансальский	0,83	0,90	0,97	0,86	0,87	0,87	0,93	1	0,84
ROM	румынский	0,82	0,88	0,86	0,84	0,84	0,87	0,87	0,84	1

Таблица 3. Доли совпадений основных списков романских языков[10]

В результате анализа были выявлены всего 5 триад с несогласованными значениями долей совпадений. Список этих триад, а также соответствующие им ошибки несогласованности представлены в табл. 4. Среди них 3 триады с ошибкой в два слова и 2 триады с ошибкой в одно слово.

Обозначения языков			Доли совпадений			Ошибка согласованности E_c
A	B	C	N_{AB}	N_{BC}	N_{AC}	
ESP	VAL	PRV	0,96	0,93	0,87	0,02
FRA	PRT	VAL	0,83	0,95	0,90	0,02
PRT	VAL	PRV	0,95	0,93	0,86	0,02
FRA	ESP	VAL	0,85	0,96	0,90	0,01
GAL	VAL	PRV	0,95	0,93	0,87	0,01

Таблица 4. Результаты анализа согласованности исходных данных для романских языков (указаны только несогласованные триады)

[10] Доли совпадений приводятся по данным из архива проекта «Глобальная лексикостатистическая база данных» (starlingdb.org/new100/main.htm).

С учётом того, что доля несогласованных триад составляет всего 6% от их общего количества (5 из 84), а величина ошибки не превышает 2 слова, исходную лексикостатистическую матрицу романских языков (табл. 3) можно считать хорошо согласованной. Тем не менее, попробуем установить причину несогласованности в указанных пяти триадах, опираясь на результаты проверки, приведённые в табл. 4.

В силу того, что взаимная согласованность всех долей совпадений в триаде языков нарушается при изменении любого из трёх значений, мы не можем установить, какое значение является недостоверным в отдельно взятой триаде. Однако, если проанализировать частотность вхождения отдельных языков (или пар языков) в несогласованные триады, можно косвенно определить, в каких ОС с наибольшей вероятностью содержатся недостоверные этимологические оценки.

В частности, при рассмотрении списка триад в табл. 4 мы обнаруживаем, что во все несогласованные триады входит валенсийское наречие каталанского (VAL), а вторым по частотности является провансальский (PRV), который встречается трижды. Это означает, что несогласованность триад, вероятнее всего, обусловлена недостоверными долями совпадений этих языков между собой (в триадах, где они встречаются вместе) или с другими идиомами (в остальных случаях). Следовательно, для устранения несогласованности исходных данных нам следует в первую очередь искать неточности, допущенные при сравнительном анализе валенсийского и провансальского основных списков[11]. При этом значение ошибки согласованности (в правом столбце табл. 4) показывает, на сколько слов (или процентов) мы должны скорректировать доли совпадений между идиомами, чтобы они стали согласованными.

Рассмотрим теперь более сложную ситуацию на примере группы тюркских языков, включающей 31 идиом. Проверка исходной таблицы долей совпадений (см. табл. 8 в Приложении 2), показывает, что согласованность данной группы заметно ниже, чем в случае с романскими языками[12]. Об этом свидетельствует как больший процент несогласованных триад (10% от общего количества), так и — что более существенно — повышение абсолютной величины ошибок согласованности.

Как видно из распределения в табл. 5, большинство несогласованных триад (343 из 467) имеют минимальные ошибки в 1–2 слова. Однако мы также фиксируем большое число ошибок в 3 слова (70 триад), 4 слова (34 триады), 5 слов (14 триад) и даже 6 слов (4 триады). При этом максимальное значение несогласованности достигает 7 и 8 слов! Такие значительные ошибки могут не только снизить точность лексикостатистических расчётов, но и существенно повлиять на их качественные результаты — например, при построении генеалогического дерева и определении принадлежности языка к той или иной языковой подгруппе[13].

[11] Отметим, что анализ частотности позволяет установить именно наиболее вероятных «виновников» несогласованности, однако нельзя исключать, что в отдельных случаях причиной рассогласования могут быть доли совпадений идиомов, которые встречаются всего в одной или нескольких несогласованных триадах.

[12] Можно предположить, что в любой лексикостатистической таблице, составленной по реальным языковым данным, всегда присутствует некоторое количество несогласованных значений, и чем больше группа рассматриваемых языков, тем больше будет её несогласованность. Однако это предположение ошибочно: например, при проверке таблицы совпадений для 16 дардских языков все 560 триад оказались согласованными (Васильев & Коган 2013: 158–159).

[13] В частности, хорошо известно, что метод «ближайших соседей» и его модификации, которые широко применяются для построения генетических классификаций, очень чувствительны к завышенным значениям долей совпадений. По этой причине даже единичные ошибочные значения в таблице исходных данных зачастую приводят к масштабным изменениям всей структуры дерева. См., например, Васильев 2010; Васильев & Коган 2013; Васильев & Саенко 2020.

Ошибка согласованности E_c, слов	Количество триад с ошибкой E_c
8	1
7	1
6	4
5	14
4	34
3	70
2	122
1	221
Всего	467

Таблица 5. Распределение триад по величине ошибки согласованности (тюркские языки)

Чтобы установить наиболее вероятные источники несогласованности в исходных данных, проанализируем список несогласованных триад (табл. 9 в Приложении 3) и выделим языки, которые встречаются в них чаще всего. Для удобства представим полученные результаты в виде двух разных таблиц: для отдельных языков (табл. 6) и для пар языков (табл. 7).

Язык	Число вхождений	Язык	Число вхождений	Язык	Число вхождений	Язык	Число вхождений
UZB	188	AZB	47	ALT	33	SJG	22
KLP	109	TRM	47	TUB	29	CHV	21
QUM	79	BLK	46	SHR	29	UIG	20
NOG	78	GAG	40	SAL	29	TOF	19
BAS	78	KRX	36	KMD	27	HAK	18
TAT	73	XAL	35	TUV	26	KRM	18
KAZ	71	TRK	35	KRQ	22	JAK	15
KRG	54	KUU	35	ATU	22		

Таблица 6. Список тюркских языков, входящих в несогласованные триады, упорядоченный по числу вхождений

При рассмотрении табл. 6 видно, что абсолютным лидером индивидуального списка является узбекский язык (UZB), который входит почти в 40% всех несогласованных триад (188 из 467), в том числе в триады с наибольшими ошибками согласованности 8, 7 и 6 слов (см. табл. 9). Вторым по частотности стал каракалпакский (KLP), входящий в состав 109 триад (около четверти от общего количества). Эти данные указывают на необходимость проверки и уточнения долей совпадения узбекского и каракалпакского ОС с остальными списками.

Для выявления конкретных пар языков с недостоверными значениями N целесообразно использовать табл. 7. При этом в первую очередь следует обращать внимание на частотные пары языков, входящие в триады с большой ошибкой согласованности (3, 4, 5 и более слов). К таким парам, например, относятся: KLP–KAZ, UZB–AZB, UZB–KAZ, UZB–SAL, BAS–KLP, UZB–KUU, UZB–GAG[14].

[14] Показательно, что во все эти пары входят узбекский либо каракалпакский языки, чаще всего встречающиеся в несогласованных триадах (см. табл. 6). Также отметим, что высокая частотность пар KLP-KAZ и

Как уже говорилось выше, с помощью критерия транзитивности невозможно однозначно определить, какое из трёх значений N является причиной несогласованности в триаде языков. Более того, недостоверными могут оказаться сразу два или все три значения. Поэтому данные о количестве вхождений отдельных языков или пар идиомов в несогласованные триады являются не строгой, а вероятностной оценкой. Тем не менее, они позволяют сделать процесс поиска недостоверных значений в исходных данных целенаправленным и системным, что многократно упрощает и ускоряет работу исследователя.

Обозначения пары языков		N_{AB}	Число вхождений	Распределение общего числа вхождений по значению ошибки E_c							
A	B			0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
KLP	KAZ	1.00	27	10	4	10	2	0	1	0	0
NOG	QUM	1.00	22	11	8	3	0	0	0	0	0
UZB	AZB	0.95	21	0	4	4	2	8	1	1	1
UZB	KAZ	0.99	20	4	4	4	5	1	1	0	1
BLK	QUM	0.98	20	10	7	1	2	0	0	0	0
UZB	SAL	0.95	19	3	8	2	5	1	0	0	0
UZB	QUM	0.98	18	6	7	5	0	0	0	0	0
BAS	KLP	0.98	18	4	5	3	4	2	0	0	0
UZB	KUU	0.93	17	4	7	3	2	1	0	0	0
TAT	BAS	0.99	16	11	3	2	0	0	0	0	0
UZB	TRM	0.97	16	5	6	4	1	0	0	0	0
NOG	KLP	0.99	16	9	4	2	1	0	0	0	0
UZB	BLK	0.95	16	6	6	2	1	1	0	0	0
UZB	GAG	0.90	16	4	6	1	4	1	0	0	0
UZB	ATU	0.95	15	6	4	4	1	0	0	0	0
UZB	SHR	0.91	15	3	5	5	1	0	1	0	0
UZB	KRX	0.96	15	2	8	3	2	0	0	0	0
…											

Таблица 7. Список пар тюркских языков, входящих в несогласованные триады, упорядоченный по числу вхождений (фрагмент)[15]

Рассмотренные выше примеры хорошо демонстрируют основные преимущества описанной методики, обусловленные её формальным характером. К ним относится простота и универсальность[16] применения, объективность получаемых результатов, а также возможность не только качественной, но и количественной оценки достоверности исходных данных. При этом все необходимые вычисления могут быть легко реализованы в виде компьютерной программы и выполняться полностью автоматически, что открыва-

NOG-QUM объясняется полным совпадением их ОС (N=1, табл. 7). В этом случае критерий транзитивности будет выполняться только тогда, когда доли совпадений каждого из двух языков с любым третьим языком попарно равны.

[15] Таблица с полным списком всех пар приводится в сопровождающих материалах, доступных на сайте журнала.

[16] Методика может применяться к исходным лексикостатистическим данным, полученным любым способом, и не влияет на выбор моделей и методов, используемых в ходе дальнейшего исследования.

ет широкие возможности для её практического использования как отдельно, так и в составе уже существующих лексикостатистических приложений.

В то же время, вместе с достоинствами такого формального подхода следует также отметить его ограничения. Прежде всего, нужно учитывать, что согласованность долей совпадений не гарантирует их достоверности[17]. Использование критерия транзитивности позволяет нам выявить только те значения, которые привели к ошибкам согласованности исходных данных, однако это не означает, что в них не содержится других неточностей или ошибок.

Во-вторых, предлагаемая методика является по сути «диагностическим» инструментом, с помощью которого можно быстро оценить согласованность исходной таблицы и установить недостоверные значения, нуждающиеся в пересмотре. Однако выяснение причин несогласованности и собственно корректировка долей совпадений невозможны без детального этимологического анализа основных списков.

Таким образом, на практике наиболее эффективным и целесообразным представляется комбинированное применение формального и содержательного подходов, при котором сначала проводится автоматизированная проверка исходных данных на согласованность, а затем несогласованные доли совпадений, выявленные в ходе проверки, анализируются и корректируются вручную специалистом[18]. При необходимости эта процедура может быть повторена несколько раз до достижения полной согласованности исходных данных, либо уменьшения количества и величины ошибок согласованности до приемлемых значений.

Приложение 1
Доказательство необходимого и достаточного условия согласованности оценок долей совпадений ОС для трёх языков

Докажем, что необходимым и достаточным условием согласованности долей совпадений между ОС трёх языков является согласованность долей совпадений для любой одной пары языков (AB, BC или AC) по отношению к третьему языку (C, A или B).

1. Условие согласованности оценок N_{AB}, N_{BC}, N_{AC} для трёх языков A, B и C не выполняется по определению, если это условие не выполняется хотя бы для одной пары языков по отношению к соответствующему третьему языку.

Необходимое условие доказано.

2. Допустим, что условие согласованности выполняется для пары языков A и C:

$$N_{AB} + N_{BC} - 1 \le N_{AC} \le 1 - |N_{AB} - N_{BC}|. \quad (1)$$

Для компактного изложения рассуждений перепишем (1) в виде:

$$a + b - 1 \le c \le 1 - |a - b|,$$

где обозначения определяются почленным сопоставлением (1) и (2).

[17] В строгом смысле согласованность долей совпадений является необходимым, но не достаточным условием их достоверности.

[18] Либо соответствующие идиомы целиком исключаются из рассмотрения, если причину несогласованности установить не удалось.

Для случая a≥b (2) примет вид:

$$a + b - 1 \le c \le 1 - a + b,$$

от которого можно перейти к двум неравенствам:

$$a - c - 1 \le -b; \quad (3)$$
$$-b \le 1 - a - c. \quad (4)$$

Из (4) получаем

$$a + c - 1 \le b, \quad (5)$$

и (3) и (4) можно переписать в виде:

$$a \le 1 - b + c;$$
$$a \le 1 + b - c,$$

который соответствует выражению

$$a \le 1 - |b - c| \quad (6)$$

для произвольных b и c.

Раскрывая (2) для случая a≤b, получим аналогично:

$$a + b - 1 \le c \le 1 - b + a, \quad (7)$$
$$b - c - 1 \le -a;$$
$$-a \le 1 - b - c.$$

и, далее,

$$b + c - 1 \le a, \quad (8)$$

а также

$$b \le 1 - a + c;$$
$$b \le 1 + a - c,$$

что соответствует

$$b \le 1 - |a - c| \quad (9)$$

для произвольных a и c.

Таким образом, при выполнении неравенства (2) для пары языков A и C:

$$a + b - 1 \le c \le 1 - |a - b|, \quad (2)$$

выполняются условия (5), (6), (8), (9):

$$b + c - 1 \le a, \quad (8)$$
$$a \le 1 - |b - c| \quad (6)$$
$$a + c - 1 \le b, \quad (5)$$
$$b \le 1 - |a - c|, \quad (9)$$

которые соответствуют искомым условиям согласованности оценок для пар языков A и B, B и C по отношению к третьему языку C и A:

$$N_{BC} + N_{AC} - 1 \le N_{AB} \le 1 - |N_{BC} - N_{AC}|;$$
$$N_{AB} + N_{AC} - 1 \le N_{BC} \le 1 - |N_{AB} - N_{AC}|.$$

Достаточное условие доказано.

Приложение 2

№	LANG	TRK	TAT	UZB	UIG	SJG	AZB	TRM	HAK	CHV	JAK	TUV	KRG	NOG	BAS
1	TRK	1	0,86	0,9	0,85	0,8	0,94	0,92	0,81	0,73	0,74	0,77	0,88	0,88	0,82
2	TAT	0,86	1	0,98	0,97	0,85	0,88	0,91	0,89	0,77	0,77	0,8	0,98	0,99	0,99
3	UZB	0,9	0,98	1	0,96	0,9	0,95	0,97	0,91	0,79	0,79	0,84	0,99	0,98	0,96
4	UIG	0,85	0,97	0,96	1	0,9	0,88	0,91	0,9	0,73	0,77	0,82	0,95	0,96	0,94
5	SJG	0,8	0,85	0,9	0,9	1	0,82	0,86	0,86	0,71	0,78	0,83	0,86	0,87	0,83
6	AZB	0,94	0,88	0,95	0,88	0,82	1	0,96	0,82	0,78	0,74	0,76	0,89	0,88	0,84
7	TRM	0,92	0,91	0,97	0,91	0,86	0,96	1	0,88	0,77	0,77	0,81	0,94	0,95	0,91
8	HAK	0,81	0,89	0,91	0,9	0,86	0,82	0,88	1	0,71	0,79	0,88	0,9	0,88	0,88
9	CHV	0,73	0,77	0,79	0,73	0,71	0,78	0,77	0,71	1	0,69	0,67	0,77	0,77	0,75
10	JAK	0,74	0,77	0,79	0,77	0,78	0,74	0,77	0,79	0,69	1	0,76	0,79	0,78	0,78
11	TUV	0,77	0,8	0,84	0,82	0,83	0,76	0,81	0,88	0,67	0,76	1	0,85	0,81	0,79
12	KRG	0,88	0,98	0,99	0,95	0,86	0,89	0,94	0,9	0,77	0,79	0,85	1	0,97	0,96
13	NOG	0,88	0,99	0,98	0,96	0,87	0,88	0,95	0,88	0,77	0,78	0,81	0,97	1	0,97
14	BAS	0,82	0,99	0,96	0,94	0,83	0,84	0,91	0,88	0,75	0,78	0,79	0,96	0,97	1
15	BLK	0,83	0,93	0,95	0,9	0,83	0,85	0,89	0,86	0,73	0,8	0,8	0,93	0,96	0,93
16	GAG	0,98	0,84	0,9	0,86	0,78	0,92	0,9	0,79	0,73	0,75	0,74	0,88	0,86	0,81
17	KRM	0,86	0,94	0,95	0,92	0,84	0,88	0,92	0,86	0,72	0,75	0,79	0,94	0,92	0,93
18	KLP	0,89	0,98	0,99	0,96	0,87	0,89	0,95	0,93	0,77	0,78	0,85	0,99	0,99	0,98
19	ATU	0,87	0,9	0,95	0,92	0,88	0,88	0,89	0,87	0,76	0,84	0,79	0,91	0,9	0,89
20	SAL	0,85	0,89	0,95	0,92	0,84	0,9	0,92	0,84	0,76	0,74	0,75	0,9	0,91	0,87
21	SHR	0,79	0,86	0,91	0,9	0,85	0,8	0,85	0,95	0,67	0,76	0,85	0,89	0,87	0,86
22	TOF	0,76	0,8	0,83	0,82	0,81	0,75	0,79	0,87	0,68	0,77	0,97	0,84	0,82	0,79
23	KAZ	0,86	0,97	0,99	0,95	0,85	0,86	0,94	0,9	0,75	0,75	0,79	0,99	0,98	0,95
24	QUM	0,87	0,96	0,98	0,95	0,86	0,91	0,93	0,88	0,79	0,77	0,79	0,94	1	0,97
25	KRX	0,88	0,92	0,96	0,93	0,9	0,89	0,92	0,91	0,78	0,86	0,86	0,92	0,9	0,9
26	KRQ	0,92	0,95	0,94	0,94	0,85	0,92	0,91	0,88	0,72	0,74	0,78	0,93	0,95	0,93
27	ALT	0,82	0,92	0,95	0,94	0,86	0,85	0,89	0,93	0,72	0,76	0,88	0,94	0,92	0,91
28	XAL	0,86	0,82	0,9	0,86	0,8	0,9	0,92	0,8	0,77	0,76	0,73	0,85	0,85	0,81
29	KUU	0,81	0,89	0,93	0,91	0,85	0,83	0,88	0,9	0,69	0,74	0,84	0,92	0,9	0,86
30	KMD	0,77	0,86	0,88	0,89	0,87	0,78	0,86	0,88	0,69	0,73	0,84	0,89	0,87	0,88
31	TUB	0,81	0,91	0,92	0,94	0,86	0,83	0,88	0,91	0,71	0,77	0,86	0,93	0,91	0,93

BLK	GAG	KRM	KLP	ATU	SAL	SHR	TOF	KAZ	QUM	KRX	KRQ	ALT	XAL	KUU	KMD	TUB
0,83	0,98	0,86	0,89	0,87	0,85	0,79	0,76	0,86	0,87	0,88	0,92	0,82	0,86	0,81	0,77	0,81
0,93	0,84	0,94	0,98	0,9	0,89	0,86	0,8	0,97	0,96	0,92	0,95	0,92	0,82	0,89	0,86	0,91
0,95	0,9	0,95	0,99	0,95	0,95	0,91	0,83	0,99	0,98	0,96	0,94	0,95	0,9	0,93	0,88	0,92
0,9	0,86	0,92	0,96	0,92	0,92	0,9	0,82	0,95	0,95	0,93	0,94	0,94	0,86	0,91	0,89	0,94
0,83	0,78	0,84	0,87	0,88	0,84	0,85	0,81	0,85	0,86	0,9	0,85	0,86	0,8	0,85	0,87	0,86
0,85	0,92	0,88	0,89	0,88	0,9	0,8	0,75	0,86	0,91	0,89	0,92	0,85	0,9	0,83	0,78	0,83
0,89	0,9	0,92	0,95	0,89	0,92	0,85	0,79	0,94	0,93	0,92	0,91	0,89	0,92	0,88	0,86	0,88
0,86	0,79	0,86	0,93	0,87	0,84	0,95	0,87	0,9	0,88	0,91	0,88	0,93	0,8	0,9	0,88	0,91
0,73	0,73	0,72	0,77	0,76	0,76	0,67	0,68	0,75	0,79	0,78	0,72	0,72	0,77	0,69	0,69	0,71
0,8	0,75	0,75	0,78	0,84	0,74	0,76	0,77	0,75	0,77	0,86	0,74	0,76	0,76	0,74	0,73	0,77
0,8	0,74	0,79	0,85	0,79	0,75	0,85	0,97	0,79	0,79	0,86	0,78	0,88	0,73	0,84	0,84	0,86
0,93	0,88	0,94	0,99	0,91	0,9	0,89	0,84	0,99	0,94	0,92	0,93	0,94	0,85	0,92	0,89	0,93
0,96	0,86	0,92	0,99	0,9	0,91	0,87	0,82	0,98	1	0,9	0,95	0,92	0,85	0,9	0,87	0,91
0,93	0,81	0,93	0,98	0,89	0,87	0,86	0,79	0,95	0,97	0,9	0,93	0,91	0,81	0,86	0,88	0,93
1	0,83	0,93	0,94	0,9	0,87	0,84	0,8	0,93	0,98	0,89	0,9	0,89	0,81	0,86	0,83	0,88
0,83	1	0,84	0,88	0,85	0,84	0,77	0,74	0,85	0,86	0,88	0,9	0,82	0,84	0,79	0,78	0,8
0,93	0,84	1	0,93	0,89	0,9	0,87	0,79	0,92	0,94	0,89	0,95	0,9	0,86	0,87	0,86	0,91
0,94	0,88	0,93	1	0,92	0,9	0,9	0,85	1	0,97	0,94	0,94	0,95	0,87	0,92	0,89	0,93
0,9	0,85	0,89	0,92	1	0,89	0,85	0,81	0,9	0,91	0,97	0,88	0,89	0,88	0,88	0,84	0,86
0,87	0,84	0,9	0,9	0,89	1	0,84	0,76	0,89	0,91	0,89	0,89	0,88	0,92	0,85	0,83	0,85
0,84	0,77	0,87	0,9	0,85	0,84	1	0,82	0,87	0,86	0,9	0,86	0,91	0,79	0,9	0,88	0,9
0,8	0,74	0,79	0,85	0,81	0,76	0,82	1	0,81	0,81	0,83	0,77	0,88	0,72	0,84	0,84	0,86
0,93	0,85	0,92	1	0,9	0,89	0,87	0,81	1	0,96	0,91	0,93	0,92	0,83	0,9	0,88	0,9
0,98	0,86	0,94	0,97	0,91	0,91	0,86	0,81	0,96	1	0,91	0,94	0,92	0,85	0,89	0,86	0,89
0,89	0,88	0,89	0,94	0,97	0,89	0,9	0,83	0,91	0,91	1	0,91	0,89	0,88	0,87	0,85	0,88
0,9	0,9	0,95	0,94	0,88	0,89	0,86	0,77	0,93	0,94	0,91	1	0,89	0,88	0,86	0,84	0,89
0,89	0,82	0,9	0,95	0,89	0,88	0,91	0,88	0,92	0,92	0,89	0,89	1	0,82	0,96	0,95	0,96
0,81	0,84	0,86	0,87	0,88	0,92	0,79	0,72	0,83	0,85	0,88	0,88	0,82	1	0,79	0,78	0,8
0,86	0,79	0,87	0,92	0,88	0,85	0,9	0,84	0,9	0,89	0,87	0,86	0,96	0,79	1	0,96	0,96
0,83	0,78	0,86	0,89	0,84	0,83	0,88	0,84	0,88	0,86	0,85	0,84	0,95	0,78	0,96	1	0,96
0,88	0,8	0,91	0,93	0,86	0,85	0,9	0,86	0,9	0,89	0,88	0,89	0,96	0,8	0,96	0,96	1

Таблица 8. Доли совпадений основных списков тюркских языков[19]

[19] Доли совпадений приводятся по данным из архива проекта «Глобальная лексикостатистическая база данных» (starlingdb.org/new100/main.htm).

Приложение 3

Обозначения языков			Доли совпадений			Ошибка согласованности E_c
A	B	C	N_{AB}	N_{BC}	N_{AC}	
UZB	AZB	KAZ	0.95	0.86	0.99	0.08
UZB	AZB	BAS	0.95	0.84	0.96	0.07
UZB	KAZ	XAL	0.99	0.83	0.90	0.06
UZB	AZB	SHR	0.95	0.80	0.91	0.06
TUV	**KLP**	KAZ	0.85	1,00	0.79	0.06
TAT	**UZB**	XAL	0.98	0.90	0.82	0.06
UZB	AZB	KUU	0.95	0.83	0.93	0.05
TRK	BAS	**KLP**	0.82	0.98	0.89	0.05
UZB	AZB	ALT	0.95	0.85	0.95	0.05
UZB	AZB	KMD	0.95	0.78	0.88	0.05
TAT	**UZB**	AZB	0.98	0.95	0.88	0.05
UZB	BAS	GAG	0.96	0.81	0.90	0.05
UZB	BAS	XAL	0.96	0.81	0.90	0.05
BAS	GAG	**KLP**	0.81	0.88	0.98	0.05
UZB	SAL	KAZ	0.95	0.89	0.99	0.05
UZB	AZB	KRG	0.95	0.89	0.99	0.05
UZB	AZB	NOG	0.95	0.88	0.98	0.05
…						

Таблица 9. Результаты анализа согласованности исходных данных для тюркских языков (фрагмент таблицы)[20]

Обозначения языков

TRK — турецкий, TAT — татарский, UZB — узбекский, UIG — уйгурский, SJG — сарыюгурский, AZB — азербайджанский, TRM — туркменский, HAK — хакасский, CHV — чувашский, JAK — якутский, TUV — тувинский, KRG — киргизский, NOG — ногайский, BAS — башкирский, BLK — балкарский, GAG — гагаузский, KRM — караимский, KLP — каракалпакский, ATU — древнетюркский, SAL — саларский, SHR — шорский, TOF — тофаларский, KAZ — казахский, QUM — кумыкский, KRX — караханидский, KRQ — крымский караимский, ALT — алтайский, XAL — халаджский, KUU — челканский (куу), KMD — кумандинский, TUB — тубаларский.

Литература

Бурлак, С. А., С. А. Старостин. 2005. *Сравнительно-историческое языкознание*. Москва: «Академия».

Васильев, М. Е. 2010. Об использовании лексического критерия для построения генеалогической классификации. *Бюллетень Общества востоковедов РАН* 17: 530–572.

Васильев, М. Е., А. И. Коган. 2013. К вопросу о восточнодардской языковой общности. *Вопросы языкового родства* 16(117): 149–177.

[20] Таблица с полным списком несогласованных триад приводится в сопровождающих материалах, доступных на сайте журнала.

Васильев, М. Е., М. Н. Саенко. 2020. Анализ топологии и оценка точности лексикостатистических классификаций (на примере славянских языков). *Вопросы языкового родства* 18/4: 320–347.

Милитарёв, А. Ю. 2000. О дохлой лошади, оказавшейся скакуном. *Знание — сила* 4: 20–31.

Соловьёв, В. Д., А. В. Дыбо, Г. С. Старостин. 2010. Типологическая схожесть языков как метод изучения языковой эволюции. Дискуссия. *Вопросы языкового родства* 4: 177–198.

Старостин, С. А. 2007. Сравнительно-историческое языкознание и лексикостатистика. В: С. А. Старостин. *Труды по языкознанию:* 407–447. Москва: «Языки славянских культур».

References

Burlak, S. A., S. A. Starostin. 2005. *Sravnitel'no-istoricheskoe jazykoznanie.* Moskva: «Akademija».

Militarev, Alexander. 2005. Once more about glottochronology and the comparative method: the Omotic-Afrasian case. In: Ilya Smirnov (ed.). *Orientalia et Classica. Trudy Instituta vostochnyx kultur i antichnosti:* 339–408. Moscow: RSUH Publishers.

Militarev, A. Ju. 2000. O doxloj loshadi, okazavshejs'a skakunom. *Znanie — sila* 4: 20–31.

Solovjev, V. D., A. V. Dybo, G. S. Starostin. 2010. Tipologicheskaja sxozhest' jazykov kak metod izuchenija jazykovoj evol'utsii. Diskussija. *Journal of Language Relationship* 4: 177–198.

Starostin, George. 2010. Preliminary lexicostatistics as a basis for language classification: A new approach. *Journal of Language Relationship* 3: 79–116.

Starostin, S. A. 2007. Sravnitel'no-istoricheskoe jazykoznanie i leksikostatistika. In: S. Starostin. *Trudy po jazykoznaniju:* 407–447. Moskva: «Jazyki slav'anskix kul'tur».

Starostin, Sergey. 2000. Comparative-historical linguistics and lexicostatistics. In: Colin Renfrew, April McMahon, Larry Trask (eds.). *Time Depth in Historical Linguistics*, Vol. 1: 223–259. Cambridge: The McDonald Institute for Archaeological Research Publications.

Vasilyev, M. E. 2010. Ob ispol'zovanii leksicheskogo kriterija dl'a postrojenija genealogicheskoj klassifikatsii. *B'ulleten' Obshchestva vostokovedov RAN* 17: 530–572.

Vasilyev, M. E., A. I. Kogan. 2013. K voprosu o vostochnodardskoj jazykovoj obshchnosti. *Journal of Language Relationship* 16(117): 149–177.

Vasilyev, M. E., M. N. Saenko. 2020. Analiz topologii i otsenka tochnosti leksikostatisticheskix klassifikatsij (na primere slav'anskix jazykov). *Journal of Language Relationship* 18/4: 320–347.

Vydrin, Valentin. 2009. On the Problem of the Proto-Mande Homeland. *Journal of Language Relationship* 1: 107–142.

Mikhail Vasilyev. Evaluating the reliability of source data in lexicostatistical analysis

The article addresses the reliability of the source data used in comparative-historical linguistics for lexicostatistical purposes (to obtain genealogical classifications and linguistic dating). Such data, usually represented by a table of cognacy percentages between the basic wordlists of languages, may contain errors or inaccuracies, stemming from the complexity and subjectivity of the etymological analysis procedure, which may significantly reduce the reliability of lexicostatistical calculations. To solve this issue, a formal methodology is proposed based on the criterion of consistency (or transitivity) of percentage values in the initial lexicostatistical table. Applying this criterion enables the identification of unreliable values in the source data, as well as the numerical estimation of data inconsistency in each case. The advantages of the proposed approach include its simplicity and versatility, the objectivity of the results, and ease of implementation as a computer application. Testing the methodology on the lexicostatistical data of Romance and Turkic languages proves its applicability and practical efficiency while examining both small and large language groups.

Keywords: lexicostatistics; glottochronology; distance matrix; consistency criterion.