

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 system¹.

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: ن گ بن ژ ړ ږ ډ ځ خ چ ټ پ

Vowels: ۱ ۲ ۳; Vowel Point: ۰

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 system².)

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	<i>See note 1</i>	-	<i>See note 1</i>		
2	ب	ب	ب	ب	0628	b	0062	بغلان	بَغْلان	Baghlān
3	پ	پ	پ	پ	067E	p	0070	پوتکی	پوٹکی	Pōtakay
4	ت	ت	ت	ت	062A	t	0074	شیرین تگاب	شیرین تَغاب	Shīrīn Tagāb
5	characters not available in Unicode ³			ت	067C	ṭ	1E6F	پښتون کوټ	پښتون کَوټ	Pashtūn Kōt
6	ث	ث	ث	ث	062B	ṯ	0073+0304	ثابر	ثَابِر	Sābir
7	ج	ج	ج	ج	062C	j	006A	جلال آباد	جَلال آباد	Jalālābād
8	چ	چ	چ	چ	0686	ch	0063+0068	چاریکار	چَارِیکار	Chārīkāṛ
9	ح	ح	ح	ح ⁴	0681	dz ^{Note2}	0064+007A	خدران	خَدْران	Dzadrān
10	خ	خ	خ	خ	0685	ts ^{Note2}	0074+0073	خوکی	خَوکی	Tsowkēy
11	ح	ح	ح	ح	062D	ḥ	1E29	حضرت امام	حَضْرَتِ اِمَام	Ḥazrat-e Imām
12	خ	خ	خ	خ	062E	kh	006B+0068	خوست	خوست	Khōst
13	د		د		062F	d	0064	سپین بولدک	سپین بَوْلَدک	Spīn Bōldak
14	character not available in Unicode ³			د	0689	ḍ	1E0F	دند و پتان	دَنْد و پَتان	Dand wa Patān
15	ذ		ذ		0630	z̤	007A+0304	گذرگاه نور	گُذَرگاه نور	Gužargāh-e Nūr
16	ر		ر		0631	r	0072	کندهار	کَنْدَهَار	Kandahār
17	character not available in Unicode ³			ر	0693	ṛ	1E5F	آندر	اَنْدَر	Andar
18	ز		ز		0632	z	007A	کندز	کُنْدز	Kunduz
19	ژ		ژ		0698	zh	007A+0068	میر اسلم ژرنده	میر اَسْلَم ژَرَنده	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 ح 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			ږ	0696	zh	007A+035F+0068	بریره	بِرِرَہ	Zhīrah
21	س	س	س	س	0633	s	0073	سمنگان	سَمَنگان	Samangān
22	ش	ش	ش	ش	0634	sh	0073+0068	مزار شریف	مَزارِ شَرِیف	Mazār-e Sharīf
23	characters not available in Unicode ³			ښ	069A	sh	0073+035F+0068	کښته کلا	کښتَہ کلا	Kshêtah Kalā
24	ص	ص	ص	ص	0635	ṣ	015F	قیصار	قَیصار	Qayṣār
25	ض	ض	ض	ض	0636	ẓ	1E95	فیض آباد	فَیض آباد	Faīẓābād ^{Note 5}
26	ط	ط	ط	ط	0637	ṭ	0163	حضرت سلطان	حَضَرَتِ سُلطان	Ḥazrat-e Sulṭān
27	ظ	ظ	ظ	ظ	0638	ẓ	007A+0327	ظاهر کلا	ظَاهِر کَلا	Zāhir Kalā
28	ع	ع	ع	ع	0639	‘	2018	پل علم	پُلِ عَلم	Pul-e ‘Alam
29	غ	غ	غ	غ	063A	gh	0067+0068	غزني	غَزَنِي	Ghaznī
30	ف	ف	ف	ف	0641	f	0066	مزار شریف	مَزارِ شَرِیف	Mazār-e Sharīf
31	ق	ق	ق	ق	0642	q	0071	قیصار	قَیصار	Qayṣār
32	ک	ک	ک	ک	06A9	k	006B	کندهار	کَنَدَهار	Kandahār
33	گ	گ	گ	گ, گ	06AF	g	0067	گردېز	گَرْدَېز	Gardēz
34	ل	ل	ل	ل	0644	l	006C	کابل	کَابُل	Kābul
35	م	م	م	م	0645	m	006D	میمنه	مَیْمَنَہ	Maīmanah
36	ن	ن	ن	ن	0646	n	006E	خان آباد	خان آباد	Khānābād
37	characters not available in Unicode ³			ڼ	06BC	ṇ	1E49	مانی	مانِی	Māṇēy
38	و			و	0648	w	0077	واخان	واخان	Wākhān
39	ه	ه	ه	ه	0647	h	0068	کندهار	کَنَدَهار	Kandahār
40	ي, ی	ی	ی	ي, ی	0649, 064A	y	0079	ینگی قلعه	یَنگِی قَلْعَہ	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	اَ	064E	a	0061	جلال آباد	جَلال آباد	Jalālābād
2	اِ	0627	ā see note 1	0101	کابل	کَابُل	Kābul
3	آ	0622	ā see note 1	0101	آب بند	آب بَند	Āb Band
4	اِی	0650	i	0069	پل حصار	پُلِ حِصار	Pul-e Ḥiṣār
5	ی (ي)	06CC (064A)	ī	012B	غزني	غَزني	Ghaznī
6	ې	06D0	ē	0113	گردېز	گَرْدېز	Gardēz
7	اِی	06CC	ay, aī	0061+0079, 0061+012B	میوند میدان شهر	مِیَوَند مَیدان شَهر	Maywand, Maīdān Shahr
8	او	0648	ow	006F+0077	جوزجان	جَوَزجان	Jowzjān
9	و	0659	ê	00EA	گردون	گَرْدون	Gêrdôn
10	ی	06CD	êy	00EA+0079	خوکی	خَوکی	Tsowkêy
11	و	064F	u	0075	کابل	کَابُل	Kābul
12	و	0648	ō, ū	014D, 016B	سپين بولدك بالا بلوك	سپين بولَدك بالا بُلوك	Spīn Bōldak Bālā Bulūk
13	ء	0621	ʾ see note 4	2019	هوائى ډگر	هَوائى ډَگر	Hawāʾī Ḍagar
	ء as izāfah	0674	-e, -ye ^{see note 9}		قلعه نو	قَلْعَه نَو	Qalʾah-ye Now
14	ئ	064A+0654	êy	00EA+0079	- ⁵	-	-
15	ى	0649+0670	á	00E1	موسى خېل	موسى خَېل	Mūsá Khēl

Numerals									
٠	١	٢	٣	٤	٥	٦	٧	٨	٩
0	1	2	3	4	5	6	7	8	9

Although Perso-Arabic script is written from right to left, numerical expressions, e.g. ١٩٦٨ → 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:

- 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., *میرِ اَسْلَمَ ژرَنده* → *Mīr Aslam Zhrandah*
- When it carries a *maddah* (آ) (see vowel table, row 3), it represents *ā*; e.g., *آب بند* → *Āb Band*.
- Medially and finally it represents *ā* (see table 2, row 2); e.g., *مانی* → *Mānêy*
- 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* (ع) and *dzē* (غ) may be romanized *ṭs* and *ḏz* (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 *که*, *زه*, *سه*, and *گه* 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:

- 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’*.
- Yeh* with *hamzah* (ئ) should be romanized *êy*, unless it represents the compound (*izāfah*) 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” (اِ) 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*, *ṯ*, *d*, *ḏ*, *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 السَّرَاج* → *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 *Ḥājī* without and *‘Abbās* with the doubled consonant. The doubled *y* consonant is almost always retained, as in *Sayyid* or *Qayyūm*.

9. The *izāfah* 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. غر حصار → *Ghar-e Ḥiṣār*; قلعه نو → *Qal‘ah-ye Now*.

10. The character sequence خو, when followed by ا or ی, 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)
Ď (U+0044+0031)	ď (U+0064+00031)
Ě (U+0112)	ě (U+0113)
Ê (U+00CA)	ê (U+00EA)
Ħ (U+1E28)	ħ (U+1E29)
Ī (U+012A)	ī (U+012B)
Ņ (U+004E+0304)	ņ (U+004E+0304)
Ō (U+014C)	ō (U+014D)
Ŕ (U+0052+0031)	ŗ (U+0072+0031)
Ş (U+015E)	ş (U+015F)
Š (U+0053+0304)	š (U+0073+0304)
Ţ (U+0054+0031)	ţ (U+0074+0031)
Ț (U+0162)	ț (U+0163)
Ū (U+016A)	ū (U+016B)
Ẑ (U+005A+0327)	ẑ (U+007A+0327)
Ẓ (U+005A+0304)	ẓ (U+007A+0304)
Ẕ (U+005A+0331)	ẕ (U+007A+0331)
ẐH (U+005A+0048+035F)	ẑh (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.