L2/15-236

Title: Proposal to Encode the Tocharian Script
Author: Lee Wilson (ttlwilson@hotmail.com)
Date: 2015-10-09

1 Introduction

Tocharian is a Brahmi-based script historically used to write the Tocharian languages (ISO 639-3: xto, txb), traditionally referred to as Tocharian A and Tocharian B, which belong to the Tocharian branch of the Indo-European language family. The script was in use primarily during the 8th century CE by the people inhabiting the northern edge of the Tarim Basin of the Taklamakan Desert in what is now Xinjiang in western China, an area formerly known as Turkestan.

The Tocharian script is attested in over 4,000 extant manuscripts that were discovered in the early 20th century in the Tarim Basin. The discovery was significant for two reasons: first, it revealed a previously unknown branch of the Indo-European language family, and second, it marked the easternmost geographical extent of that language family.

The most distinctive aspect of Tocharian script is the vowel transcribed as ä, commonly known as the *Fremdvokal*, and the modifications to the script that it entails. This vowel is indicated with a vowel sign not employed in scripts outside of the Tarim Basin region. However, this vowel sign is not commonly used when writing Tocharian. Instead, sequences of a consonant followed by the *Fremdvokal* are most typically indicated by 11 CV signs known as *Fremdzeichen*, which are unique to Tocharian.

2 Background

The Tocharian script is a north-eastern descendant of Brahmi script, related to Khotanese, Tibetan, and Siddham scripts, which was used along the Tarim Basin in the Taklamakan Desert in what is now Xinjiang in western China.

Tocharian script was historically used to write the Tocharian languages (ISO 639-3: xto, txb), traditionally referred to as Tocharian A and Tocharian B, which belong to the Tocharian branch of the Indo-European language family. The script was in use primarily during the 8th century CE by the people inhabiting the northern edge of the Tarim Basin.

Tocharian script is attested in over 4,000 extant manuscripts that were discovered in the early 20th century in the Tarim Basin, primarily in Kucha, Karasahr, and Turfan.

3 The Issue of Representing Tocharian in Unicode

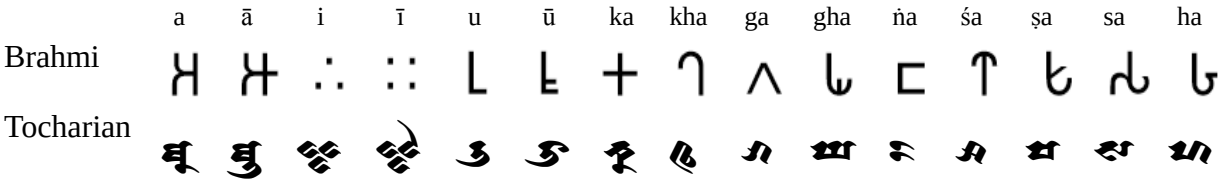
The primary issue facing any proposal to encode Tocharian in the UCS is that of unification with the Brahmi script. While it is true that many sources on Tocharian often refer to its script as "Tocharian Brahmi" or some similar appellation, the Tocharian script nevertheless presents several differences from Brahmi as laid out in the UCS in terms of glyph shapes, character repertoire, and rendering behaviours in particular. With this in mind, there are two main models for representing Tocharian in the UCS:

1. Encoding Tocharian as an independent script
2. Encoding Tocharian as a subset of Brahmi

3.1 Assessment of the Models for Representing Tocharian

Due to the traditional description of Tocharian script as a variant of Brahmi, their similar character repertoire, and the numerous Sanskrit loanwords found in Tocharian, some may argue that Tocharian is simply a regional variation of Brahmi and is accordingly a candidate for being encoded as a subset of Brahmi. In such a case, the distinctive elements of Tocharian would need to managed at the presentation level through fonts and by encoding characters unique to Tocharian as Brahmi extensions. This approach poses problems, which are outlined below:

- *Failure to provide a plain text solution*: The Brahmi script as represented in the UCS is based on Aśokan Brahmi from the 3rd century BCE. The first and most obvious issue facing the encoding of Tocharian as a subset of Brahmi is the visual dissimilarity of Tocharian and Brahmi characters. Nearly all characters in Tocharian are considerably different from their Brahmi counterparts, as illustrated in the following selection of letters:



  Any reader of Brahmi-encoded Tocharian texts would be required to obtain a Brahmi font with character design based on Tocharian in order to read the texts, and would subsequently be unable to view Aśokan Brahmi texts properly. As a result, considering Tocharian as a subset of Brahmi fails to provide a means for plain text representation of the script.

- *Fundamental differences in structure*: Tocharian, while descended from Brahmi, employs several structural forms that are not used in Brahmi, such as stacked aksaras and the bar virama, which will be outlined below. As such, Tocharian script cannot be accurately rendered with Brahmi encoding.

Considering these problems, model 1, encoding Tocharian as an independent script, appears to be the best option.

4 Structure

4.1 Introduction

Tocharian script has typically been referred to as a modified form of Brahmi, indicating that people have traditionally considered this script simply to be a form of Brahmi. Although its structure and functionality is indeed clearly within the Brahmic tradition, the Tocharian script is nevertheless significantly different in a number of ways from the Aśokan Brahmi currently encoded, both in terms of glyph shape and orthographic conventions.

As is typical with Brahmic scripts, each letter indicates a consonant followed by the inherent vowel *a* by default. However, unlike scripts such as Devanagari, there is no visual element that is removed when a letter is used in a conjunct. The vowel is silenced either by a subscript conjunct or the *virāma*.

The most obviously different aspect of Tocharian is the use of eleven consonant signs, traditionally referred to as *Fremdzeichen*, which serve the dual function of representing a consonant plus the vowel *ä*, and to stand in place of the consonant plus virama (the deciding factor of use being the age of the manuscript; later manuscripts do not use *Fremdzeichen* alone to indicate consonant + virama).

Tocharian also employs unique compounding and virama usage which will be explained below.

4.2 Representative glyphs

The fonts used in this document were created by the author and are based on the documents preserved in the International Dunhuang Project.

4.3 Character Names

The characters are named in accordance with the UCS convention for Brahmi-based scripts, with the exception of the vowel AE and EI. The rationale for the spelling AE is that the *Fremdvokal* is traditionally transcribed ä, and ae is the typical replacement for ä in 7-bit ASCII contexts.

4.4 Directionality

The script is written from left to right.

4.5 Vowels

There are 13 independent vowel signs:

 VOWEL LETTER A  VOWEL LETTER UU  VOWEL LETTER O

 VOWEL LETTER AA
 VOWEL LETTER VOCALIC R
 VOWEL LETTER AU
 VOWEL LETTER I
 VOWEL LETTER VOCALIC RR
 VOWEL LETTER AE
 VOWEL LETTER II
 VOWEL LETTER E
 VOWEL LETTER U
 VOWEL LETTER AI

The vowels VOCALIC L and VOCALIC LL are not attested in any Tocharian texts, but spaces have been left available in the code block in case of future discovery.

4.6 Vowel Signs

There are 12 dependent vowel signs:

 VOWEL SIGN AA
 VOWEL SIGN UU
 VOWEL SIGN AI
 VOWEL SIGN I
 VOWEL SIGN VOCALIC R
 VOWEL SIGN O
 VOWEL SIGN II
 VOWEL SIGN VOCALIC RR
 VOWEL SIGN AU
 VOWEL SIGN U
 VOWEL SIGN E
 VOWEL SIGN AE

 VOWEL SIGN AE indicates the vowel /ɨ/ (Krause and Slocum, 2014). The transcription <ä> is standard.

4.7 Consonants

There are 44 consonant letters:

 KA
 TTA
 PA
 SSA
 WAE
 KHA
 TTHA
 PHA
 SA
 SHAE
 GA
 DDA
 BA
 HA
 SSAE
 GHA
 DDHA
 BHA
 BA
 SAE
 NGA
 NNA
 MA
 TAE
 CA
 TA
 YA
 NAE

 CHA  THA  RA  PAE

 JA  DA  LA  MAE

 JHA  DHA  VA  RAE

 NYA  NA  SHA  LAE

All letters bear the inherent vowel *a*. This vowel may be silenced with  VIRAMA or through the use of conjuncts, to be explained below.

Note that the default forms of the letters  TA and  NA are not consistently differentiated in Tocharian manuscripts. However, their combinations with certain vowels and subscript consonants remain distinct, requiring the letters to be encoded separately.

4.8 Various signs

There are 4 various signs:

 ANUSVARA  VISARGA  JIHVAMULIYA  UPADHMANIYA

4.9 Numbers

There are 20 numbers:

	ONE		SIX		TWENTY		SEVENTY
	TWO		SEVEN		THIRTY		EIGHTY
	THREE		EIGHT		FORTY		NINETY
	FOUR		NINE		FIFTY		ONE HUNDRED
	FIVE		TEN		SIXTY		ONE THOUSAND

Numbers for “two hundred” and “three hundred” also exist, but they are transparent combinations of the digit for one hundred and the digits for multiples of one. Numbers beyond 300 are written in horizontal sequences, e.g.  400.

In numbers 11, 21, etc., the number  ONE always stacks vertically, appearing above of the previous number, e.g.  *91*. This only occurs with ONE, all other numbers being formed horizontally, e.g. *92*,  *43*,  *24*, etc.

The numbers 200, 300, and 11, 21, through 91 require special encoding to allow for the modifications above described. There are a number of potential solutions. The first is to encode each one separately, but this seems unnecessary, as they can readily be created through glyph combination. The second option is to employ the virama to merge the letters, but as this solution was proposed and rejected for Brahmi, it is best avoided. The solution is to use a dedicated number joiner. As Brahmi already has a number joiner for this purpose, it could be employed, but perhaps a script-specific number joiner may be the better option. This number joiner would, however, only be used for a total of 11 joined numbers (11, 21, 31, 41, 51, 61, 71, 81, 91, 200, 300), so it may not be useful enough to warrant incorporation as a separate character.

4.10 Vowel signs (matras)

Each vowel letter has a corresponding vowel sign. Vowel signs can be found above, below, or to the right of the consonant letter. Vowel signs that appear below the letter often initiate

changes in the vowel sign, the consonant letter, or both. The vowel signs ◌ AA, ◌ U, and ◌ UU also takes on several contextual forms, and the consonant letter ऌ LA takes on irregular forms.

4.10.1 Contextual forms of vowel signs

AA The vowel sign ◌ AA has various contextual forms, outlined below:

1 When combined with open-topped consonants and certain others:

	ghā	(GHA, vowel sign ◌ AA)
	ñā	(NYA, vowel sign ◌ AA)
	pā	(NYA, vowel sign ◌ AA)
	phā	(NYA, vowel sign ◌ AA)
	mā	(NYA, vowel sign ◌ AA)
	yā	(YA, vowel sign ◌ AA)
	ṣā	(YA, vowel sign ◌ AA)
	sā	(YA, vowel sign ◌ AA)
	hā	(YA, vowel sign ◌ AA)

2 A smaller variant occurs with certain round-topped letters:

	khā	(KHA, vowel sign ◌ AA)
	gā	(GA, vowel sign ◌ AA)
	dhā	(DHA, vowel sign ◌ AA)
	śā	(SHA, vowel sign ◌ AA)

3. A tall superscript form also appears with certain letters:

	ṅā	(NGA, vowel sign ◌ AA)
	jā	(JA, vowel sign ◌ AA)
	ṭā	(TTA, vowel sign ◌ AA)
	ṇā	(NNA, vowel sign ◌ AA)

U/UU The vowel signs ◌ U and ◌ UU have three contextual variations, outlined below:

1. They both take a distinct form on letters that already have descenders that resemble U. This form also appears on DA:


ku (KA, vowel sign U)
jhu (JHA, vowel sign U)
ḍu (DDA, vowel sign U)
du (DA, vowel sign U)
ru (RA, vowel sign U)


kū (KA, vowel sign UU)
jhū (JHA, vowel sign UU)
ḍū (DDA, vowel sign UU)
dū (DA, vowel sign UU)
rū (RA, vowel sign UU)

2. The second form is similar to the first, and only occurs with subscript RA:


pra (PA, RA)
pru (PA, RA, vowel sign U)
prū (PA, RA, vowel sign UU)

3. Forms superficially resembling the independent vowels O and AU appear in combination with certain letters:


tu (TA, vowel sign U)
bhu (BHA, vowel sign U)
gu (GA, vowel sign U)
śu (SHA, vowel sign U)


tū (TA, vowel sign UU)
bhū (BHA, vowel sign UU)
gū (GA, vowel sign UU)
śū (SHA, vowel sign UU)

VOCALIC R AND RR Similar to the vowels U and UU, when these signs attach to a

consonant with a descender, it is deleted:


kṛ (KA, vowel sign VOCALIC R)

jhṛ (JHA, vowel sign VOCALIC R)

ḍṛ (DDA, vowel sign VOCALIC R)


kṝ (KA, vowel sign VOCALIC RR)

jhṝ (JHA, vowel sign VOCALIC RR)

ḍṝ (DDA, vowel sign VOCALIC RR)

Note that these do not occur with RA.

They also cause minor variation in the forms of NA. and BHA:


nṛ (NA, vowel sign VOCALIC R)

bhṛ (BHA, vowel sign VOCALIC R)


nṝ (NA, vowel sign VOCALIC RR)

bhṝ (BHA, vowel sign VOCALIC RR)

LA The consonant letter LA induces a number of irregular vowel sign forms:


li (LA, vowel sign I)


lī (LA, vowel sign II)


le (LA, vowel sign E)


lai (LA, vowel sign AI)


lo (LA, vowel sign O)


lau (LA, vowel sign AU)

I/II/E/AI/O/AU On open topped letters, these vowel signs appear one ascender to the left of the right ascender. In the case of AI, the two elements of the vowel sign appear on different ascenders. Examples:


ghi (GHA, vowel sign I)
pī (PA, vowel sign II)
hī (HA, vowel sign II)
se (SA, vowel sign E)
mai (MA, vowel sign AI)
ṣo (SSA, vowel sign O)
pau (PHA, vowel sign AU)

4.11 Conjuncts

Subscripts are employed to indicate consonant clusters. Most subscripts are relatively transparent and easily identifiable. There are nevertheless some subscripts that differ to a greater or lesser degree from their base forms.

Conjuncts typically comprise between 2 and 4 consonant letters, though there is theoretically no limit:


jña (JA, NYA)
kṣtsa (KA, SSA, TA, SA)

4.11.1 Variation in subscript glyph shapes

YA and RA form subscripts that are entirely dissimilar to their base forms, while VA is also slightly different:


pya (PA, YA)
pra (PA, RA)
pva (PA, VA)

Several other letters gain a supporting bar in subscript form by which they attach to the base letter:


dga (DA, GA)
ṣṭha (SSA, TTHA)
ddha (DA, DHA)
ntha (NA, THA)
wśa (WAE, SHA)

All subscripts with head-like serif element lose it in subscript form:

 *lla* (LA, LA)

 *csa* (CA, SA)

 *kma* (KA, MA)

 *lta* (LA, TA)

The position of subscripts in relation to the base consonants to which they attach is entirely dependent on the specific characters involved. Every base and subscript form has an invariable connection point used in the formation of conjuncts. As a result, some subscripts appear directly below the base, while others appear partially or almost fully to the right:

 *ṣṭha* (SSA, TTHA)

 *kma* (KA, MA)

 *sya* (SA, YA)

 *csa* (CA, SA)

This invariable positioning has consequences for subscript  YA, as it typically extends somewhat above the base of the glyph to which it is attached. When the location of the connection point of the base glyph makes this impossible, the height of subscript YA is truncated. Compare its length and height in the following conjuncts:

 *sya* (SA, YA)

 *kya* (KA, YA)

The conjunct *kka* employs an abbreviated form of the subscript:

 *kka* (KA, KA)

Tocharian has one letter,  *tsa*, which appears frequently and could also be classified as an akhand. It is noteworthy that in this particular conjunct,  TA has a combining form resembling that of NA. This likely arises from *tsa* appearing frequently, and NA having the simpler combining form (see 4.11.2 for discussion of the subscript forms of TA and NA).

4.11.2 Variation in base glyph shapes

Conjuncts can also initiate changes in the form of the base consonant. This is most noticeable in the base conjunct forms of consonant letters with descenders. Just as they lose their

descenders when combining with subscript vowel signs, so do they lose them in consonant conjuncts, e.g.:

 *kla* (KA, LA)
 *ḍva* (DDA, VA)

This also occurs with the letter RA, but with an important difference: namely, that it acts as a typical repha. The form of RA appears above the writing line and attaches to the full base form of a letter:

 *rha* (RA, HA)
 *rṇa* (RA, NNA)

All vowel signs aside from those that attach to the bottom of consonants must attach to the repha:

 *rgo* (RA, GA, vowel sign O)
 *rnā* (RA, NA, vowel sign AA)

The letter LA has an irregular form when it combines with repha RA:

 *rla* (RA, LA)

Repha does not occur with YA; instead, a regular conjunct is formed:

 *rya* (RA, YA)

The letters TA and NA have unique alterations in shape. The alteration in the base form can differentiate the two letters:

 *twa* (TA, WAE)
 *nwa* (NA, WAE)

This is not always a reliable guide, however, as TA occasionally resembles the combining form of NA:

 *tsa* (TA, SA)

As mentioned in 4.11.10, this may best be handled as an akhand ligature.

4.11.3 Aksara conjuncts

Unusually, two consonant signs, each bearing its own a vowel sign, can be combined into a single conjunct. This is most commonly found with the sequence *ku*, which represents the Tocharian consonant /k^{w}/, but it also occurs for metrical rather than phonological reasons (Hitch 2012: 282) (see Figure 4 i, j, k). Examples:

 *ku* (KA, vowel sign U)

 *ce* (CA, vowel sign E)

 *küce*

 *wi* (WAE, vowel sign I)

 *nā* (NA, vowel sign AA)

 *wïnā*

As can be seen, the vowel sign from the subscripted aksara is moved to a more convenient location.

4.12 Virama

There is 1 virama:

 VIRAMA

Tocharian employs a form of the virama that functions exactly as viramas in other Indic scripts. However, it also employs a second, far more commonly-occurring form of virama that appears visually as a horizontal or diagonal bar that precedes the marked letter or conjunct and connects it with the preceding, vowel-bearing letter or conjunct. The distinction between the two virama is mostly context-based, as, typically, *Fremdzeichen* take the bar virama while standard letters take the standard virama, though there are some exceptions (see Figure 4).

Example:

 *śal* (SHA, LAE, VIRAMA)

This also occurs with final consonant clusters that include a *Fremdzeichen*:

 *lkānt* (LA, VIRAMA, KA NA, VIRAMA, TAE, VIRAMA)

It is important to note that the bar virama can attach to any portion of the previous aksara, including the base consonant, subscript, or vowel sign.

Occasionally, a consonant may bear a redundant standard virama in addition to the bar virama:

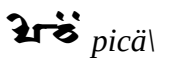
 *mor* (MA, vowel sign O, RAE, VIRAMA, VIRAMA)

This is, however, optional; the redundant standard virama is typically absent.

The bar virama should be treated as an alternate form of the standard virama. Though it appears before the letter it modifies, this is common in Brahmic scripts (cf. vowel signs in Devanagari, Thai, etc.). Bar virama is the standard form used with *Fremdzeichen*, while the standard virama appears on regular consonants. The exception to this are the letters CA and NYA, which lack *Fremdzeichen* variants.

 *teñ* (TA, NYA, VIRAMA)

Occasionally, the vowel sign AE will appear on a letter carrying a bar virama:

 *picä* (PA, vowel sign I, CA, VIRAMA, vowel sign AE)

This is largely restricted to CA and NYA, but does occur on some other letters as well:

 *koyä* (KA, vowel sign O, YA, VIRAMA, vowel sign AE)

The proposed implementation is:

- *Fremdzeichen* with virama: virama is realized as bar virama
- *Fremdzeichen* with two viramas: first virama is realized as bar virama, second as standard virama

- standard consonant with virama: virama is realized as standard virama
- standard consonant with virama and vowel sign AE: virama is realized as bar virama.

4.13 Subscript Independent Vowel Letters

In distinct contrast to most Brahmic scripts (but with precedent in e.g. Khmer), Tocharian indicates some diphthongs through the use of subscript independent vowel signs, which are also necessarily marked with virama (see Figure 7 e, f, g, h).

If the base letter has a subscript, the virama is straight and attaches to the subscript. If it does not, the virama angles up to attach to the base consonant. Examples:

 *klye-u* (KA, LA, YA, vowel sign E U)

 *lo-i* (LA, vowel sign O, I)

Notice that the subscript I takes a different form.

4.14 Nasalization

The languages do not have nasalization *per se*, but the script nevertheless employ *anusvāra* both for nasal consonants and for transcription of Sanskrit nasalization. It appears immediately above the base consonant letter.

 *ghaṃ* (GHA, ANUSVARA)

 *ṅkmāṃ* (NGA, KA, MA, vowel sign AA, ANUSVARA)

4.15 Aspiration

Tocharian employs three signs for aspiration: the *visarga* sign, which appears to the right of the base consonant sign, and the *jihvāmūlīya* and *upadhmānīya*, which respectively indicate velar and labial allophones of *h*. These differ from visarga in that they act as letters and form conjuncts with the preceding consonant letter.

 *khaḥ* (KHA, VISARGA)

 *ẖka* (JIHVAMULIYA, KA)

 *ḫpa* (UPADHMANIYA, PA)

In Tocharian texts, VISARGA is relatively common, but JIHVAMULIYA and UPADHMANIYA are exceedingly rare.

4.16 Punctuation

There are four punctuation marks:

 DANDA

 DOUBLE DANDA

 PUNCTUATION DOT

 PUNCTUATION DOUBLE DOT

5 Character data

5.1 Character Properties

Tocharian character properties are as follows:

```
11E00;TOCHARIAN LETTER A;Lo;0;L;;;;;N;;;;;
11E01;TOCHARIAN LETTER AA;Lo;0;L;;;;;N;;;;;
11E02;TOCHARIAN LETTER I;Lo;0;L;;;;;N;;;;;
11E03;TOCHARIAN LETTER II;Lo;0;L;;;;;N;;;;;
11E04;TOCHARIAN LETTER U;Lo;0;L;;;;;N;;;;;
11E05;TOCHARIAN LETTER UU;Lo;0;L;;;;;N;;;;;
11E06;TOCHARIAN LETTER VOCALIC R;Lo;0;L;;;;;N;;;;;
11E07;TOCHARIAN LETTER VOCALIC RR;Lo;0;L;;;;;N;;;;;
11E08;<RESERVED>
11E09;<RESERVED>
11E0A;TOCHARIAN LETTER E;Lo;0;L;;;;;N;;;;;
11E0B;TOCHARIAN LETTER AI;Lo;0;L;;;;;N;;;;;
11E0C;TOCHARIAN LETTER O;Lo;0;L;;;;;N;;;;;
11E0D;TOCHARIAN LETTER AU;Lo;0;L;;;;;N;;;;;
11E0E;TOCHARIAN LETTER AE;Lo;0;L;;;;;N;;;;;
11E0F;<THIS POSITION SHALL NOT BE USED>
11E10;TOCHARIAN LETTER KA;Lo;0;L;;;;;N;;;;;
11E11;TOCHARIAN LETTER KHA;Lo;0;L;;;;;N;;;;;
11E12;TOCHARIAN LETTER GA;Lo;0;L;;;;;N;;;;;
11E13;TOCHARIAN LETTER GHA;Lo;0;L;;;;;N;;;;;
11E14;TOCHARIAN LETTER NGA;Lo;0;L;;;;;N;;;;;
11E15;TOCHARIAN LETTER CA;Lo;0;L;;;;;N;;;;;
11E16;TOCHARIAN LETTER CHA;Lo;0;L;;;;;N;;;;;
11E17;TOCHARIAN LETTER JA;Lo;0;L;;;;;N;;;;;
11E18;TOCHARIAN LETTER JHA;Lo;0;L;;;;;N;;;;;
11E19;TOCHARIAN LETTER NYA;Lo;0;L;;;;;N;;;;;
11E1A;TOCHARIAN LETTER TTA;Lo;0;L;;;;;N;;;;;
11E1B;TOCHARIAN LETTER TTHA;Lo;0;L;;;;;N;;;;;
11E1C;TOCHARIAN LETTER DDA;Lo;0;L;;;;;N;;;;;
11E1D;TOCHARIAN LETTER DDHA;Lo;0;L;;;;;N;;;;;
11E1E;TOCHARIAN LETTER NNA;Lo;0;L;;;;;N;;;;;
11E1F;TOCHARIAN LETTER TA;Lo;0;L;;;;;N;;;;;
11E20;TOCHARIAN LETTER THA;Lo;0;L;;;;;N;;;;;
11E21;TOCHARIAN LETTER DA;Lo;0;L;;;;;N;;;;;
11E22;TOCHARIAN LETTER DHA;Lo;0;L;;;;;N;;;;;
11E23;TOCHARIAN LETTER NA;Lo;0;L;;;;;N;;;;;
11E24;TOCHARIAN LETTER PA;Lo;0;L;;;;;N;;;;;
11E25;TOCHARIAN LETTER PHA;Lo;0;L;;;;;N;;;;;
11E26;TOCHARIAN LETTER BA;Lo;0;L;;;;;N;;;;;
11E27;TOCHARIAN LETTER BHA;Lo;0;L;;;;;N;;;;;
11E28;TOCHARIAN LETTER MA;Lo;0;L;;;;;N;;;;;
11E29;TOCHARIAN LETTER YA;Lo;0;L;;;;;N;;;;;
11E2A;TOCHARIAN LETTER RA;Lo;0;L;;;;;N;;;;;
11E2B;TOCHARIAN LETTER LA;Lo;0;L;;;;;N;;;;;
11E2C;TOCHARIAN LETTER VA;Lo;0;L;;;;;N;;;;;
11E2D;TOCHARIAN LETTER SHA;Lo;0;L;;;;;N;;;;;
11E2E;TOCHARIAN LETTER SSA;Lo;0;L;;;;;N;;;;;
11E2F;TOCHARIAN LETTER SA;Lo;0;L;;;;;N;;;;;
11E30;TOCHARIAN LETTER HA;Lo;0;L;;;;;N;;;;;
11E31;TOCHARIAN VOWEL SIGN AA;Mc;0;L;;;;;N;;;;;
11E32;TOCHARIAN VOWEL SIGN I;Mn;0;NSM;;;;;N;;;;;
11E33;TOCHARIAN VOWEL SIGN II;Mn;0;NSM;;;;;N;;;;;
11E34;TOCHARIAN VOWEL SIGN U;Mn;0;NSM;;;;;N;;;;;
11E35;TOCHARIAN VOWEL SIGN UU;Mn;0;NSM;;;;;N;;;;;
11E36;TOCHARIAN VOWEL SIGN VOCALIC R;Mn;0;NSM;;;;;N;;;;;
11E37;TOCHARIAN VOWEL SIGN VOCALIC RR;Mn;0;NSM;;;;;N;;;;;
11E38;<RESERVED>
11E39 <RESERVED>
11E3A;TOCHARIAN VOWEL SIGN E;Mn;0;NSM;;;;;N;;;;;
```

```
11E3B;TOCHARIAN VOWEL SIGN AI;Mn;0;NSM;;;;;N;;;;;
11E3C;TOCHARIAN VOWEL SIGN O;Mn;0;NSM;;;;;N;;;;;
11E3D;TOCHARIAN VOWEL SIGN AU;Mn;0;NSM;;;;;N;;;;;
11E3E;TOCHARIAN VOWEL SIGN AE;Mn;0;NSM;;;;;N;;;;;
11E3F;<THIS POSITION SHALL NOT BE USED>
11E40;TOCHARIAN LETTER KAE;Lo;0;L;;;;;N;;;;;
11E41;TOCHARIAN LETTER TAE;Lo;0;L;;;;;N;;;;;
11E42;TOCHARIAN LETTER NAE;Lo;0;L;;;;;N;;;;;
11E43;TOCHARIAN LETTER PAE;Lo;0;L;;;;;N;;;;;
11E44;TOCHARIAN LETTER MAE;Lo;0;L;;;;;N;;;;;
11E45;TOCHARIAN LETTER RAE;Lo;0;L;;;;;N;;;;;
11E46;TOCHARIAN LETTER LAE;Lo;0;L;;;;;N;;;;;
11E47;TOCHARIAN LETTER WAE;Lo;0;L;;;;;N;;;;;
11E48;TOCHARIAN LETTER SHAE;Lo;0;L;;;;;N;;;;;
11E49;TOCHARIAN LETTER SSAE;Lo;0;L;;;;;N;;;;;
11E4A;TOCHARIAN LETTER SAE;Lo;0;L;;;;;N;;;;;
11E4B;TOCHARIAN SIGN ANUSVARA;Mn;0;NSM;;;;;N;;;;;
11E4C;TOCHARIAN SIGN VISARGA;Mc;0;L;;;;;N;;;;;
11E4D;TOCHARIAN SIGN JIHVAMULIYA;Lo;0;L;;;;;N;;;;;
11E4E;TOCHARIAN SIGN UPADHMANIYA;Lo;0;L;;;;;N;;;;;
11E4F;TOCHARIAN VIRAMA;Mn;9;L;;;;;N;;;;;
11E50;TOCHARIAN NUMBER ONE;No;0;L;;;;1;N;;;;;
11E51;TOCHARIAN NUMBER TWO;No;0;L;;;;2;N;;;;;
11E52;TOCHARIAN NUMBER THREE;No;0;L;;;;3;N;;;;;
11E53;TOCHARIAN NUMBER FOUR;No;0;L;;;;4;N;;;;;
11E54;TOCHARIAN NUMBER FIVE;No;0;L;;;;5;N;;;;;
11E55;TOCHARIAN NUMBER SIX;No;0;L;;;;6;N;;;;;
11E56;TOCHARIAN NUMBER SEVEN;No;0;L;;;;7;N;;;;;
11E57;TOCHARIAN NUMBER EIGHT;No;0;L;;;;8;N;;;;;
11E58;TOCHARIAN NUMBER NINE;No;0;L;;;;9;N;;;;;
11E59;TOCHARIAN NUMBER TEN;No;0;L;;;;10;N;;;;;
11E5A;TOCHARIAN NUMBER TWENTY;No;0;L;;;;20;N;;;;;
11E5B;TOCHARIAN NUMBER THIRTY;No;0;L;;;;30;N;;;;;
11E5C;TOCHARIAN NUMBER FORTY;No;0;L;;;;40;N;;;;;
11E5D;TOCHARIAN NUMBER FIFTY;No;0;L;;;;50;N;;;;;
11E5E;TOCHARIAN NUMBER SIXTY;No;0;L;;;;60;N;;;;;
11E5F;TOCHARIAN NUMBER SEVENTY;No;0;L;;;;70;N;;;;;
11E60;TOCHARIAN NUMBER EIGHTY;No;0;L;;;;80;N;;;;;
11E61;TOCHARIAN NUMBER NINETY;No;0;L;;;;90;N;;;;;
11E62;TOCHARIAN NUMBER ONE HUNDRED;No;0;L;;;;100;N;;;;;
11E63;TOCHARIAN NUMBER ONE THOUSAND;No;0;L;;;;1000;N;;;;;
11E64;TOCHARIAN DANDA;Po;0;L;;;;;N;;;;;
11E65;TOCHARIAN DOUBLE DANDA;Po;0;L;;;;;N;;;;;
11E66;TOCHARIAN PUNCTUATION DOT;Po;0;L;;;;;N;;;;;
11E67;TOCHARIAN PUNCTUATION DOUBLE DOT;Po;0;L;;;;;N;;;;;
```

5.2 Syllabic Categories

```
# Indic_Syllabic_Category=Vowel_Independent
11E00..11E0E    ; Vowel_Independent  # Lo [13] TOCHARIAN LETTER A..LETTER AE

# Indic_Syllabic_Category=Consonant
11E10..11E30    ; Consonant               # Lo [33] LETTER KA..LETTER HA
11E40..11E4A    ; Consonant               # Lo [11] LETTER KAE..LETTER SAE

# Indic_Syllabic_Category=Vowel_Dependent
11E31           ; Vowel_Dependent         # Mc [1] SIGN AA
11E32..11E3E    ; Vowel_Dependent         # Mn [11] SIGN I..SIGN AE

# Indic_Syllabic_Category=Bindu
11E4B           ; Bindu                   # Mn [1] SIGN ANUSVARA

# Indic_Syllabic_Category=Visarga
11E4C           ; Visarga                 # Mc [1] SIGN VISARGA
```

```
# Indic_Syllabic_Category=Consonant_With_Stacker
1134D..11E4E    ; Consonant_With_Stacker  # Lo [2] SIGN JIHVAMULIYA..SIGN UPADHMANIYA

# Indic_Syllabic_Category=Virama
1134F           ; Virama                  # Mn [1] SIGN VIRAMA

# Indic_Syllabic_Category=Number
11E50..11E63    ; Number                  # Lo [20] NUMBER ONE..NUMBER ONE THOUSAND
```

5.3 Positional Categories

```
# Indic_Positional_Category=Right
11E31           ; Right    # Mc [1] SIGN AA

# Indic_Positional_Category=Top
11E32..11E33    ; Top      # Mn [2] SIGN I..SIGN II
11E3A..11E3E    ; Top      # Mn [5] SIGN E..SIGN AE

# Indic_Positional_Category=Bottom
11E34..11E37    ; Bottom   # Mn [4] SIGN U..SIGN VOCALIC RR
```

6 Code charts

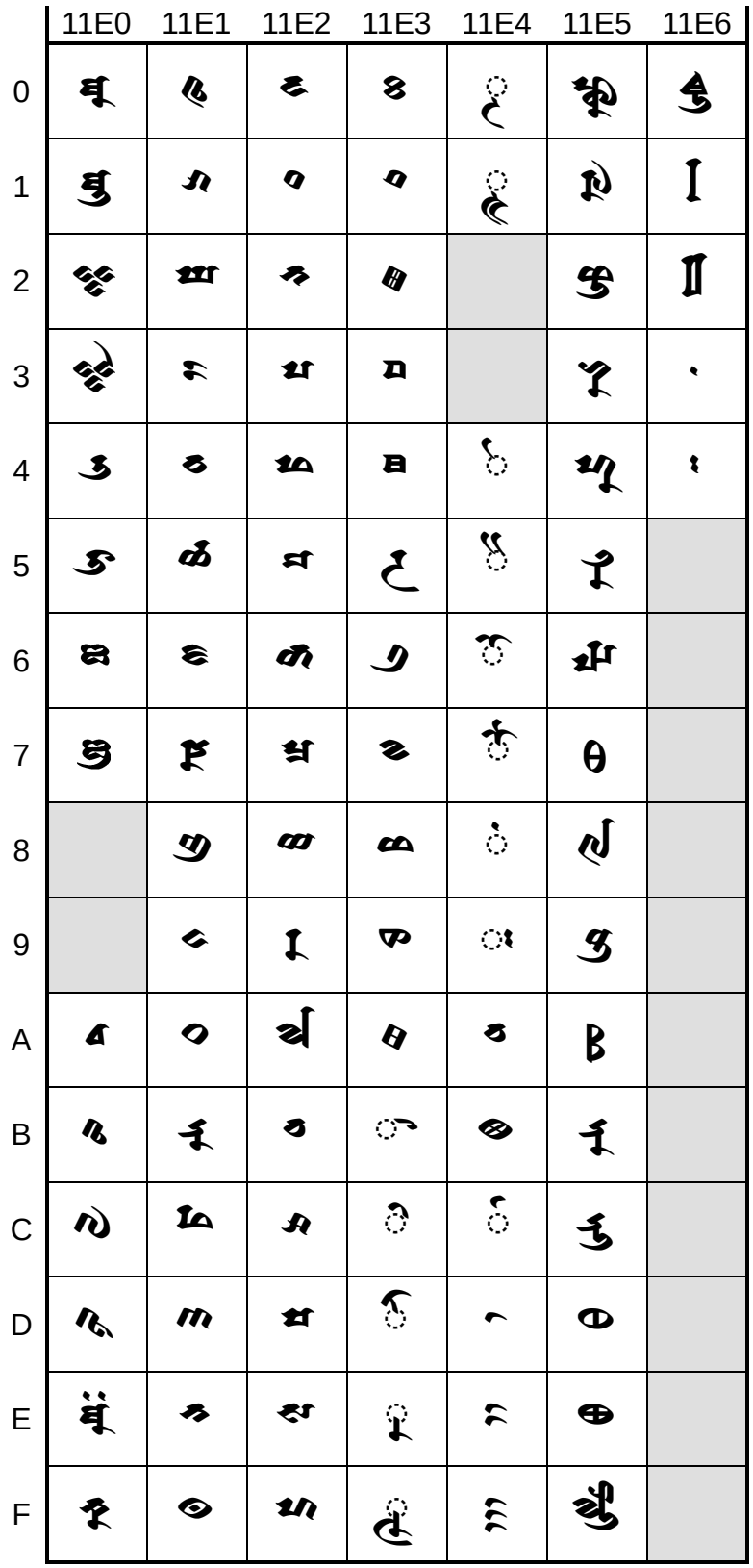


Figure 1: Proposed code chart for Tocharian

Independent vowels

11E00 TOCHARIAN LETTER A
11E01 TOCHARIAN LETTER AA
11E02 TOCHARIAN LETTER I
11E03 TOCHARIAN LETTER II
11E04 TOCHARIAN LETTER U
11E05 TOCHARIAN LETTER UU
11E06 TOCHARIAN LETTER VOCALIC R
11E07 TOCHARIAN LETTER VOCALIC RR
11E08 <RESERVED>
11E09 <RESERVED>
11E0A TOCHARIAN LETTER E
11E0B TOCHARIAN LETTER AI
11E0C TOCHARIAN LETTER O
11E0D TOCHARIAN LETTER AU
11E0E TOCHARIAN LETTER AE

Consonants

11E10 TOCHARIAN LETTER KA
11E11 TOCHARIAN LETTER KHA
11E12 TOCHARIAN LETTER GA
11E13 TOCHARIAN LETTER GHA
11E14 TOCHARIAN LETTER NGA
11E15 TOCHARIAN LETTER CA
11E16 TOCHARIAN LETTER CHA
11E17 TOCHARIAN LETTER JA
11E18 TOCHARIAN LETTER JHA
11E19 TOCHARIAN LETTER NYA
11E1A TOCHARIAN LETTER TTA
11E1B TOCHARIAN LETTER TTHA
11E1C TOCHARIAN LETTER DDA
11E1D TOCHARIAN LETTER DDHA
11E1E TOCHARIAN LETTER NNA
11E1F TOCHARIAN LETTER TA
11E20 TOCHARIAN LETTER THA
11E21 TOCHARIAN LETTER DA
11E22 TOCHARIAN LETTER DHA
11E23 TOCHARIAN LETTER NA
11E24 TOCHARIAN LETTER PA
11E25 TOCHARIAN LETTER PHA
11E26 TOCHARIAN LETTER BA
11E27 TOCHARIAN LETTER BHA
11E28 TOCHARIAN LETTER MA
11E29 TOCHARIAN LETTER YA
11E2A TOCHARIAN LETTER RA
11E2B TOCHARIAN LETTER LA
11E2C TOCHARIAN LETTER VA
11E2D TOCHARIAN LETTER SHA
11E2E TOCHARIAN LETTER SSA
11E2F TOCHARIAN LETTER SA
11E30 TOCHARIAN LETTER HA

Dependent vowel signs

11E31 TOCHARIAN SIGN AA
11E32 TOCHARIAN SIGN I
11E33 TOCHARIAN SIGN II
11E34 TOCHARIAN SIGN U
11E35 TOCHARIAN SIGN UU
11E36 TOCHARIAN SIGN VOCALIC R
11E37 TOCHARIAN SIGN VOCALIC RR
11E38 <RESERVED>
11E39 <RESERVED>
11E3A TOCHARIAN SIGN E
11E3B TOCHARIAN SIGN AI
11E3C TOCHARIAN SIGN O
11E3D TOCHARIAN SIGN AU

Special Consonants

11E40 TOCHARIAN LETTER KAE
11E41 TOCHARIAN LETTER TAE
11E42 TOCHARIAN LETTER NAE
11E43 TOCHARIAN LETTER PAE
11E44 TOCHARIAN LETTER MAE
11E45 TOCHARIAN LETTER RAE
11E46 TOCHARIAN LETTER LAE
11E47 TOCHARIAN LETTER WAE
11E48 TOCHARIAN LETTER SHAE
11E49 TOCHARIAN LETTER SSAE
11E4A TOCHARIAN LETTER SAE

Various signs

11E4B TOCHARIAN SIGN ANUSVARA
11E4C TOCHARIAN SIGN VISARGA
11E4D TOCHARIAN SIGN JIHVAMULIYA
11E4E TOCHARIAN SIGN UPADHMANIYA

Virama

11E4E TOCHARIAN VIRAMA

Numbers

11E50 TOCHARIAN NUMBER ONE
11E51 TOCHARIAN NUMBER TWO
11E52 TOCHARIAN NUMBER THREE
11E53 TOCHARIAN NUMBER FOUR
11E54 TOCHARIAN NUMBER FIVE
11E55 TOCHARIAN NUMBER SIX
11E56 TOCHARIAN NUMBER SEVEN
11E57 TOCHARIAN NUMBER EIGHT
11E58 TOCHARIAN NUMBER NINE
11E59 TOCHARIAN NUMBER TEN
11E5A TOCHARIAN NUMBER TWENTY
11E5B TOCHARIAN NUMBER THIRTY
11E5C TOCHARIAN NUMBER FORTY
11E5D TOCHARIAN NUMBER FIFTY
11E5E TOCHARIAN NUMBER SIXTY
11E5F TOCHARIAN NUMBER SEVENTY
11E60 TOCHARIAN NUMBER EIGHTY
11E61 TOCHARIAN NUMBER NINETY
11E62 TOCHARIAN NUMBER ONE HUNDRED
11E63 TOCHARIAN NUMBER ONE THOUSAND

Punctuation

11E64 TOCHARIAN PUNCTUATION DANDA
11E65 TOCHARIAN PUNCTUATION DOUBLE DANDA

11E66 · TOCHARIAN PUNCTUATION DOT
11E67 ⁝ TOCHARIAN PUNCTUATION DOUBLE DOT

Figure 2: Proposed names list for Tocharian

7 Samples

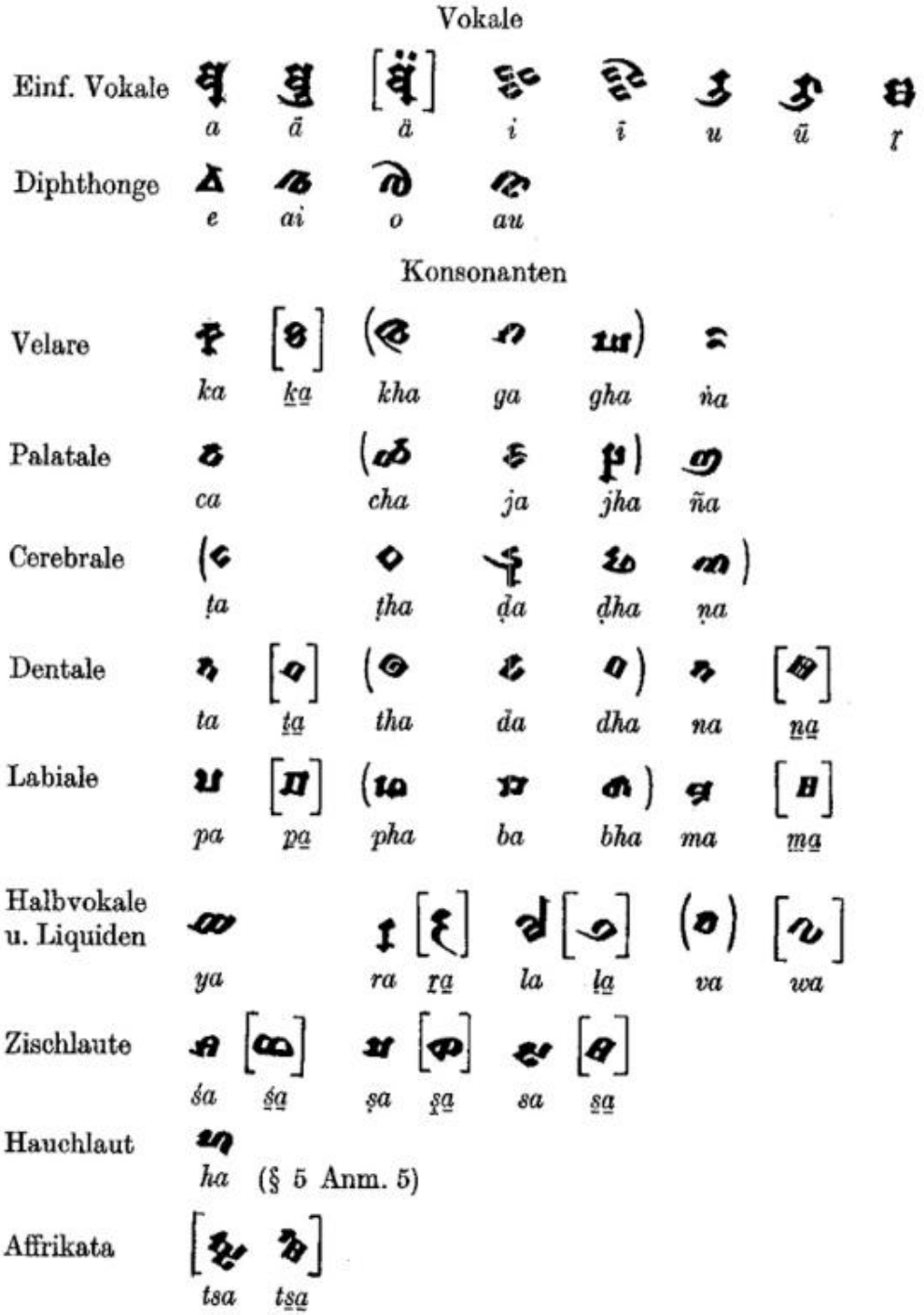


Figure 3: A table of the basic letters of Tocharian (from Krause and Thomas 1960:41, Malzahn 2007b:227-8).

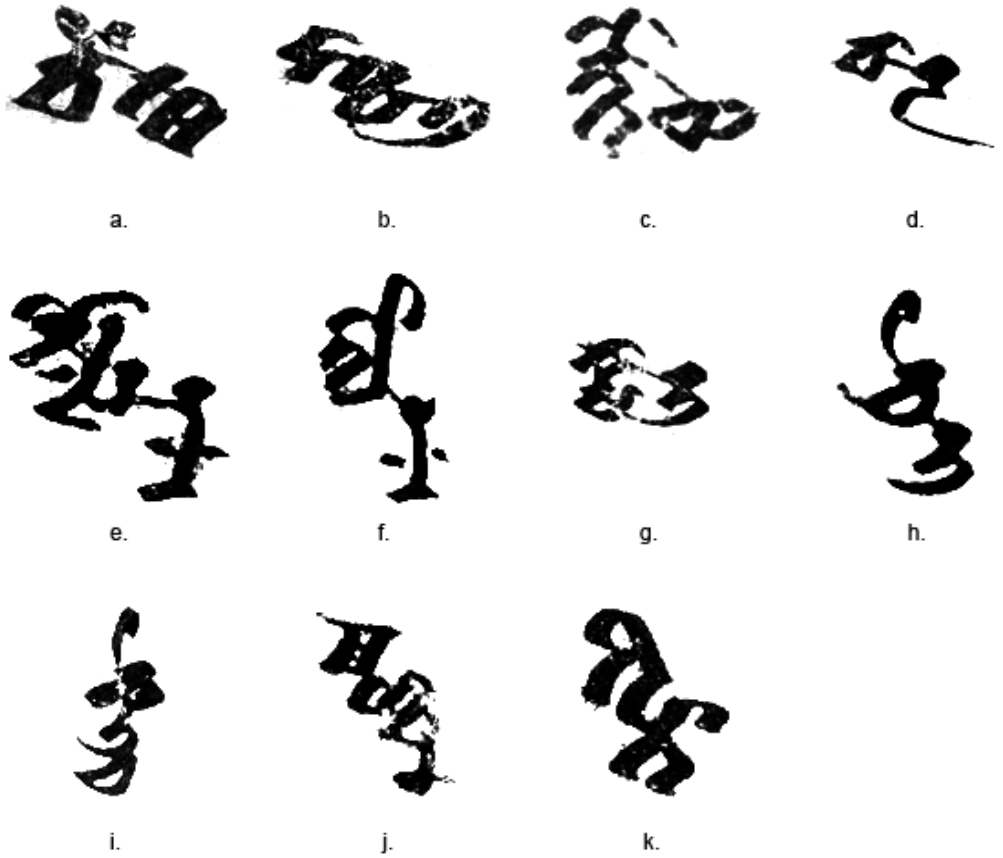


Figure 4: examples of bar virama, subscript independent vowel letters, and stacked aksaras. a. *ceṃts*, b. *ṅk̲äl̲*, c. *ttos̲*, d. *cār̲*, e. *ssoî*, f. *loî*, g. *ksāû*, h. *ceû*, i *küce*, j. *mañcu*, k. *winā*.

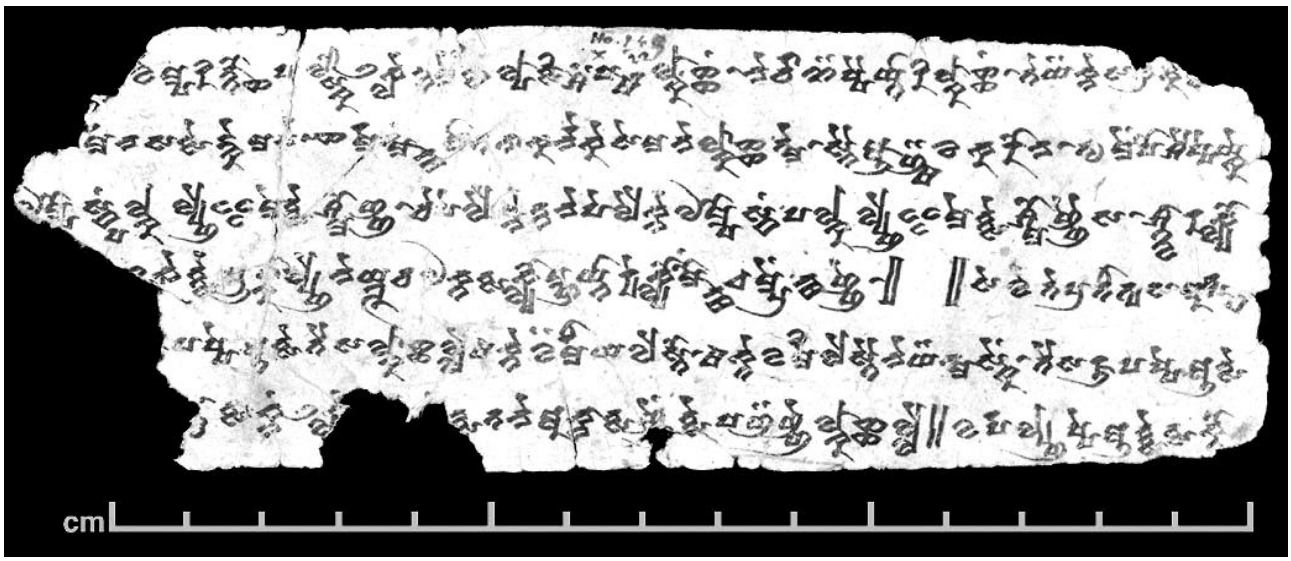


Figure 5: Original Tocharian manuscript showing natural text (from International Dunhuang Project).

Figure 6: Original Tocharian manuscript displaying a list of velar and palatal conjuncts.

Figure 7: Original Tocharian manuscript displaying a list of palatal and retroflex conjuncts.

Figure 8: Original Tocharian manuscript displaying a list of dental and bilabial conjuncts.

Figure 9: Original Tocharian manuscript displaying a list of bilabial, liquid, and fricative conjuncts.

Figure 10: Original Tocharian manuscript displaying a list of bilabial, liquid, and fricative conjuncts.

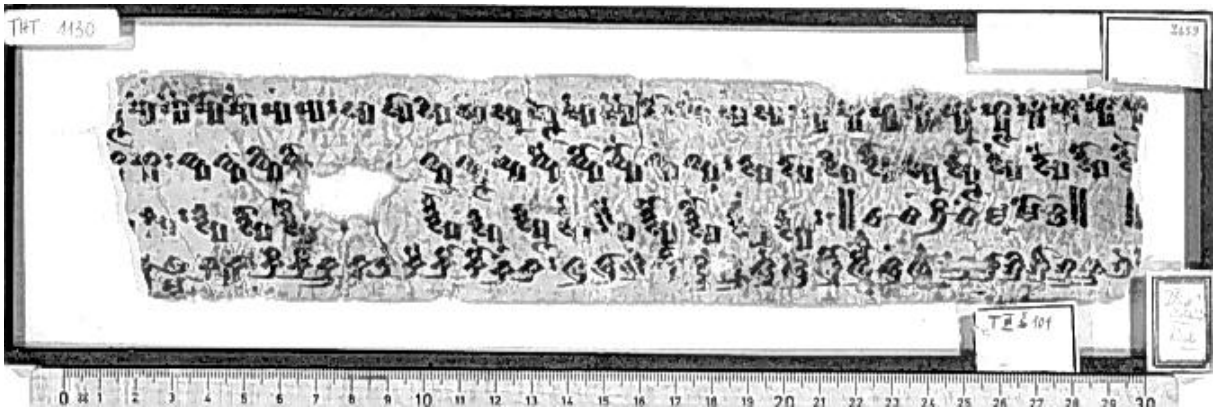


Figure 11: Original Tocharian manuscript displaying a list of fricative, affricative, and velar conjuncts.

8 References

Hitch, Doug. “Review: Melanie Malzahn (ed.) *Instrumenta Tocharica*” In *Tocharian and Indo-European Studies*, vol 13. Jens Elmegård Rasmussen, Michaël Peyrot, Thomas Olander, eds. 2012, pp 277-290. Copenhagen: Museum Tusculanum Press.

International Dunhuang Project. http://idp.bl.uk/. Accessed August 2014.

Krause, Todd B., Jonathan Slocum. Tocharian Online. May 13, 2014. http://www.utexas.edu/cola/centers/lrc/eieol/tokol-1-X.html. Accessed July, 2014.

Krause, Wolfgang, Werner Thomas. *Tocharisches Elementarbuch*.1960. Heidelberg: Carl Winter - Universitätsverlag.

Mladjov, Ian. “Tang China” http://www.historyandcivilization.com/Maps---Tables---Chinese-History.html. Accessed July 2014.

TITUS. “The Tocharian “Alphabet””. http://titus.fkidg1.uni-frankfurt.de/didact/idg/toch/tochbr.htm. Accessed July 2014.

---. “Tocharian Manuscripts”. http://titus.fkidg1.uni-frankfurt.de/texte/tocharic/tht.htm. Accessed August 2014.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646[1]
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:** *Preliminary Proposal to Encode the Tocharian Script*
2. Requester's name: *Lee Wilson (ttlwilson@hotmail.com)*
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *2014-10-09*
5. Requester's reference (if applicable):
6. Choose one of the following:
   This is a complete proposal: *yes*
   (or) More information will be provided later:

B. Technical – General

1. Choose one of the following:
   a. This proposal is for a new script (set of characters): *yes*
      Proposed name of script: *Tocharian*
   b. The proposal is for addition of character(s) to an existing block:
      Name of the existing block:
2. Number of characters in proposal: *97*
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 X ; 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? *Yes*
   a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document? *Yes*
   b. Are the character shapes attached in a legible form suitable for review? *Yes*
5. Fonts related:
   a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?
   Lee Wilson (TrueType or OpenType format)
   b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):
   Lee Wilson (ttlwilson@hotmail.com)
6. References:
   a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided? *Yes*
   b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached? *No*
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)?
   No

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? *No*
   If YES explain
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.)? *n/a*
   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? *extinct*
   Reference:
4. The context of use for the proposed characters (type of use; common or rare) *rare*
   Reference:
5. Are the proposed characters in current use by the user community? *No*
   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?
   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)? *Yes*
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence? *No*
   If YES, is a rationale for its inclusion provided?
   If YES, reference:
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? *No*
    If YES, is a rationale for its inclusion provided?
    If YES, reference:
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? *Yes*
    If YES, reference: *Combining signs*
    Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *No*
    If YES, reference:
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) *Virama*
    see proposal for details
13. Does the proposal contain any Ideographic compatibility characters? *No*
    If YES, are the equivalent corresponding unified ideographic characters identified?
    If YES, reference: