Checked for validity and accuracy – November 2017

ROMANIZATION OF PASHTO

BGN/PCGN 1968 System, 2017 Revision

Pashto is an Indo-Iranian language and is one of two nationally official languages in Afghanistan and one of five regionally recognised languages in Pakistan. The romanization system presented here may be applied to all Pashto geographical names. Although the BGN/PCGN policy for geographical names in Afghanistan is to apply the *BGN/PCGN national system of romanization for Afghanistan* (2007), which incorporates Dari elements, when applied to a Pashto geographical name, the romanized results of the BGN/PCGN national system for Afghanistan are the same as those of this Pashto romanization system[1].

The Pashto alphabet uses a modified form of the Perso-Arabic script, and contains twelve additional consonants not present in standard Arabic, as well as three additional vowel characters and an additional vowel point.

Consonants: ټ ډ ړ ږ ښ ګ ڼ ځ څ ځ ځ ځ

Vowels: ۍ ې ے; Vowel Point: ◌ٗ

The points used in Arabic to mark short vowels and certain other diacritical marks are not written in Pashto. Consequently, a reference source may sometimes be required to aid correct identification of the standard spellings and proper vowels and elimination of dialectal and idiosyncratic variations. In the interests of clarity, a column showing vowel pointing from Arabic to indicate short vowels has been included in the examples below, alongside the unpointed form that will usually be encountered. However it should be noted that the pronunciation of short vowels will vary.

(*Note: it is recommended that a font such as Scheherazade, available from www.sil.org, which includes the Unicode extended Arabic sub-range, be used to view this system*[2].)

[1] The two systems are designed to be complementary. The national system for Afghanistan was developed to be broader in scope in order to apply to geographical names in Afghanistan regardless of their language origin; however, when applied to a Pashto name the national system equally conveys the particularities of Pashto captured by this language-specific system.

[2] Please note that the identification of a particular font does not represent an endorsement of any specific product or manufacturer.

Table 1: Consonant Characters

	Script				Unicode codepoint (Independent)	Romanization	Roman Unicode codepoint (lower case)	Example		
	Final	Medial	Initial	Independent				Unpointed script	Pointed script	Roman script
1	ا		ا		0627	*See note 1*	-	*See note 1*		
2	ب	ـبـ	بـ	ب	0628	b	0062	بغلان	بَغلان	Baghlān
3	پ	ـپـ	پـ	پ	067E	p	0070	پوټکی	پوټَکَی	Pōṯakay
4	ت	ـتـ	تـ	ت	062A	t	0074	شیرین تگاب	شیرین تَگاب	Shīrīn Tagāb
5	*characters not available in Unicode*[3]			ټ	067C	ṯ	1E6F	پښتون کوټ	پَښتون کوټ	Paṯhtūn Kōṯ
6	ث	ـثـ	ثـ	ث	062B	s̄	0073+0304	ثابر	ثابِر	Sābir
7	ج	ـجـ	جـ	ج	062C	j	006A	جلال آباد	جَلال آباد	Jalālābād
8	چ	ـچـ	چـ	چ	0686	ch	0063+0068	چاریکار	چاریکار	Chārīkār
9	ځ	ـځـ	ځـ	ځ [4]	0681	dz[Note2]	0064+007A	ځدران	ځَدرانِ	Dzadrāṉ
10	څ	ـڅـ	څـ	څ	0685	ts[Note2]	0074+0073	څوکی	څَوکی	Tsowkêy
11	ح	ـحـ	حـ	ح	062D	ḥ	1E29	حضرت إمام	حَضرَت إمَام	Ḥaẓrat-e Imām
12	خ	ـخـ	خـ	خ	062E	kh	006B+0068	خوست	خوست	Khōst
13	د		د		062F	d	0064	سپین بولدك	سپین بولدَك	Spīn Bōldak
14	*character not available in Unicode*[3]		ډ		0689	ḍ	1E0F	ډند و پتان	ډَند وَ پَتان	Ḍanḍ wa Patān
15	ذ		ذ		0630	z̄	007A+0304	گذرگاۀ نور	گُذَرگاۀ نور	Guẓargāh-e Nūr
16	ر		ر		0631	r	0072	کندهار	کَنْدَهار	Kandahār
17	*character not available in Unicode*[3]		ړ		0693	ṛ	1E5F	أندړ	أَنْدَړ	Andaṛ
18	ز		ز		0632	z	007A	کندز	کُندُز	Kunduz
19	ژ		ژ		0698	zh	007A+0068	میر أسلم ژرنده	میر أَسلَم ژرَندَه	Mīr Aslam Zhrandah

[3] These characters are not available with a single Unicode codepoint, so cannot be displayed here. When typing, the independent character's codepoint will automatically display with the appropriate word-medial or word-final form where so appearing in a word.

[4] The variant form ځ is seen infrequently, and does not have a single Unicode codepoint.

	Script				Unicode codepoint (Independent)	Romanization	Roman Unicode codepoint (lower case)	Example		
								Unpointed script	Pointed script	Roman script
20	*character not available in Unicode*[3]		ﮊ		0696	zh	007A+035F +0068	ږيره	ږيرَه	Zẖīrah
21	س	سـ	ـسـ	س	0633	s	0073	سمنگان	سَمَنگان	Samangān
22	ش	شـ	ـشـ	ش	0634	sh	0073+0068	مزار شريف	مَزارِ شَريف	Mazār-e Sharīf
23	*characters not available in Unicode*[3]			ښ	069A	sh	0073+035F +0068	کښته کلا	کښتَه کَلا	Kshêtah Kalā
24	ص	صـ	ـصـ	ص	0635	ṣ	015F	قيصار	قَيصار	Qayṣār
25	ض	ضـ	ـضـ	ض	0636	ẓ	1E95	فيض آباد	فَيض آباد	Faīẓābād[Note 5]
26	ط	طـ	ـطـ	ط	0637	ṭ	0163	حضرت سلطان	حَضرَتِ سُلطان	Ḥaẓrat-e Sulṭān
27	ظ	ظـ	ـظـ	ظ	0638	ẓ	007A+0327	ظاهر کلا	ظاهِر کَلا	Ẓāhir Kalā
28	ع	عـ	ـعـ	ع	0639	‘	2018	پل علم	پُلِ عَلَم	Pul-e ‘Alam
29	غ	غـ	ـغـ	غ	063A	gh	0067+0068	غزني	غَزني	Ghaznī
30	ف	فـ	ـفـ	ف	0641	f	0066	مزار شريف	مَزارِ شَريف	Mazār-e Sharīf
31	ق	قـ	ـقـ	ق	0642	q	0071	قيصار	قَيصار	Qayṣār
32	ك	كـ	ـكـ	ک	06A9	k	006B	کندهار	کَندَهار	Kandahār
33	گ	گـ	ـگـ	گ, گ	06AF	g	0067	گردېز	گَردېز	Gardēz
34	ل	لـ	ـلـ	ل	0644	l	006C	کابل	کابُل	Kābul
35	م	مـ	ـمـ	م	0645	m	006D	ميمنه	مَيمَنَه	Maīmanah
36	ن	نـ	ـنـ	ن	0646	n	006E	خان آباد	خان آباد	Khānābād
37	*characters not available in Unicode*[3]			ڼ	06BC	n̲	1E49	مانۍ	مانۍ	Mān̲êy
38	و		ـو	و	0648	w	0077	واخان	واخان	Wākhān
39	ه	ھ, ہ	ـھـ	ه	0647	h	0068	کندهار	کَندَهار	Kandahār
40	ى, ي	ۍ	ېـ	ى, ي	0649, 064A	y	0079	ينگي قلعه	يَنگي قَلعَه	Yangī Qal‘ah

Table 2: Vowel, diphthong and diacritical characters

	Script (independent form)	Unicode codepoint	Romanization	Roman Unicode codepoint (lower case)	Example		
					Unpointed script	Pointed script	Roman script
1	◌َ	064E	a	0061	جلال آباد	جَلال آباد	Jalālābād
2	ا	0627	ā see note 1	0101	كابل	كابُل	Kābul
3	آ	0622	ā see note 1	0101	آب بند	آب بَند	Āb Band
4	◌ِ	0650	i	0069	پل حصار	پُلِ حِصار	Pul-e Ḥişār
5	ى (ي)	06CC (064A)	ī	012B	غزني	غَزني	Ghaznī
6	ې	06D0	ē	0113	گردېز	گَردېز	Gardēz
7	◌َى	06CC	ay, aī	0061+0079, 0061+012B	ميوند ميدان شهر	مَيوَند مَيدان شَهر	Maywand, Maīdān Shahr
8	◌َو	0648	ow	006F+0077	جوزجان	جَوزجان	Jowzjān
9	◌َ	0659	ê	00EA	گردون	گَردون	Gêrdōn
10	ى	06CD	êy	00EA+0079	څوکۍ	څَوکۍ	Tsowkêy
11	◌ُ	064F	u	0075	كابل	كابُل	Kābul
12	و	0648	ō, ū	014D, 016B	سپين بولدك بالا بلوك	سپين بولدَك بالا بُلوك	Spīn Bōldak Bālā Bulūk
13	ء	0621	’ see note 4	2019	هوائی ډګر	هَوائی ډَګر	Hawā’ī Ḏagar
	ء as izāfah	0674	-e, -ye see note 9		قلعه نو	قَلعَهٔ نَو	Qal‘ah-ye Now
14	ئ	064A+0654	êy	00EA+0079	-[5]	-	-
15	ىٰ	0649+0670	á	00E1	موسى خېل	موسیٰ خېل	Mūsá Khēl

Numerals									
۰	۱	۲	۳	۴	۵	۶	۷	۸	۹
0	1	2	3	4	5	6	7	8	9

Although Perso-Arabic script is written from right to left, numerical expressions, e.g. ۱۹۶۸ → 1968, are written from left to right.

[5] Only occurs at the end of verbs, so unlikely to be encountered in geographical names.

NOTES

1. *Alif* (ا) should be romanized as follows:

a. Initially, it indicates that the word begins with a vowel or diphthong; the *alif* itself is not romanized, but rather the short vowel it "carries" is romanized; e.g., میر أسلَم ژرَندَه → Mīr Aslam Zhrandah

b. When it carries a *maddah* (آ) (see vowel table, row 3), it represents *ā*; e.g., آب بَند → Āb Band.

c. Medially and finally it represents *ā* (see table 2, row 2); e.g., مانۍ → Māṉêy

d. Medially and finally in words of Arabic origin, *alif* may serve as the bearer of *hamzah*, e.g. رأس → *ra's*. See also note 4.

2. The characters tsē (څ) and dzē (ځ) may be romanized t͡s and d͡z (the combining double breve (Unicode 0361) appearing over the digraph) when for special reasons it is desired that confusion be avoided between ت (t) plus س (s) and between د (d) plus ز (z), respectively.

3. Occasionally the character sequences کھ, زھ, سھ, and گھ occur. They may be romanized *k·h, z·h, s·h,* and *g·h* in order to differentiate these romanizations from the digraphs *kh*, *zh*, *sh*, and *gh*, which are used to represent the characters خ, ژ, ش, and غ respectively.

4. *Hamzah* (ء) should be romanized as follows:

a. In word-initial position, where it will appear either above or below *alif* (أ or إ), it indicates a short vowel and should not itself be romanized. In other positions it should be romanized by an apostrophe, e.g. جُزء → *juz'*.

b. *Yeh* with *hamzah* (ئ) should be romanized *ê*y, unless it represents the compound (*iẓāfah*) morpheme, in which case it is romanized according to note 9 below.

5. The division of words utilized in Pashto writing is followed in romanization, except that the elements *–ābād*, *-khwā*, *-shahr*, *-zādah*, *-zay* and *-ullāh* are always romanized as part of the preceding word, e.g. رَحْمَت آباد → Raḥmatābād and رَحْمَت الله → Raḥmatullāh. However, when the word for God (الله) appears as a standalone word it should be written Allāh. Note also the "dagger alif" (ٰ) above the second ل (*lām*) in the word الله; this, like the short vowels, is not written in Pashto but should be romanized *ā*, like a full-size alif. Persian derivational endings such as *–vand* and endings of Turkish origin such as *–lar*, *-lī*, *-lū*, *-i*, *-u*, *-si*, and *–su*, should be written together with the preceding word.

6. The Pashto preposition د should be romanized *dê* in agreement with its pronunciation, despite the fact that it is sometimes pointed with *kasrah* (ِ).

7. In names of Arabic origin, the *l* of the definite article *al*/*ul* is assimilated before the 'sun letters' *t*, *s̄*, *d*, *z̄*, *r*, *z*, *s*, *sh*, *ṣ*, *ẓ*, *ṭ*, *ẓ*, *l* and *n*. In romanization, the article will be written *al* or its assimilated equivalent in name-initial position but *ul* or its assimilated equivalent elsewhere; the article should be separated from the name it precedes and should not be capitalized, except at the beginning of a name, e.g. جَبَل السَراج → *Jabal us Sarāj*.

8. In Arabic names, a *shaddah,* ّ is used to denote the doubling of a particular consonant character, e.g. مُحَمَّد → *Muḥammad*. However, in Pashto this 'doubling' is frequently omitted in both Perso-Arabic script and the resulting romanization. Guidance on doubling may be taken from an authoritative names source, such as an Afghan government source or Pashto dictionary; for example, it is usual to see *Ḫājī* without and *'Abbās* with the doubled consonant. The doubled *y* consonant is almost always retained, as in *Sayyid* or *Qayyūm*.

9. The *iẓāfah* morpheme is not a grammatical feature of Pashto and, if encountered in a linguistically hybrid geographical name (i.e. combining features of both Pashto and Dari), it should be treated according to the BGN/PCGN national system of romanization for Afghanistan, 2007, as *-e*, unless the preceding word ends with a silent heh (ه) or a vowel when it should be shown *-ye*, e.g. غرِ حِصار → *Ghar-e Ḫiṣār*; قَلعَهٔ نَو → *Qal'ah-ye Now*.

10. The character sequence خو, when followed by ا or ى, should be romanized *khw*, although the *w* is either not pronounced, or only weakly pronounced; e.g. خواجه → *khwājah*.

11. An inventory of letter-diacritic combinations in addition to the unmodified letters of the basic Roman script is:

‘ (U+2018)	’ (U+2019)
Ā (U+0100)	ā (U+0101)
Á (U+00C1)	á (U+00E1)
D̲ (U+0044+0031)	d̲ (U+0064+00031)
Ē (U+0112)	ē (U+0113)
Ê (U+00CA)	ê (U+00EA)
Ḥ (U+1E28)	ḥ (U+1E29)
Ī (U+012A)	ī (U+012B)
N̈ (U+004E+0304)	n̈ (U+004E+0304)
Ō (U+014C)	ō (U+014D)
R̲ (U+0052+0031)	r̲ (U+0072+0031)
Ş (U+015E)	ş (U+015F)
S̄ (U+0053+0304)	s̄ (U+0073+0304)
T̲ (U+0054+0031)	t̲ (U+0074+0031)
Ţ (U+0162)	ţ (U+0163)
Ū (U+016A)	ū (U+016B)
Ẓ (U+005A+0327)	ẓ (U+007A+0327)
Z̄ (U+005A+0304)	z̄ (U+007A+0304)
Z̲ (U+005A+0331)	z̲ (U+007A+0331)
Z̲H (U+005A+0048+035F)	z̲h (U+007A+0068+035F)

12. The Romanization columns show only lowercase forms but, when romanizing, uppercase and lowercase Roman letters as appropriate should be used.