

Preliminary proposal for Turkestan Brāhmī within the Brāhmī block

11 September 2026

Francesco Catoni, catonifdev, gmail.com

1. Introduction

North Turkestan Brāhmī (NTB) and South Turkestan Brāhmī (STB) are two closely related variants of the Brāhmī script used between the 5th and 11th century in Turkestan (modern Xinjiang), to write, chiefly in Buddhist contexts, in various languages. Aside from Sanskrit, STB was used to write primarily in Khotanese, as well as rarely Middle Chinese and Old Turkic,¹ while NTB was used to write Tocharian A, Tocharian B, Tumshukese, Uigur and less commonly Sogdian. The use of Brāhmī for all these languages, besides Sanskrit, lacks sufficient Unicode support.²

Another Turkestan Brāhmī variant is that of Rouran, the earliest attested Mongolic language. As the research in the field is still premature, this proposal does not ask for Rouran-specific codepoints. However, attestations from the Khüis Tolgoi inscription will occasionally be mentioned throughout the proposal whenever they can be of further evidence to the signs present in the other languages.

1.1. Abbreviations

BCJ = Brāhmī conjunct joiner
 BSS = Brāhmī *Sonderzeichen* selector
 Chin. = Middle Chinese (ltc)
 Khot. = Khotanese (kho)
 NTB = North Turkestan Brāhmī
 Skr. = Sanskrit (sa)
 Sogd. = Sogdian (sog)
 STB = South Turkestan Brāhmī
 TochA = Tocharian A (xto)
 TochB = Tocharian B (txb)
 Tumsh. = Tumshukese, a.k.a. Tumshuqese (xtq)
 Turk. = Old Turkic from Khotan (oui)
 Uig. = Uigur, a.k.a. Old Uyghur (oui)

1.2. Other conventions used

Throughout the document each new character or functionality proposed is accompanied by images of a conjunct (*akṣara*) that either contains it or is relevant to the section. The images

1. Emmerick & Róna-Tas (1992) also mention Tibetan.

2. As a clarification, note that Sogdian Brāhmī and Uigur Brāhmī should not be confused respectively with the Sogdian script (Sogd), of Syriac origin, and the Uigur script (Ougr), its derivation, both already supported by Unicode.

are accompanied by their identifier, often a shelfmark according to the established abbreviations, followed by the source from which the scan image was taken, denoted with a copyright symbol ©.

None of the transliterations given were our original research: Tocharian A and Tocharian B are from CEToM, Uigur is from von Gabain (1954) and Maue (1996 and 2015), Tumshukese is from Maue (2009), Sogdian is from Maue & Sims-Williams (2024), Khotanese is from Bailey (1951 and 1969) and Skjærvø (2015), Middle Chinese is from Emmerick & Pulleyblank (1993), Old Turkic from Khotan is from Emmerick & Róna-Tas (1992). All visible viramas, regardless of their shape, are transliterated as <.>.

After the transliteration and transcription there is always the sequence of codepoints by which the example in question is proposed to be encoded. To make the proposed encoding easier to read, the term BRAHMI is omitted from the codepoint names, and the sequence relevant to the section is highlighted either in **GREEN** for newly proposed codepoints or in **YELLOW** if the codepoints are already encoded or proposed in another section. Sequences in **RED** describe encoding advised against.

2. Treatment

There have been previous proposals to provide encoding support for these languages both within the Brāhmī block (Baums & Glass 2003: L2/03/249-R), as a separate “Turkestani” block (Wilson 2014: L2/14-192) and as language-specific Unicode blocks (Wilson 2015: L2/15-022, L2/15-023, L2/15-236).

In our opinion the points presented by Wilson do not meet the requirements for encoding these scripts outside of the Brāhmī block (U+11000–1107F), which in light of the many languages and styles they were written in appears to be the most sustainable solution, leaving the visual discrepancy up to the font stage. It is noted that the concept of a “Tocharian script”, as the proposals by Wilson describes it, is entirely absent from academia,³ where the script is only ever called “Brāhmī”. It is only when there is a need to differentiate it from other variants of Brāhmī that the appellations “North Turkestan Brāhmī” and “South Turkestan Brāhmī” are used, and only when there is a need to differentiate the different uses of the script between different languages that terminology like “Tocharian Brāhmī”, “Uigur Brāhmī”, etc. is employed. It is therefore evident that the script, despite its differences both in appearance and functionalities from the Gupta script⁴ it derives from, is conceptually perceived as Brāhmī, and technical limitations should be the only compelling arguments for creating specific blocks for the two variants treated here.

The proposal will therefore only focus on the missing characters and on the features of the script which raise technical difficulties within the Brāhmī block, and will not go over the appearance of the various signs and sequences which already have a straightforward encoding within the Indian Brāhmī system, for which we defer to Wilson’s proposals.⁵ We nonetheless hope to have made it clear throughout the proposal which features are present in North and which in South Turkestan Brāhmī, with a summarised table at the end of the proposal, as to make this document easily usable even were the multiple blocks solution chosen.

3. Or at least from the works consulted, for which see the bibliography.

4. Also known as “Late Brāhmī”.

5. Our only correction to it is regarding the claim that the “vowels VOCALIC L and VOCALIC LL are not attested in any Tocharian texts”, for which see § 8.3. The rare glyphs for the numerals 10 000 and 100 000 are also missing, though they are not covered here either as we believe they are best treated in a separate proposal.

3. Tocharian special signs

Among the technical difficulties encountered, the hardest to ignore is that these variants of Brāhmī made use of special characters for distinct phonemes which could not be well-represented by any of the Indian signs.

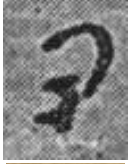
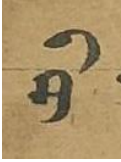
3.1. <ä>

This vowel is known as the *Fremdvokal*. When after a consonant, it appears on top of the conjunct and resembles a trema, while its independent shape is like that of independent <a> with that same diacritic on top. Common to STB, NTB and even Rouran,⁶ it likely derives from the visarga of the Kharoṣṭhi variant used for Khotan Prakrit,⁷ and it is found in all the languages treated here. In some contexts, in STB, the two dots can appear vertically on top of one another, and in later manuscripts they join together, eventually becoming hardly distinguishable from <i> in Late Khotanese Cursive texts. Following Wilson, the name proposed is VOWEL SIGN AE for the diacritic and LETTER AE for the independent vowel.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochA	A 273 recto, l. 4 © IDP		ä	ä	LETTER AE
TochB	IOL Toch 127 recto, l. 3 © IDP		yä	yä	LETTER YA + VOWEL SIGN AE
Tumsh.	TS 21 verso, l. 1 © DTA		ndä	—	LETTER NA + VIRAMA + LETTER DA + VOWEL SIGN AE
Uig.	U 6828 recto, l. 2 © DTA		yyo-hk ^h ä	yok	LETTER YA + VIRAMA + LETTER YA + VOWEL SIGN O + BAR VIRAMA + LETTER HA + VIRAMA + LETTER KHA + VOWEL SIGN AE
Khot.	IOL Khot 75/3 recto, l. 1 © IDP		ḍä	ḍä	LETTER DDA + VOWEL SIGN AE

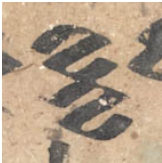
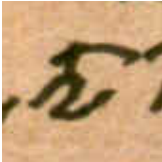
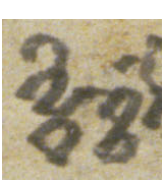
6. Maue 2017, § 2.3.1, pp. 11f.

7. See Peyrot (2025) for further discussion as well as other possibilities.

Khot.	Or 6393/2, l. 4 © IDP		ysä	ysä	LETTER YA + VIRAMA + LETTER SA + VOWEL SIGN AE
Chin.	IOL Khot S 7, l. 91 © IDP apud Sims- Williams 2001		rä	— ⁸	LETTER RA + VOWEL SIGN AE
Turk.	P 2892 verso, l. 166 © BnF		śä	š	LETTER SHA + VOWEL SIGN AE

3.2. <w>

The consonant traditionally transliterated as <w> in Tocharian is believed to have sounded as IPA [ʍ].⁹ We propose the name LETTER WA.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochA	A 277 recto, l. 6 © IDP		wi	wi	LETTER WA + VOWEL SIGN I
TochB	IOL Toch 47 verso, l. 5 © IDP		wwa	wwa	LETTER WA + VIRAMA + LETTER WA
Tumsh.	TS 25, l. 4 © DTA		rwā	—	LETTER RA + VIRAMA + LETTER WA
Uig.	U 6865 recto, l. 2 © DTA		wa	va	LETTER WA
Sogd.	bs 2 verso, l. 3 © DTA		wd ^{hi} -k	vδīk	LETTER WA + VIRAMA + LETTER DHA + VOWEL SIGN I + BAR VIRAMA + LETTER KA + BSS + VIRAMA

3.3. The *Sonderzeichen*

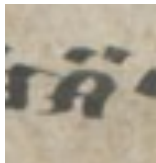
The label *Fremdzeichen*, or “special sign”, can generally be used to refer to any letter found in these languages not found in the original Indian Brāhmī. The ones that stand for a separate

8. The sequence <śve rä> stood for 說.

consonant value, like <w> above and the Konow signs below, have no choice but to be encoded as independent codepoints. However, there is a class of special signs introduced in Tocharian, which in order to differentiate them from the former will be called *Sonderzeichen* here instead,¹⁰ that are not used to describe new consonant values, but rather as equivalents of their base letters implying either the <ä> vowel or no vowel at all in the second conjunct of a bar virāma (§ 4.1). The former function of these signs was rare in the scripts derived from Tocharian Brāhmī, while the latter survived very well into Uigur and Sogdian.¹¹

While both Baums & Glass and Wilson proposed to encode these signs atomically as 11 new codepoints, the suggestion given here is to treat them as combinations of their base forms followed by a non-spacing mark, SONDERZEICHEN SELECTOR, henceforth abbreviated as BSS. This conclusion is reached on the grounds that these letters do not differ from their base forms in their consonantal phoneme, but rather on its implied vowel, so encoding them as two characters appears closer to the usual Unicode approach to Indic scripts. The increase of difficulty in the font implementation this approach would entail appears minimal in comparison to the already very complicated behaviour of the script. Nevertheless, since this is a technical matter, the question is left open for the SEW to give insight for in case an atomic approach would be preferable.

It could be argued that the codepoint VOWEL SIGN AE already serves the purpose of this proposed selector, since in theory the two should never contrast, and are semantically equivalent. Contrary to this solution is the fact that in practice they do contrast even just within Tocharian. In scholarly transcriptions these signs are transcribed distinctly by underlining the consonant and the following <a> whenever present.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochB	PK AS 5C recto, l. 2 © BnF		śä	śä	LETTER SHA + VOWEL SIGN AE
TochB	PK AS 6 verso, l. 7 © BnF		śä	śä	LETTER SHA + BSS
TochB	PK AS 70 recto, l. 1 © BnF		nä	nä	LETTER NA + VOWEL SIGN AE
TochA	A 338 recto, l. 6 © IDP		nä	nä	LETTER NA + BSS

Moreover, these forms are also inconsistently used within clusters, where treating the selector as a vowel sign would break the conjunct. Granted that in such a situation the

9. For the various views on it, see Maue & Sims-Williams (2024), § 2.2.1.2, p. 29. Note however he misreferenced Reuter (1925), § 78, p. 227, who actually proposed something along the lines of [ŭ].

10. Following Maue (2015): p. 503, fn. 4.

11. Maue & Sims-Williams 2024: § 2.2.1.1, pp. 28f.

conjunct joiner proposed below (§ 4.2) could be used, the proposed solution here is cleaner all factors considered.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochB	PK AS 7C recto, l. 2 © BnF		ṣpā	ṣpā	LETTER SSA + BSS + VIRAMA + LETTER PA + BSS
TochA	A 283 verso, l. 4 © IDP		ptṣa-k	ptsäk	LETTER PA + BSS + VIRAMA + LETTER TA + VIRAMA + LETTER SA + BSS + BAR VIRAMA + LETTER KA + BSS

The list of combinations with the *Sonderzeichen* selector, images taken from Wilson, is the following:

Name Without BSS With BSS

KA		
TA		
NA		
PA		
MA		
RA		
SHA		
SSA		



The *Sonderzeichen* <la> is excluded from this list, for which see § 8.3.

4. Other technical difficulties

Aside from missing base characters, there are other issues.

4.1. The bar virāma

A very common feature of NTB is the bar virāma, a line connecting two conjuncts indicating that the second conjunct has no implied vowel, sharing the function with the virāma. The last consonant of the second conjunct usually takes on its contextual *Sonderzeichen* form if present, and may be adorned with a “redundant” visible virāma above.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochA	A 110 recto, l. 1 © IDP		rpu- <u>s</u>	rpus	LETTER RA + VIRAMA + LETTER PA + VOWEL SIGN U + BAR VIRAMA + LETTER SA + BSS
TochB	IOL Toch 246 recto, l. 3 © IDP		sa-ñ <u>s</u>	sāñ	LETTER SA + BSS + BAR VIRAMA + LETTER NYA + VIRAMA
Uig.	U 6867 verso, l. 1 © DTA		syā- <u>r</u>	sār	LETTER SA + VIRAMA + LETTER YA + VOWEL SIGN AA + BAR VIRAMA + LETTER RA + BSS + VIRAMA
Uig.	Ch/U 6092, l. 4 ¹² © DTA		sva- <u>r</u>	svar	LETTER SA + VIRAMA + LETTER VA + BAR VIRAMA + LETTER RA + BSS + VIRAMA

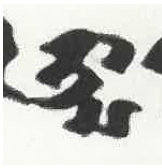
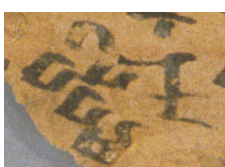
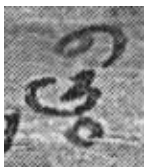
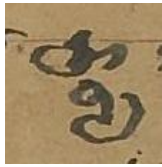
Considering a contextual variant of VIRAMA to be digitised after the consonant as Wilson suggested not only proves to be a great challenge to be made coherent with the existing behaviour of the sign, but also fails to provide a way to represent the redundant visible virāmas. As such, a separate codepoint BAR VIRAMA is proposed to be digitised between the two conjuncts, i.e. its logical position as it appears on the manuscript.

4.2. Joining conjuncts

The end of a conjunct is generally marked by the first vowel sign, but there are many cases in Turkestan Brāhmī where a conjunct can continue regardless. We propose a non-spacing

12. See Figure 2 for a description of the manuscript.

mark to be placed after the vowel. While the ZERO WIDTH JOINER (U+200D) could be made to work as such, it is already used for a different scope in Indic scripts. Following the naming practice of the NUMBER JOINER (U+1107F), we propose the addition of a CONJUNCT JOINER, henceforth abbreviated as BCJ. This control character, however, would never contrast with VIRAMA, which only appears between consonants or at the end of a conjunct,¹³ while BCJ would be used in vowel + consonant sequences. Nonetheless, *virāma* means “stop”, and using it in a context in which it does not denote the absence of a vowel would be confusing. Whether to encode this simply as VIRAMA or as a new codepoint is left as an open question for the SEW to give insight upon. Throughout this document proposed encodings are given with BCJ.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochA	YQ I.2 recto, l. 1 © Xinjiang Museum apud CEToM		upā	u _u pā	LETTER U + BCJ + LETTER PA + VOWEL SIGN AA
TochB	THT 286 recto, l. 4 © IDP		kuse	k _u se	LETTER KA + VOWEL SIGN U + BCJ + LETTER SA + VOWEL SIGN E
Uig.	U 6864 verso, l. 1 © DTA		oya	ö	LETTER O + BCJ + LETTER YA
Uig.	U 6831 recto, l. 4 © DTA		uyu	ü	LETTER U + BCJ + LETTER YA + VOWEL SIGN U
Uig.	U 6873a verso, l. 2 © DTA		isi-g ₁	isig	LETTER I + BCJ + LETTER SA + VOWEL SIGN I + BAR VIRAMA + LETTER G1A + VIRAMA
Chin.	IOL Khot S 7, l. 93 © IDP apud Sims- Williams 2001		uvī	為	LETTER U + BCJ + LETTER VA + VOWEL SIGN II
Turk.	P 2892 verso, l. 177 © BnF		ūvū	ü	LETTER UU + BCJ + LETTER VA + VOWEL SIGN UU

5. Tumshukese special signs

To represent its sounds not present in Tocharian, Tumshukese made use of twelve new signs, some of which are later also found in Uigur and Sogdian. These signs are conveniently

all found in the verso of TS 5,¹⁴ already published in Konow (1935), the publication that essentially discovered Tumshukese, after which the signs are traditionally called “Konow signs no. 1–12.”

For five out of the twelve signs the proposed names indicate no more than their Konow number, matching the <x_n> notation, while for the other seven the name is more descriptive about the sign’s supposed value, where it can be reasonably ascertained. Still it should be noted that the Unicode codepoint itself is tied to the Konow numbering and not to its currently accepted value, so if some of these readings in future research turn out to be inaccurate or get swapped around, the codepoints proposed here should still be used for the glyphs that they represent in this proposal. To make this more evident, appropriate annotations to the chart “Konow sign no. X” after each of the seven codepoints with a phonemic name are proposed.

5.1. <z>

Konow sign no. 1, traditionally transliterated as <z>, is attested in Tumshukese, Uigur and Sogdian, where it stands for [z]. We propose LETTER ZA.

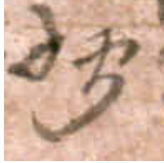
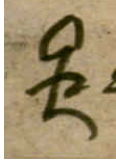
Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	TS 25, l. 3 © DTA		zba	—	LETTER ZA + VIRAMA + LETTER BA
Tumsh.	TS 31, l. 6 © DTA		rza	—	LETTER RA + VIRAMA + LETTER ZA
Uig.	U 5216 recto, l. 3 © DTA		zum̐	zun	LETTER ZA + VOWEL SIGN U + ANUSVARA
Uig.	Mz 208 recto, l. 4 © DTA		czyaṃ	čzün	LETTER CA + VIRAMA + LETTER ZA + VIRAMA + LETTER YA + ANUSVARA
Sogd.	Mz 639 recto, l. 6 © DTA		zi-rk̕	zirk	LETTER ZA + VOWEL SIGN I + BAR VIRAMA + LETTER RA + VIRAMA + LETTER KA + BSS
Sogd.	bs 9 verso, l. 3 © DTA		zā	[[zā]]	LETTER ZA + VOWEL SIGN AA

13. To avoid visible viramas from merging conjuncts together, the ZERO WIDTH NON-JOINER (U+200C) should be used.

14. See Figure 1.

5.2. <g₁>

Konow sign no. 2, present in Tumshukese, Uigur and Sogdian, was traditionally transliterated as <g₁> or <γ> and originally stood for the voiced guttural fricative [ɣ]. We propose a new codepoint LETTER G1A.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	TS 31, l. 2 © DTA		g ₁ u	—	LETTER G1A + VOWEL SIGN U
Tumsh.	TS 28, l. 31 © DTA		g ₁ i	—	LETTER G1A + VOWEL SIGN I
Uig.	U 6873a verso, l. 4 © DTA		g ₁ k ₁ a	gka	LETTER G1A + VIRAMA + LETTER K1A
Uig.	U 6839 verso, l. 3 © DTA		ši-g ₁	šig	LETTER SHA + VOWEL SIGN I + BAR VIRAMA + LETTER G1A + VIRAMA
Sogd.	Mz 639 recto, l. 3 © DTA		g ₁ a	γə	LETTER G1A

Furthermore, sign no. 2 of the Khüis Tolgoi inscription has also been tentatively identified as this <g₁>.¹⁵

5.3. <g₂> or <x₃>

Konow sign no. 3, common in Tumshukese. Also present in Uigur and Sogdian, though less commonly, where it appears to be used as an allograph of <g₁>, hence the traditional transliteration <g₂>. From this it was generally assumed that the sign must have stood for some sort of voiced guttural in Tumshukese as well, though the characteristic that made it distinctive from <g₁> was unclear. Recently Maue proposed for its value in Tumshukese to actually to be [d̪z], and that Uigur and Sogdian, lacking the sound, had no need to distinguish it from the similar-looking <g₁>.¹⁶ We propose to name the codepoint LETTER X3A.

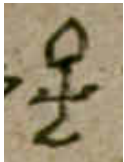
15. Maue 2017: §2.2.1.1, p. 4ff.

16. See Maue & Sims-Williams 2024: § 2.2.2, p. 32

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	TS 15 verso, l. 4 © DTA		x ₃ ä	—	LETTER X3A + VOWEL SIGN AE
Tumsh.	TS 19 verso, l. 2 © DTA		x ₃ i	—	LETTER X3A + VOWEL SIGN I
Sogd.	bs 1 recto, l. 1 © DTA		mg ₂ om	myōn	LETTER MA + VIRAMA + LETTER X3A + VOWEL SIGN O + ANUSVARA
Sogd.	bs 1 recto, l. 1 © DTA		g ₂ u	yu	LETTER X3A + VOWEL SIGN U

5.4. <z>

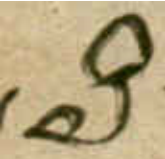
Konow sign no. 4, present in Tumshukese, Uigur and Sogdian. By analogy of the name of LETTER SSA we propose a new codepoint with the name LETTER ZZA.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	bi 37 recto, l. 2 © DTA		za	—	LETTER ZZA
Tumsh.	TS 27, l. 1 © DTA		zdi	—	LETTER ZZA + VIRAMA + LETTER DA + VOWEL SIGN I
Uig.	Mz 647 b recto, l. 5 © DTA		zum	žun	LETTER ZZA + VOWEL SIGN U + ANUSVARA
Sogd.	Mz 639 verso, l. 1 © DTA		zā	žā	LETTER ZZA + VOWEL SIGN AA

5.5. <x₅>

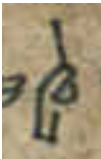
Konow sign no. 5, present only in Tumshukese. Konow proposed that it stood for a palatised <ž> or <š>,¹⁷ while Skjærvø showed how it can be made to derive from earlier intervocalic *š and considered [ʒ] the most likely value for it.¹⁸ We propose a new codepoint LETTER X5A.

17. Konow 1935: p. 10

Language	Location	Image	Transl.	Proposed encoding
Tumsh.	bs 3 recto, l. 2 © DTA		x _{5a}	LETTER X5A
Tumsh.	TS 25, l. 24 © DTA		x _{5i}	LETTER X5A + VOWEL SIGN I
Tumsh.	IOL Khot 203/11 verso, l. 4 © IDP		x _{5a}	LETTER X5A

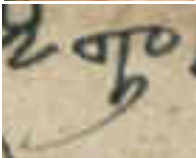
5.6. <x₆>

Konow sign no. 6, present only in Tumshukese. Konow tentatively analysed it as <z>,¹⁹ while Dragoni showed it to be the outcome of earlier intervocalic *θ after a front vowel, and listed [ç] and [j] as possible phonetic values,²⁰ though this is not universally accepted.²¹ We propose the codepoint LETTER X6A.

Language	Location	Image	Transl.	Proposed encoding
Tumsh.	TS 39, l. 7 © DTA		x _{6ye}	LETTER X6A + VIRAMA + LETTER YA + VOWEL SIGN E

5.7. <d₁>

Konow sign no. 7, present in Tumshukese, Uigur and Sogdian. Often also transcribed as <δ> or <ḍ>. We propose the creation of a new codepoint as LETTER D1A.

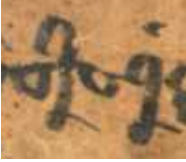
Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	TS 4 verso, l. 4 © DTA		d _{1ä}	—	LETTER D1A + VOWEL SIGN AE
Tumsh.	TS 40, l. 2 © DTA		d _{1u}	—	LETTER D1A + VOWEL SIGN U

18. Skjærvø 1987: § 9, p. 84

19. Konow 1935: p. 10

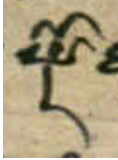
20. Dragoni 2020: § 10, p. 218

21. Maue & Sims-Williams 2024: § 2.2.2, p. 32

Uig.	Mz 190 verso, l. 3 © DTA		d ₁ i-z	d ₁ z	LETTER D1A + VOWEL SIGN I + BAR VIRAMA + LETTER ZA + VIRAMA
Uig.	Mz 615 recto, l. 6 © DTA		rd ₁ yā	rdä	LETTER RA + VIRAMA + LETTER D1A + VIRAMA + LETTER YA + VOWEL SIGN AA
Sogd.	THT 2172 recto, l. 2 © IDP		d ₁ e	—	LETTER D1A + VOWEL SIGN E
Sogd.	Mz 639 recto, l. 4 © DTA		d ₁ wa	δva	LETTER D1A + VIRAMA + LETTER WA

5.8. <x₈>

Konow sign no. 8, present only in Tumshukese. Maue & Ogihara proposed it to represent [d̥z],²² and later Maue analysed it as [ts^h].²³ We propose a new codepoint LETTER X8A.

Language	Location	Image	Transl.	Proposed encoding
Tumsh.	TS 23, l. 12 © DTA		x ₈ ā	LETTER X8A + VOWEL SIGN AA
Tumsh.	TS 26, l. 4 © DTA		x ₈ O	LETTER X8A + VOWEL SIGN O
Tumsh.	TS 21c verso, l. 1 © DTA		x ₈ O	LETTER X8A + VOWEL SIGN O

5.9. <x₉>

Konow sign no. 9, present in Tumshukese and Sogdian. Konow did not have other attestations besides TS 5,²⁴ and Hitch assumed it to be a variant of <d₁>.²⁵ Duàn Qíng proposed a reading <yy> after finding more attestations,²⁶ and Maue & Sims-Williams analysed it as <h^w>.²⁷ We propose a new codepoint LETTER X9A.

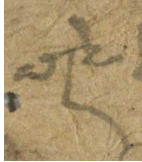
22. Maue & Ogihara 2017: § 3, p. 424

23. Maue & Sims-Williams 2024: § 2.2.2, p. 32

24. Konow 1935: p. 776

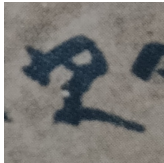
25. Hitch 1981: pp. 62, 66

26. Duàn 2021

Language	Location	Image	Transl.	Proposed encoding
Tumsh.	bi 37 recto, l. 4 © DTA		x ₉ e	LETTER X9A + VOWEL SIGN E

5.10. <ž>

Konow sign no. 10, present in Tumshukese, where it is in phonological opposition with <ž>, and later also in Uigur and Sogdian, where it is used interchangeably with <ž>.²⁸ Following the name of LETTER SHA, we propose a new codepoint called LETTER ZHA.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	TS 25, l. 12 © DTA		žo	—	LETTER ZHA + VOWEL SIGN O
Tumsh.	TS 38, l. 1 © DTA		ži	—	LETTER ZHA + VOWEL SIGN I
Sogd.	bs 6 recto, l. 5 © DTA apud Maue & Sims-Williams 2024		ža	ža	LETTER ZHA

5.11. <v₁>

Konow sign no. 11, found in Tumshukese, Uigur and Sogdian. Initially interpreted by Konow <dz>,²⁹ Maue recognised it to stand for [w],³⁰ and this was widely recognised. Initially also transliterated as <v̲>, the transliteration <v₁> became standard. We propose a new codepoint named LETTER V1A.

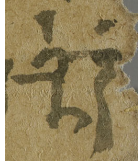
Language	Location	Image	Transl.	Transcr.	Proposed encoding
Tumsh.	TS 46 recto, l. 2 © DTA		v ₁ a	—	LETTER V1A
Tumsh.	TS 26, l. 10 © DTA		v ₁ e	—	LETTER V1A + VOWEL SIGN E

27. Maue & Sims-Williams 2024: § 2.2.2, p. 53

28. Maue & Sims-Williams 2024: p. 34, fn. 67

29. Konow 1935: passim

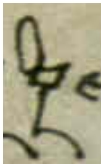
30. Maue 1996: p. 215ff.

Uig.	Mz 682d recto, l. 3 © DTA		hv ₁ ā	hwa	LETTER HA + VIRAMA + LETTER V1A + VOWEL SIGN AA
Sogd.	bs 7 recto, l. 2 © DTA		v ₁ a	wa	LETTER V1A
Sogd.	bs 9 verso, l. 3 © DTA		v ₁ e-ṛ	wer	LETTER V1A + VOWEL SIGN E + BAR VIRAMA + LETTER RA + BSS + VIRAMA

Sign no. 4 of the Khüis Tolgoi inscription, of unclear interpretation, has also been tentatively identified as either <ṛ̃> or this <v₁>.³¹

5.12. <χš>

Konow sign no. 12, found only in Tumshukese, transliterated as <χš> or <xš> already since Konow. We propose a new codepoint named LETTER XSHA.

Language	Location	Image	Transl.	Proposed encoding
Tumsh.	TS 1 recto, l. 1 © DTA		χše	LETTER XSHA + VOWEL SIGN E
Tumsh.	TS 22, l. 6 © DTA		χši	LETTER XSHA + VOWEL SIGN
Tumsh.	TS 25, l. 1 © DTA		χša	LETTER XSHA

Sign no. 3 of the Khüis Tolgoi inscription has been suggested to be either a <kṣ> ligature or, though with a less elaborate appearance, this Tumshukese <χš>.³²

6. Khotanese special signs

Though the main focus of this proposal is NTB, once its treatment is properly established STB would only a couple of codepoints away from being supported as well. For this it is treated in this same proposal.

The script, aside from Khotanese and Sanskrit, was also used to write in Middle Chinese (IOL Khot S 7 = Ch 00120) and Old Turkic (P 2892), which are given here mostly for curiosity, as

31. Maue 2017: §2.2.1.3, p. 10

32. Maue 2017: §2.2.1.2, pp. 9f.

they do not differ in neither appearance nor functionalities from their contemporary Late Khotanese Cursive manuscripts.

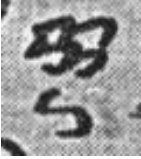
6.1. <ei>

Khotanese employed a new vowel sign, which resembled a St. Andrew's cross, traditionally transcribed <ei> which stood for /aə/.³³ We propose a new codepoint VOWEL SIGN EI. We were not able to attest an independent vowel shape which would have required a new codepoint LETTER EI,³⁴ so we propose to reserve a codepoint for it without defining it.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Khot.	IOL Khot 14/2 recto, l. 2 © IDP		nei	nei	LETTER NA + VOWEL SIGN EI
Khot.	IOL Khot 9/12 recto, l. 3 © IDP		ysei	ysei	LETTER YA + VIRAMA + LETTER SA + VOWEL SIGN EI

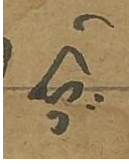
6.2. The hook

Khotanese employed a “subscript hook” below a conjunct to imply that its consonant was to be pronounced with an accompanying breathiness.³⁵ We propose the codepoint HOOK BELOW.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Khot.	IOL Khot 13/3 recto, l. 2 © IDP		pä'	pä'	LETTER PA + VOWEL SIGN AE + HOOK BELOW
Khot.	IOL Khot 16/10 recto, l. 1 © IDP		va'	va'	LETTER VA + HOOK BELOW
Khot.	Or 6399/2.7, l. 1 © IDP		kva'	kva'	LETTER KA + VIRAMA + LETTER VA + HOOK BELOW
Chin.	IOL Khot S 7, l. 88 © IDP apud Sims-Williams 2001		śū'	如	LETTER SHA + VOWEL SIGN UU + HOOK BELOW

33. Emmerick *et al.* 2024: § I.4.1, p. 17

34. This letter is found in Wilson's proposal, but no evidence is given.

Turk.	P 2892 verso, l. 173 © BnF		hām':	yo	LETTER HA + VOWEL SIGN AA + ANUSVARA + HOOK BELOW + VISARGA
-------	-------------------------------	---	-------	----	---

7. Other less straightforward special signs

Aside from the signs already in Tumshukese, there was little innovation of the script for Uigur and Sogdian. However there are three phenomena worth noting.

7.1. <k₁>

Uigur repurposed the Tocharian <kk> ligature³⁶ to stand for the voiceless guttural occlusive around back vowels, traditionally transcribed <k₁>, <q> or <ḱ>. This fact would not by itself merit a separate codepoint, if it weren't that legitimate <kk> sequences in Uigur do not appear as that ligature as they do in Tocharian, which excludes the possibility of encoding <k₁> simply as LETTER KA + VIRAMA + LETTER KA. We hence propose the creation of a new codepoint LETTER K1A. A font made to support NTB should hence take into consideration in which language the text was tagged as to know how to properly display the LETTER KA + VIRAMA + LETTER KA sequence.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochB	THT 89 recto, l. 5 © IDP		kka	kka	LETTER KA + VIRAMA + LETTER KA
Skr. ³⁷	Mz 682f recto, l. 3 © Maue 1996		kka	kka	LETTER KA + VIRAMA + LETTER KA
Uig.	Mz 635 verso, l. 5 © DTA		k ₁ i	—	LETTER K1A + VOWEL SIGN I
Uig.	MIK III 418, l. 1 ³⁸ © Maue 1996		lk ₁ u	lku	LETTER LA + VIRAMA + LETTER K1A + VOWEL SIGN U

35. Emmerick *et al.* 2024: § I.4.3, p. 17

36. The term “ligature” here appears inappropriate, since all conjuncts are naturally ligated. What is meant here is that the sequence had an irregular appearance, since the second <k> appeared as a horizontal stroke, closer to its original Aśokan Brāhmī form.

37. In a Sanskrit-Uigur bilingual text.

38. A stone inscription.

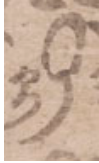
Uig.	Mz 192 recto, l. 2 © DTA		kkyā	gkă	LETTER KA + VIRAMA + LETTER KA + VIRAMA + LETTER YA
Uig.	U 5214 verso, l. 4 © DTA		kkyā	gkă	LETTER KA + VIRAMA + LETTER KA + VIRAMA + LETTER YA + VOWEL SIGN AA

Furthermore, sign no. 1 of the Khüis Tolgoi inscription has also been tentatively identified as this <k₁>.³⁹

7.2. <rr>

Not dissimilar to the <kk> ~ <k₁> case described above is the case of the sign traditionally transcribed as <rr>, found in Tumshukese, Uigur, Khotanese and perhaps Sogdian, standing for some rhotic. When at the beginning of a conjunct, i.e. in a “radical” context, its appearance is that of an expected <r> + <r> sequence, i.e. the initial shape followed by the final shape,⁴⁰ and in fact this sequence is attested in Tocharian in the same appearance.

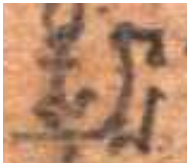
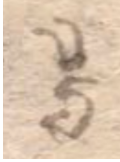
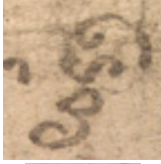
It is unexpected however that when preceded by another consonant, i.e. in a “dependent” context, the sign appears as a looping final <r>. General academic consensus until recently emphasised how this sign is to be treated as a separate sign derived by alteration of <r>, and not as a <r> + <r> sequence. Contrary to this, a recent article argues how this is best analysable as originating as a ligature.⁴¹ Nonetheless, the sign contrasts with LETTER RA + VIRAMA + LETTER RA appearing in Khotanese, though rarely, as the isolated form with an initial, or “superscript”, form above it. We hence propose a new codepoint LETTER RRA.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochB	THT 350 verso, l. 6 © IDP		rrā	rrā	LETTER RA + VIRAMA + LETTER RA + VOWEL SIGN AA
TochB	Ot 12, l. 12 © IDP		rri	rri	LETTER RA + VIRAMA + LETTER RA + VOWEL SIGN I
Tumsh.	TS 29, l. 7 © DTA		rre	—	LETTER RRA + VOWEL SIGN E

39. Maue 2017: §2.2.1.1, pp. 4ff.

40. When in this sequence, the final <r> shape is also used preconsonantly. Since this is the only context in which <r> can occur medially, this is not an exception to any rule, and it is presumable that a medial <r> in another context would have also had the final shape.

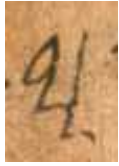
41. Del Tomba 2025. See pp. 40f. for the opposing views of nearly all publications preceding it.

Uig.	U 5208 verso, l. 12 © DTA		rryā	rä	LETTER RRA + VIRAMA + LETTER YA + VOWEL SIGN AA
Khot.	IOL Khot 14/1 recto, l. 4 © IDP		rrū	rrū	LETTER RRA + VOWEL UU
Khot.	Or 11344/15 recto, l. 1 © IDP		rrvī	rrvī	LETTER RRA + VIRAMA + LETTER VA + LETTER II
Khot.	IOL Khot 130/2 verso, l. 4 © IDP		dr̥rām	dr̥rā	LETTER DA + VIRAMA + LETTER RRA + VOWEL SIGN AA + ANUSVARA
Khot.	Or 11344/14, l. 2 © IDP		dr̥rai	dr̥rai	LETTER DA + VIRAMA + LETTER RRA + VOWEL SIGN AI
Khot.	Or 11344/10, l. 2 © IDP		hvr̥rī	hvr̥rī	LETTER HA + VIRAMA + LETTER RRA + VOWEL SIGN II
Khot.	IOL Khot 132/3 verso, l. 4 © IDP		r-ra	r-ra	LETTER RA + VIRAMA + LETTER RA
Khot.	IOL Khot 132/3 verso, l. 5 © IDP		r-ra	r-ra	LETTER RA + VIRAMA + LETTER RA

7.3. Uigur Trägerzeichen

This circular sign, called *Trägerzeichen* by Maue,⁴² i.e. “carrier sign”, is attested a handful of times seemingly only in a single Uigur document (U 6889 = T III M 140), which in general has a very unconventional use of the Brāhmī script for the language. The sign appears in the conjuncts <’yu> and <’ya>. Maue notes the resemblance both in appearance and in function with <𑖦> in the Tibetan script, which is yet another Brahmic script in which the Uigur language was written in by Buddhists. Matching the Tibetan codepoint name, we propose a new codepoint LETTER -A.

42. Maue 2015: p. 33f.

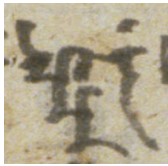
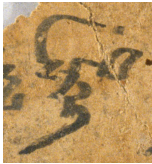
Language	Location	Image	Transl.	Transcr.	Proposed encoding
Uig.	U 6889 recto, l. 4 © DTA		'yu	ü	LETTER -A + VIRAMA + LETTER YA + VOWEL SIGN U
Uig.	U 6889 verso, l. 1 © DTA		'yu	ü	LETTER -A + VIRAMA + LETTER YA + VOWEL SIGN U

8. Final remarks

The following sections do not require any intervention from Unicode, but are noted here for encoders of the manuscripts and font creators, about features of the script which may pose questions about their encoding.

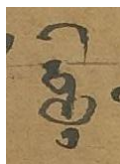
8.1. Double vowel signs

Both NTB and STB employed double vowel signs in a single conjunct, which is generally not allowed by Indian Brāhmī. The conjunct joiner BCJ should not be employed for these cases, for which we suggest to simply encode the vowel signs one after the other.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
TochA	A 355 verso, l. 3 © IDP		krue ⁴³	kr _u e	LETTER KA + VIRAMA + LETTER RA + VOWEL SIGN U + VOWEL SIGN E
Uig.	U 5214 recto, l. 5 © DTA		hwa	hwa	LETTER HA + VIRAMA + LETTER WA + VOWEL SIGN U + VOWEL SIGN AA
Sogd.	bs 2 verso, l. 5 © DTA		hyue- _r	xü(w)er or x _u er	LETTER HA + VIRAMA + LETTER YA + VOWEL SIGN U + VOWEL SIGN E + BAR VIRAMA + LETTER RA + BSS + VIRAMA
Sogd.	bs 7 verso, l. 2 © DTA		skuā- _t	sku(w)āt or skwāt	LETTER SA + VIRAMA + LETTER KA + VOWEL SIGN U + VOWEL SIGN AA + BAR VIRAMA + LETTER TA + BSS + VIRAMA
Khot.	IOL Khot 13/3 recto, l. 2 © IDP		ysmuī	ysmuī	LETTER YA + VIRAMA + LETTER SA + VIRAMA + LETTER MA + VOWEL SIGN U + VOWEL SIGN II

43. The image presented also contains the entirety of the preceding conjunct, <sām>.

Turk.

P 2892 verso, l. 177
© BnF

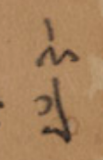
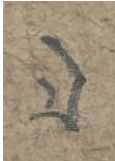
vūä'

vu

LETTER VA + VOWEL
SIGN UU + VOWEL SIGN
AE + HOOK BELOW

8.2. Khotanese <a>-carrier

Much like the *hapax legomenon* possibly Tibetan-inspired Uigur carrier (§ 7.3), it is not uncommon in Khotanese to find the independent <a> letter working as a carrier of other vowels. Traditionally transcribed with an apostrophe before the vowel, e.g. <'au>, or on top of it, e.g. <ī̇>. We propose to treat this as if it were a double vowel combination.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
Khot.	IOL Khot 130/2 verso, l. 4 © IDP		'ī	ī	LETTER A + VOWEL SIGN II
Khot.	IOL Khot S 10, l. 44 © IDP		'au	au	LETTER A + VOWEL SIGN AU
Khot.	P 4089a recto, l. 3 © BnF		'e	e	LETTER A + VOWEL SIGN E

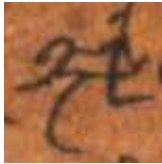
Point could be made that the independent vowel letter <ä> (§ 3.1) is formed in the same fashion, and hence does not require a separate codepoint as it could be encoded as LETTER A + VOWEL SIGN AE. This observation would not be wrong, and the only reason why a codepoint it is proposed here is to maintain consistency with the treatment of other vowels.

8.3. The so-called *Sonderzeichen* <la>

The *Sonderzeichen* for /l/, generally transcribed as <la>, corresponds in fact with the vocalic <ī̇> from the Indian Brāhmī.⁴⁴ Its appearance on the recto of TS 5, a syllabary of the Tocharian signs, in the sequence <ṛ ṛ ī̇ ī̇> rather than among the *Sonderzeichen* of the line below proves that the users of the script themselves did not consider it to be a *Sonderzeichen*, despite it functioning like one. Moreover, the long vocalic <ī̇> is also attested in Uigur in contexts where it is transliterated as <ī̇a>.

Taking these factors into consideration, as far as Unicode encoding is concerned, <la> and <ī̇a> are best never treated as *Sonderzeichen*. This means that instead of **LETTER LA + BSS** one should use **LETTER VOCALIC L**, for a post-consonantal context instead of **LETTER ?A + VIRAMA + LETTER LA + BSS** one should use **LETTER ?A + VOWEL SIGN VOCALIC L**, and for a pre-consonantal sign instead of **LETTER LA + BSS + VIRAMA + LETTER ?A** one should use **LETTER VOCALIC L + BCJ + LETTER ?A**.

44. Maue (2014) covers the issue in detail.

Language	Location	Image	Transl.	Transcr.	Proposed encoding
—	TS 5 recto, l. 1 © DTA		l̥l̥	—	LETTER VOCALIC L + LETTER VOCALIC LL
TochA	YQ II.10 verso, l. 1 © Xinjiang Museum apud CETOm		pl̥a	pl̥ä	LETTER PA + BSS + VOWEL SIGN VOCALIC L
TochB	THT 295 verso, l. 3 © IDP		l̥ka	lkä	LETTER VOCALIC L + BCJ + LETTER KA + BSS
Uig.	Mz 648a verso, l. 4 © DTA		l̥a-g̥1̥	l̥g	LETTER VOCALIC LL + BAR VIRAMA + LETTER G1A + VIRAMA
Uig.	Mz 612a verso, l. 6 © DTA		l̥ä-[g̥1̥]	l̥g	LETTER VOCALIC LL + VOWEL SIGN AE⁴⁵ + BAR VIRAMA + [LETTER G1A] + VIRAMA

9. Proposed codepoints

The number of new codepoints proposed ranges from 22 to 32. The Brāhmī block currently only has no more than 13 empty slots left, so a new block BRAHMI EXTENDED appears to be in need, for which we would not know how to suggest the position nor the size. For these reasons, this proposal does not suggest the precise codepoints of the new characters requested.

Five of the following vectorial renditions of the letters (LETTER AE, VOWEL SIGN AE, LETTER WA, VOWEL SIGN EI and HOOK BELOW) are taken from Wilson's proposals without further modification. Most of the remaining ones were created by modifying his work. Some of these may benefit from a touch-up.

Image	Name	NTB	STB
	BRAHMI LETTER AE	yes	yes
	BRAHMI VOWEL SIGN AE		

45. This case is treated as if it were a double vocalisation, mentioned on § 8.1.

	RESERVED		
	BRAHMI VOWEL SIGN EI	no	yes
	BRAHMI LETTER WA	yes	no
	BRAHMI LETTER ZA	yes	no
	BRAHMI LETTER G1A	yes	no
	BRAHMI LETTER X3A	yes	no
	BRAHMI LETTER ZZA	yes	no
	BRAHMI LETTER X5A	yes	no
	BRAHMI LETTER X6A	yes	no
	BRAHMI LETTER D1A	yes	no
	BRAHMI LETTER X8A	yes	no
	BRAHMI LETTER X9A	yes	no

	BRAHMI LETTER ZHA	yes	no
	BRAHMI LETTER V1A	yes	no
	BRAHMI LETTER XSHA	yes	no
	BRAHMI LETTER K1A	yes	no
	BRAHMI LETTER RRA	maybe ⁴⁶	yes
	BRAHMI LETTER -A	yes	no
	BRAHMI HOOK BELOW	no	yes
	BRAHMI BAR VIRAMA	yes	no
	BRAHMI CONJUNCT JOINER	maybe ⁴⁷	
	BRAHMI SONDERZEICHEN SELECTOR, or the 10 <i>Sonderzeichen</i> atomically	yes	no

10. Bibliography

BAILEY, Harold Walter (1951). *Khotanese Buddhist Texts*, London: Taylor's Foreign Press

BAILEY, Harold Walter (1969). *Khotanese Texts*, vol. 1, 2nd ed., Cambridge: University Press

46. The most compelling arguments for the inclusion of the ligature as a separate codepoint are its dependent shape and contrast with <r-r>, both of which only appear in STB. Argument for inclusion in a hypothetical separate NTB block would be consistency with STB. This problem does not poses itself if the two variants are encoded in the same Brāhmī block.

47. The character does not contrast with VIRAMA and serves a very similar purpose, see § 4.2.

- BnF: *Gallica*, Bibliothèque nationale de France, <https://gallica.bnf.fr>
- CETOM: *A Comprehensive Edition of Tocharian Manuscripts*, Universität Wien, <https://cetom.univie.ac.at>
- DEL TOMBA, Alessandro (2025). “The palaeography of *rr* and *r* in Khotanese.” In *Journal Asiatique*, vol. 313, no. 1, pp. 33–77
- DRAGONI, Federico (2020). “The Tumshuqese year of the goat and the *Fremdzeichen* x₆.” In *Journal Asiatique*, vol. 308, no. 2, pp. 215–23
- DTA: *Digitales Turfan-Archiv*, Berlin-Brandenburgischen Akademie der Wissenschaften, <https://turfan.bbaw.de/dta>
- DUÀN Qíng 段晴 (2021). “黄文弼发现的两件据史德语文书 Huáng Wénbì fāxiàn de liǎng jiàn jùshǐdéyǔ wénshū / Two Gūzdiya documents discovered by Huang Wenbi.” In 西域文史 *Xīyù wénshǐ* / *The Western Regions Literature and History*, vol. 15, pp. 1–18
- EMMERICK, Ronald E & RÓNA-TAS, András (1992). “The Turkish-Khotanese Wordlist Revisited.” In *Central Asiatic Journal*, vol. 36, no. 3/4, pp. 199–241
- EMMERICK, Ronald E. & PULLEYBLANK, Edwin G. (1993). *A Chinese Text in Central Asian Brahmi Script. New Evidence for the Pronunciation of Late Middle Chinese and Khotanese* (Serie Orientale Roma: 69), Rome: Istituto italiano per il medio ed estremo oriente
- GABAIN, Annemarie von (1954). *Türkische Turfan-Texte VIII. Texte in Brāhmīscript* (Abhandlungen der deutschen Akademie der Wissenschaften zu Berlin: 1952, no. 7)
- HITCH, Doug (1981). *Central Asian Brahmi palaeography: the relationship among the Tocharian, Khotanese, and Old Turkic Gupta script*, Calgary (MA dissertation, unpubl.)
- IDP: *International Dunhuang Programme*, British Library: <https://idp.bl.uk>, Berlin-Brandenburgischen Akademie der Wissenschaften: <https://idp.bbaw.de>, and others
- KONOW, Sten (1935). “Ein neuer Saka-Dialekt.” In *Sitzungsberichten der Preussischen Akademie der Wissenschaften. Phil.-hist. Klasse*, 1935, 20, pp. 772–823, Berlin: Walter de Gruyter
- MAUE, Dieter (1996). *Alttürkische Handschriften: Dokumente in Brāhmī und tibetischer Schrift*, vol. 1 (Verzeichnis der orientalischen Handschriften in Deutschland: 13, 9)
- MAUE, Dieter (2009). *Tumshukese Manuscripts and Inscriptions*, <https://titus.uni-frankfurt.de/texte/iranica/tumshuq/transliteration.pdf>
- MAUE, Dieter (2014). *Simply forgotten? Note on the Tocharian Fremdzeichen la*, <https://www.academia.edu/8346630>
- MAUE, Dieter (2015). *Alttürkische Handschriften: Dokumente in Brāhmī und tibetischer Schrift*, vol. 2 (Verzeichnis der orientalischen Handschriften in Deutschland: 13, 27)
- MAUE, Dieter (2017). *The Khüis Tolgoi inscription – Signs and sounds*, draft ed., <https://www.academia.edu/34589694>, later summarised in MAUE, Dieter et al. (2019). “The Khüis Tolgoi inscription.” In *Proceedings of the 60th Meeting of the Permanent International Altaistic Conference (PIAC). August 27–September 1, 2017*, pp. 73–89
- MAUE, Dieter & OGIHARA Hirotoshi (2017). “Tumshukische Miszellen III. 3. Tumshukese dental affricates.” In *Iranica*, vol. 25, *Studien zu Manichäismus, Iranistik und Zentralasienkunde im Gedenken an Werner Sundermann*, pp. 421–32
- MAUE, Dieter & SIMS-WILLIAMS, Nicholas (2024). *Sogdian documents in Brāhmī script* (Corpus Inscriptionum Iranicarum: part 2, vol. 3/6), London: School of Oriental and African Studies
- PEYROT, Michaël (2025). “On the so-called ‘Fremdvokal’ ä in Tocharian and Khotanese and its origins.” In CHING Chao-jung (ed.), *Text, script and language in Bactria and Serindia*, pp. 275–96, Wiesbaden: Reichert
- REUTERS, J. N. (1925). “Bemerkungen über die neuen Lautzeichen im Tocharischen.” In *Studia Orientalia*, vol. 1, pp. 194–237

- SANDER, Lore (1968). *Paläographisches zu den Sanskrithandschriften der Berliner Turfansammlung*, Wiesbaden: Franz Steiner
- SIMS-WILLIAMS, Ursula (2001). "The Silk Road: Trade, Travel, War and Faith." In *Bulletin of the Asia Institute, New Series*, vol. 15, pp. 163–70
- SKJÆRVØ, Prods Oktor (1987). "On the Tumshuqese *Karmavācanā* text." In *Journal of the Royal Asiatic Society*, no. 1, 1987, pp. 77–90.
- SKJÆRVØ, Prods Oktor (2013). *Khotanese Manuscripts from Chinese Turkestan in the British Library*, online ed., https://idp.bl.uk/references/Skjaervø_2013/

11. Gallery



Figure 1. TS 5 verso © DTA. This manuscript containing a list of the twelve Tumshukese characters not present in Tocharian. From left to right, their order on the manuscript corresponds to their Konow number. The recto, not shown here, is syllabary of a Tocharian Brāhmī.

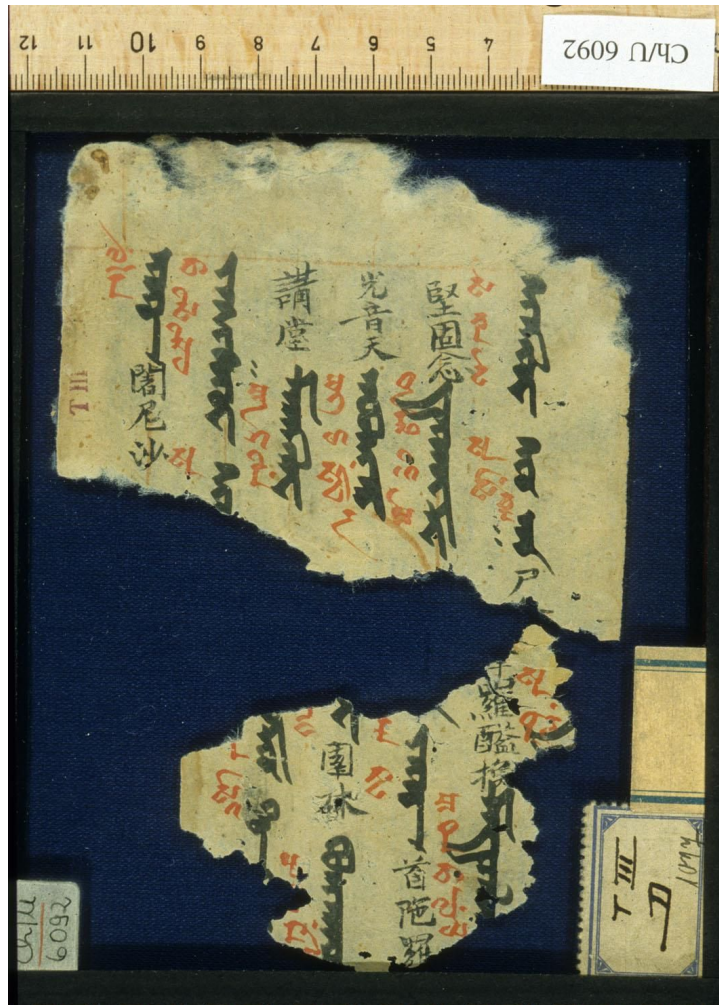


Figure 2. Ch/U 6092 recto © DTA. This is a bilingual Chinese-Uigur text, where the Uigur is written both in the Uigur script in black ink and in a Brāhmī transcription in red ink to its left, resembling a sort of “ruby”. To accomodate the Uigur script being vertical, the Brāhmī transcription is vertical as well, which makes the bar virāma on the fourth line, <sva-r̥>, also appear vertical. The bar virāma on the sixth line however, <sā-k̐>, appears horizontal as usual.

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹**

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html> for guidelines and details before filling this form.

Please ensure you are using the latest Form from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest Roadmaps.

A. Administrative

1. Title:	<i>Preliminary proposal for Turkestan Brāhmī within the Brāhmī block</i>
2. Requester's name:	<i>Francesco Catoni</i>
3. Requester type (Member body/Liaison/Individual contribution):	<i>Individual contribution</i>
4. Submission date:	<i>2026-09-11</i>
5. Requester's reference (if applicable):	
6. Choose one of the following:	
This is a complete proposal:	<i>No</i>
(or) More information will be provided later:	<i>Yes</i>

B. Technical – General

1. Choose one of the following:	
a. This proposal is for a new script (set of characters):	<i>No</i>
Proposed name of script:	
b. The proposal is for addition of character(s) to an existing block:	<i>Yes</i>
Name of the existing block:	<i>Brahmi</i>
2. Number of characters in proposal:	<i>From 22 to 32</i>
3. Proposed category (select one from below - see section 2.2 of P&P document):	
A-Contemporary ☐ B.1-Specialized (small collection) ☐ B.2-Specialized (large collection) ☐	
C-Major extinct ☒ D-Attested extinct ☐ E-Minor extinct ☐	
F-Archaic Hieroglyphic or Ideographic ☐ G-Obscure or questionable usage symbols ☐	
4. Is a repertoire including character names provided?	<i>Yes</i>
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?	<i>Yes</i>
b. Are the character shapes attached in a legible form suitable for review?	<i>Yes</i>
5. Fonts related:	
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	
6. References:	
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?	<i>Yes</i>
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	<i>Yes</i>
7. Special encoding issues:	
Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?	<i>No</i>

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at <http://www.unicode.org> for such information on other scripts. Also see Unicode Character Database (<http://www.unicode.org/reports/tr44/>) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

¹ Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	Partially
If YES explain See section 2.	
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	No
If YES, with whom?	
If YES, available relevant documents:	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	No
Reference:	
4. The context of use for the proposed characters (type of use; common or rare)	Academic
Reference:	
5. Are the proposed characters in current use by the user community?	
If YES, where? Reference:	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	No
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	No
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	Yes
If YES, is a rationale for its inclusion provided?	
If YES, reference: See sections 7.1 and 7.2.	
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	No
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?	Yes
If YES, is a rationale for its inclusion provided?	
If YES, reference: See section 4.2.	
11. Does the proposal include use of combining characters and/or use of composite sequences?	Yes
If YES, is a rationale for such use provided?	
If YES, reference: See sections 3.3, 8.1 and 8.2.	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	Yes
If YES, reference: See section 3.3.	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	Yes
If YES, describe in detail (include attachment if necessary)	
See sections 3.3 and 4.2.	
13. Does the proposal contain any Ideographic compatibility characters?	No
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	