Транслитерация памирских языков
On transliterating Pamir-language orthographies

Бахтоваршоев А.Ш.
Bakhtovarshoev A.Sh.

В статье рассмотрен вопрос о преобразовании (транслитерации) алфавитов некоторых памирских языков (шугнанского и рушанского) с кириллицы на латиницу и наоборот. Предложена и проанализирована конкретная таблица преобразования алфавитов по системам А и Б. Показано, что предложенная таблица сохраняет естественную частоту появления символов на данных языках и может быть использована в практической работе.

Ключевые слова: информационные технологии, транслитерация, памирские языки, шугнанский язык, рушанский язык

This article illustrates the transformation (transliteration) of the alphabets of two Pamir languages (Shughnani and Rushani) from Cyrillic to Latin script and vice versa. The author suggests a transliteration table for these alphabets based on two different systems and demonstrates that this table preserves the natural frequency of the letters in the specified languages and can be applied in practice.

Keywords: information technology, transliteration, Pamir languages, Shughnani, Rushani

В настоящей статье рассматриваются вопросы преобразования текстов шугнанско-рушанской языковой группы с кириллицы на латиницу и наоборот. Следует констатировать, что для данной группы языков малочисленных народов подобные средства ещё не созданы. Поэтому приходится применять разработки, существующие на сегодняшний день для конкретного языка, а именно для русского. Впрочем, большая часть таких программ применяется в режиме онлайн,

т. е. непосредственно в Интернете, что в целом не особенно удобно. При этом на многих сайтах существуют программы, преобразующие кириллицу (в частности, для русского, таджикского языков) в латиницу и наоборот, которые на жаргоне программистов обычно выступают под названием «на (с) транслит(а)», при этом вопрос о безопасности существующего на них контента, загружаемого для осуществления транслитерации, для пользователей остаётся открытым.

Транслитерация обычно применяется в следующих случаях: а) при переводе текста с одного языка на другой (иногда с использованием автоматического, машинного перевода); б) когда носители языка пользуются различными алфавитами; в) при обработке данных; г) для программ, преобразующих текст в речь; д) для веб-адресов; е) в мобильных телефонах для передачи сообщений. Очевидно, существуют и другие области применения. В случае (г) транслитерацию вынужден применять пользователь программы, если она ориентирована на какой-то конкретный алфавит. Пользователь такой программы должен «подать» ей текстовый материал только в том виде, которую она может воспринимать. Применение пункта (е) в России с 2007 г. считается нарушением законодательства, но в некоторых странах он до сих пор действует.

Обращаем внимание на тот факт, что до сих пор ни для памирских языков в целом, ни для языков шугнано-рушанской группы нет алфавита, принятого всем языковым сообществом. Имеются алфавиты, разработанные и применяемые в научных трудах специалистов-языковедов, созданные на основе иранистической международной транскрипции (например, [Эдельман 1963; Соколова 1966; Пахалина 1969; Додыхудоева 2005; Евангелие 2001а; 2001б]), а также варианты, использовавшиеся как в учебниках начальной школы, так и в научных исследованиях [Усманов, Гуломсафдаров 2009; Усманов, Кадамшоев 2009; Бахтоваршоев 2013; Карамшоев, Аламшоев 1996].

Таким образом, поскольку официально принятый единый стандартный вариант алфавита отсутствует, то целесообразен

только один путь — взять за основу один из возможных вариантов алфавита. Практическая работа с конкретным алфавитом поможет выявить ряд аспектов рассматриваемой проблемы. В данной статье мы придерживаемся, с одним дополнением, набора фонем, описанных В.С. Соколовой [Соколова 1966]. Наше дополнение состоит в том, что в алфавит введен аллофон фонемы [i], который обозначен в нём как буква «э». Этот дополненный набор фонем и есть та основа, на которой без существенных фонетических изъянов может быть записано любое выражение шугнанско-рушанской группы языков [Зарубин 1960, Соколова 1966]. Как правило, научные работы ведутся с использованием иранистической международной транскрипции, которая для «массового» использования является неудобной и затруднительной, так как одновременно применяются разные алфавиты. Более практичный подход состоит в том, чтобы применить один алфавит, который имеется в виде стандартной раскладки клавиатуры. Такой алфавит на основе кириллицы нами был составлен и подвергнут анализу [Бахтоваршоев 2013]. Этот алфавит, который мы условно называем алфавит Зарубина-Соколовой, по нашему мнению, имеет минимальные требования к программным продуктам и к квалификации пользователей операционных систем, легко может быть выучен, поэтому его «массовое» применение на практике не вызывает трудности.

На основании результатов работ [Усманов, Гуломсафдаров 2009; Бахтоваршоев 2013] можно утверждать, что тексты, напечатанные на алфавите Зарубина-Соколовой, имеют почти такие же статистические характеристики, как тексты других языков на естественных, «родных», алфавитах. Таким образом, цель данной работы состоит в том, чтобы в процессе выбора конкретных букв или комбинаций букв для транслитерации сохранить или существенно не ухудшить статистические характеристики транслитерированного текста по сравнению с оригиналом. Ведь упрощение или видоизменение записи текста не предполагает изменения его содержания и других объективных

свойств. Кроме того, простота и естественность способствует более лёгкому восприятию текста. Так как предварительно неизвестно, с каким конкретно реализованным алфавитом будут иметь дело программные продукты — с вариантом А или Б, — то мы должны рассмотреть оба варианта. Вариант А, конечно, предназначен для квалифицированных пользователей, и, скорее всего, будет применяться в исключительных случаях. Вариант Б рассчитан на рядового пользователя.

Многочисленные материалы, которые хаотически публикуются носителями языка как в виде печатной продукции, так и в Интернете, выявляют низкий уровень владения носителями правилами правописания, что происходит из-за отсутствия обучения родному языку в школе.

Несмотря на отсутствие обучения родному языку и на родном языке в школе, на протяжении XX в. были подготовлены учебные материалы для школы. Однако по тем или иным причинам подобные материалы не подходят для решения задач транслитерации, указанных выше. Так, известный «Букварь» шугнанского языка Д.К. Карамшоева и М. Аламшоева [1996] имеет весьма существенный недостаток, так как в силу недостаточного оснащения населения техническими средствами в конце 1990-х, он был ориентирован на обучение письму от руки.

Алфавит Зарубина-Соколовой, латинский вариант которого подвергнут анализу в настоящей работе, рассчитан в первую очередь, на применение информационных технологий.

Прежде чем приступить к рассмотрению вопроса о преобразовании, предварительно укажем те обозначения специалистов-языковедов, которые, ввиду их громоздкости и отличия от стандартных символов, мы заменили на более простые. [Соколова 1966; Пахалина 1969; Усманов, Кадамшоев 2009]. В свою очередь и они по сути были ориентированы на письмо от руки (или на набор на печатной машинке), и поэтому в настоящее время имеют узкую сферу применения. Выражения следующие: $\bar{\text{a}}$, $\bar{\text{o}}$, $\bar{\text{u}}$, $\check{\text{a}}$, $\check{\text{j}}$, $\check{\text{x}}$. Они сегодня, в сущности,

являются формулами (т. е. созданными и внедрёнными объектами программы Microsoft Equation; обозначим их кратко как «форм. МЕ»). Почти такие же конструкции находим в работе [Языки мира 1999]. Стандартные таблицы символов обычных текстовых редакторов операционных систем не содержат подобных выражений. В данной работе они заменены на другие символы, которые указаны ниже в таблице 1.

Таблица 1. Соответствие символов

№	Символы, принятые иранистами-языковедами	Система А (лат.-греч.)	Система Б (лат.-греч.)	Кириллица	Примечания
1	(j с кароном)	j	j	ҷ	форм. МЕ
2	(o с чертой и двумя точками)	ō	o	оо	форм. МЕ
3	(u с чертой и кружочком)	ů	uo	ӯ	форм. МЕ
4	(x с кароном)	–	x'	х	форм. МЕ
5	(γ с кароном)	γ	γ	гь	форм. МЕ
6	(ε с чертой)	ε	ε	ээ	форм. МЕ
7	ʒ	ʒ	dz	дз	ст. симв.
8	ə	ə	ə	э	ст. симв.
9	æ	æ	æ	э', ээ	ст. симв.

Автором предлагается следующий вариант преобразования:

Таблица 2. Транслитерация по А и Б

	Кириллица	Система А (лат.)	Система Б (лат.)	прим.
01	а	a	a	станд.
02	Аа/аа	ā	Aa/aa	А – ст.
03	б	b	b	станд.
04	в	v	v	станд.
05	Вь/вь	w	w	научн.
06	г	g	g	станд.
07	гь	ĝ	g′	пр. авт.

	Кириллица	**Система А (лат.)**	**Система Б (лат.)**	**прим.**
08	д	d	d	станд.
09	Дз/дз (Зь/зь)	ʒ	Dz/dz	станд.
10	Дъ/дъ	đ	Dh/dh	пр. авт.
11	е	ē	e	станд.
12	ж	ž	Zh/zh	А – ст.
13	з	z	z	станд.
14	и	i	i	станд.
15	й	y	y	станд.
16	к	k	k	станд.
17	л	l	l	станд.
18	м	m	m	станд.
19	н	n	n	станд.
20	о	o	o	станд.
21	Оо/оо	ō	Oo/oo	А – ст.
22	п	p	p	станд.
23	р	r	r	станд.
24	с	s	s	станд.
25	т	t	t	станд.
26	Тъ/тъ	Ť/ť	Th/th	пр. авт.
27	у	u	u	станд.
28	Уу/уу	ū	Uu/uu	А – ст.
29	ф	f	f	станд.
30	х	x	x	станд.
31	Хь/хь	h	h	пр. авт.
32	ц	c	c	станд.
33	ч	č	Ch/ch	станд.
34	ш	š	Sh/sh	станд.
35	ъ	′	′	станд.
36	э	e	e′	пр. авт.
37	ээ	ê	Ee/ee	пр. авт.
38	ғ	ħ	Gh/gh	пр. авт.
39	қ	q	q	станд.
40	ҷ	j	j	станд.
41	ӣ	ī	Ii/ii	пр. авт.
42	ӯ	ů	Uo/uo	Б – пр. авт.

(Пометы «ст.», «станд.» обозначают ГОСТ 7.79-2000 или ISO 9:95; «лат.» — латинские символы, «греч.» — греческие символы, «прим.» — примечание, «пр. авт.» — предложение автора статьи, */* – прописная буква/строчная буква.)

Сначала рассмотрим систему Б, так как она основана на применения стандартной английской раскладки клавиатуры, следовательно, она чаще всего и будет применяться. Заметим, что транслитератор весь текст преобразовывает за один шаг. Так как мы работаем с «неродным» транслиту текстом, то вынуждены его предварительно преобразовать к виду, «понятному» для него. Поэтому ниже опишем подготовительные шаги. Здесь и далее будем опираться на результаты и статистические данные, приведённые в работах [Усманов, Кадамшоев 2009; Бахтоваршоев 2013]. Возьмем за основу алфавит Зарубина-Соколовой, который включает 41 (39) «букву». На данном этапе оставим без комментариев стандартные замены. Замена долгих звуков *аа* (кирил.) на *aa* (лат.) и *уу* (кирил.) на *uu* (лат.) может быть осуществлена автоматически. Следующие буквы исходного шугнанского текста ***й, х, э, г, қ, ҷ, ӣ, ӯ*** должны быть предварительно заменены перед автоматическим осуществлением преобразования. Такое неудобство связано с тем, что в обычном латинском алфавите по сравнению с русским и, тем более, с шугнанским, количество букв меньше (их 26). Букву *х* (кирил.) транслитератор преобразует в *h* (аш) или *kh*, нам же нужна латинская *x* (икс). Каждая пара букв *е* и *э; й* и *и* заменяется транслитератором на одну – *е* (лат.) или *i* (лат.) соответственно. В силу этого, чтобы избежать путаницы, следует заранее заменить одну из них. Предлагается заменить «э» на латинское «е'», а «й» на «у» (игрек). Тогда при последующей транслитерации путаница исключается. «Таджикские» буквы ***г, қ, ҷ, ӣ, ӯ*** транслитератором не распознаются, так как он ориентирован на русский текст. Значит, их также надо предварительно заменить, т. е. произвести замену ғ ~ gh, қ ~ q, ҷ ~ j, ӣ ~ ii, ӯ ~ uo. Итак, после предварительной замены

восьми символов мы получаем заготовку для реализации следующего этапа работы. Подготовленный таким образом текст и подаётся на вход транслитератора.

Самую большую частоту (повторения) в шугнанском языке имеет *а,* равную приблизительно 11,01 %. Поэтому, чтобы получить сведения о представлении шугнанского текста на латинице, следует сравнить суммарную частоту букв, которые после транслитерации будут иметь одинаковую компоненту, с частотой буквы *а*. Следовательно, мы должны вычислить частоты некоторых групп букв, имеющих одинаковую компоненту и результат сравнить с частотой буквы *а*, которую примем равной 100 %. Это группы г, гь, ғ (g); д, дз, дъ (d); е, э, ээ (e); дз, з, ж (z); дъ, ж, тъ, хь, ч, ш, ҳ (h); о, ӯ (o); у, уу, ӯ (u).

Сумма частот букв ***г, гь ғ,*** которые будут иметь общую компоненту g, равна примерно 1,85 %, что меньше частоты буквы *а* почти в 6 раз; и это показывает, что выбор очень удачный. Сумма частот букв ***д, дз, дъ*** равна 7,96 %, которая показывает, что есть ещё запас примерно в 30 % (более точно: 100 — (7,96/11,01)х100 ≈ 28 %). Для следующей группы ***е, э, ээ*** сумма равна 5,32 %, запас почти 50 %. Ещё одна группа ***дз, з, ж*** имеет сумму частот 1,62 %, которая также показывает хороший результат. Сумма частот букв самой многочисленной группы ***дъ, ж, тъ, хь, ч, ш, ҳ,*** которые после перехода будут иметь компоненту *h*, приблизительно равна 5,80 %, что тоже является весьма неплохим результатом; запас больше 45 %. Следующие две группы имеют частоту 5,93 % и 6,58 % соответственно; последняя цифра имеет запас примерно 40 %. Следовательно, преобразованный текст будет иметь примерно ту же структуру, что и исходный, так как ни одна из букв не получила частоту, которая бы в разы превосходила максимальную.

Таким образом, получаем важный количественный результат: шугнанский текст на данном алфавите не будет иметь визуального перекоса по причине безусловного доминирования

каких-либо букв, так как он в целом сохраняет статистические закономерности исходного текста.

При этом, чтобы получить «стандартный» текст, после проведения процесса транслитерации надо произвести еще четыре преобразования. Для этого нужно заменить *d'* на *dh*, *x'* на *h, t'* на *th, v'* на *w*. В результате таких манипуляций шугнанский текст преобразуется с кириллицы в латиницу. Это достигается за двенадцать шагов, восемь (предварительных), до транслитерации, четыре после неё.

Здесь подчеркнём, что закономерности сохраняются для текстов, имеющие «размер» больше некоторого, вполне определённого значения, для шугнанского рассматриваемый текст должен содержать не менее 12,3 тысяч знаков или семь страниц, т. е. текст должен быть репрезентативным [Усманов, Гуломсафдаров 2009]. Если взять какую-то группу, допустим дъ, ж, тъ, хь, ч, ш, ғ (h), и расширить её явно неудачным образом, добавляя, например, аа = ah, х = kh, ӯ = uh, то результат может быть катастрофичным, так как сумма частот таким образом расширенной группы равна 12,67 %, что превосходит частоту буквы «а». Но если текст меньше репрезентативного, то статистические закономерности могут и не сохраняться, следовательно, он с вышеуказанной группой в алфавите случайно может иметь нормальные характеристики.

Конечно, можно и прямо набирать любой текст на предложенном алфавите на основе латиницы при помощи английской клавиатуры. В некоторых словах, звуки которых при наборе (записи) будут совпадать с дифтонгами dh (дъ), dz (дз), uo (уо), стоящие рядом разные буквы следует разделить апострофом, например *pod'ho, bad'zot, sad'zabuon, bu'or, mu'ol*. Это поможет избежать путаницы и произносить слова правильно, ибо новый алфавит неминуемо порождает и новые правила правописания. Следовательно, для данных языков на повестку дня выдвигается разработка общепринятого алфавита и правил правописания.

При этом надо иметь в виду, что по рекомендации ООН за 1987 г. для ***е, х, ц, ю, я*** в географических названиях используются только ***e, h, c, ju, ja*** для передачи русских названий [Транслитерация; Щерба 1940]; правила передачи таджикских географических названий автором не рассматриваются.

Так, применяя систему Б для шугнанского алфавита, получаем следующий результат:

Таблица 3. Транслитерация по системе Б

a (aqq, aql)	aa (aal, aam)	b (boruon, baat)	c (cimuud, ciiw)
ch (chaal, chok)	d (dil, duor)	dh (dhar, dhud)	dz (dzal, dzul)
e (der, ser)	ee (jeeq, teer)	e' (biide', sozde')	f (fay, fiil)
g (goz, gul)	gh (ghob, ghuok)	g' (g'ach, g'ol)	h (hoj, hoy)
i (ilm, isob)	ii (biir, tiir)	j (jaar, juur)	k (kaal, kor)
l (laag, liil)	m (maruob, maruod)	n (nosh, noxuun)	o (osh, olam)
p (puc, palla')	q (qand, qoghaz)	r (rost, rux)	s (sado, sidz)
sh (shiig, shol)	t (tillo, tor)	th (thiir, thow)	u (umr, usma')
uo (buon, uon)	uu (suur, zuur)	v (vadz, vuor)	w (waarg, woh)
x (xaat, xiit)	y (yax, yuuhk)	z (zar, zariidz)	zh (zhaash, zhow)

Заметим, что в шугнанском языке «о» всегда долгое. Напротив, рушанский язык имеет пару, краткое и долгое «о». Тогда в рушанский алфавит помимо краткого «о» мы должны дополнительно включить дограф «оо» [ō]. В то же время в рушанском отсутствует звук [ə]. Таким образом, в таблицу, соответствующую Таблице 3, для рушанского языка следует добавить долгий звук оо = [ō], а символы «ее» и «е’» удалить.

Рассмотрим теперь транслитерацию по системе А. Очевидно, что алфавитом обеспечивается первое требование — однозначность. Остаётся проанализировать частоту групп

букв, имеющих сходное начертание, т. е. отличающихся диакритическими знаками. Это такие группы, как a, ā; c, č; d, đ; e, ê, ē; g, ĝ; h, ħ; i, ī; s, š; t, ť; u, ū, ů; z, ž.

Ясно, что наибольшую частоту имеет группа «a, ā». Естественно, что она превосходит частоту использования «a» (примерно на 30 %). Результат удовлетворительный, так как явного перекоса нет. К тому же, в этой ситуации другие варианты отсутствуют. Другие группы имеют следующие результаты: c, č (2,04%); d, đ (7,84%); e, ê, ē (5,31%); g, ĝ (1,33%); h, ħ (1,70%); i, ī (7,02%); s, š (3,98%); t, ť (5,88%); u, ū, ů (6,58%); z, ž (1,49%). Как показывают эти результаты, максимальное значение имеет примерно 29 % запаса, так как 100 — (7,84/11,01) x100 ≈ 29 %. В таблице для строчной буквы ť помещена также прописная её форма, Ť с кароном, так как для этой буквы её строчная форма отличается от прописной. Таким образом, получаем количественный результат: шугнанский алфавит по предложенной системе не имеет неестественного явного доминирования каких-либо групп букв. Следовательно, он также может использоваться в практической работе в качестве действующего рабочего варианта.

Для рушанского имеем следующий результат: a, ā (17,82 %); c, č (2,04 %); d, đ (8,16 %); g, ĝ (1,68 %); h, ħ (1,65 %); i, ī (7,84 %); o, ō (5,18 %); s, š (3,40 %); t, ť (5,62 %); u, ū, ů (6,06 %); z, ž (2,41 %). Для первой группы частота больше на 21 %, и её величина ещё меньше, чем для шугнанского. Для остальных групп максимальное значение не превосходит 55 %, которое показывает, что не имеет места явное преобладание каких-то конкретных групп. Следовательно, Таблицу 2 можно использовать и для рушанского языка, заменив в нём букву ē на e, и удалив из неё ee, e'.

В качестве простого транслитератора можно использовать, например, программу Цифирица 2.2 (Cifirica 2.2) Р.В. Кошелева [Кошелев].

Литература

Бахтоваршоев А.Ш. Статистическое распределение частот букв одного варианта алфавита шугнанского языка // Доклады АН РТ, 2013. Т. 56, № 7. С. 531–533.

Додыхудоева Л.Р. Шугнанский язык // Языки Российской Федерации и соседних государств. Москва, 2005. Т. 3. С. 435–446.

Евангелие от Луки — Luqo Inǰīl. Xuγˇnůni ziv qissayen / mutarǰim R.Kh. Dodykhudoev. Maskav: Instituti tarǰumā Kitobi Muqadas, 2001a. (Евангелие от Луки (на шугнанском языке). Отрывки). Пер. *Р.Х. Додыхудоева*. Москва, 2001).

Евангелие от Луки — Luqo Inǰīl. Maskav: Instituti tarǰumā Kitobi Muqadas, 2001б. (Евангелие от Луки (на рушанском языке). Отрывки). Москва, 2001.

Зарубин И.И. Шугнанские тексты и словарь. Москва-Ленинград, 1960.

Карамшоев Д.К., Аламшоев М. Алифбо: Китоби дарсӣ барои соли аввали таҳсил ба забони шуғнонӣ (Букварь. Учебник шугнанского языка для первого года обучения) (на тадж. языке). Душанбе, 1996.

Кошелев Р.В. Цифирица 2.2 (Cifirica 2.2). Сайт: http://www.bestfree.ru/ (Последняя дата обращения: 24.01.2017). Автор предоставляет возможность безвозмездного использования программы.

Пахалина Т.Н. Памирские языки. Москва, 1969.

Соколова В.С. Шугнано-рушанская языковая группа // Языки народов СССР. Т. I. Индоевропейские языки. Москва, 1966. С. 362–397.

Транслитерация русского алфавита латиницей: http://ru.wikipedia.org/wiki/Транслитерация_русского_алфавита_латиницей (Последняя дата обращения: 24.01.2017).

Усманов З.Д., Гуломсафдаров А.Г. Статистическое распределение частот встречаемости букв в шугнанском языке // Доклады АН РТ, 2009. Т. 52, № 3. С. 187–191.

Усманов З.Д., Кадамшоев Н.У. Статистическое распределение частот встречаемости букв в рушанском языке // Доклады АН РТ, 2009. Т. 52, № 2. С. 106–110.

Щерба Л.В. Транслитерация латинскими буквами русских фамилий и географических названий // Известия АН СССР. Отд. литературы и языка. Москва, 1940. Т. I, № 3. С. 118–126.

Эдельман Д.И. О единой научной транскрипции для иранских языков. Москва-Ленинград, 1963.

Языки мира: Иранские языки. III. Восточноиранские языки. Москва, 1999.

Алигавхар Шохайдарович Бахтоваршоев
Независимый исследователь
Aligawhar Shokhaidarovich Bakhtovarshoev
Independent scholar
aligawhar07@mail.ru