Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type:	**Working Group Document**
Title:	**Comments on the Tai Don or Tai Khao encoding**
Source:	**Eiso Chan (**陈永聪**)**
Status:	**Individual Contribution to UTC**
Action:	**For consideration by UTC**
Date:	**2022-05-10**

0. Introduction

Kushim Jiang submits documents for encoding the Tai Don or Tai Khao script to UCS and Unicode. (L2/20-207, L2/22-098) In fact, there are several scripts used for Tai/Dai languages in China. This one is related to the encoded Tai Viet script according to the introductions and analyses by Kushim and Jim Brase. So, the main issue is if it's suitable to encode Tai Don or Tai Khao script in a separate block.

According to an old Vietnamese document named *Bảng chữ Thái các vùng và chữ thống nhất* (榜哛傣各墥姄哷統一, *Character table of Tai spoken area and the Unified Tai character*). It looks the document was published by Tây Bắc Autonomous Region (Khu tự trị Tây Bắc/區自治西北), and this region includes current Lai Chau Province (Tỉnh Lai Châu/省莱州), Son La Province (Tỉnh Sơn La/省山羅), some part of Yen Bai Province (Tỉnh Yên Bái/省安沛) and so on. The table shows five kinds of letterforms and the Unified Tai letterforms, which includes the letterforms from Thái Đen, Mường Lay, Phong Thổ, Phù Yên and Mộc Châu. Phong Thổ is a county under Lai Chau now. Please see Appendix A in this document.

Mr. Việt and Jim submitted L2/06-041 for encoding the Unified Tai script, and Jim submitted L2/07-039R and WG2 N3220 (L2/07-099). The Tai Viet script was adopted as TCVN 8271-6:2010 "Information Technology - Encoded Character Set - Part 6: Tai Viet Script". The current Tai Viet script is suitable to write Vietnamese Tai Dam and Tai Don languages.

Now the Tai people in Việt Nam uses the Tai Viet script and the Lattin script to write their languages at the same time. I find two Romanization methods on it. The Tai Don or Tai Khao script is used in Jinping County, Yunnan Province, PRC commonly. Please click here.

1. Mapping

Kushim kindly provides a mapping table between Tai Don or Tai Khao and Tai Viet in Section 3 in L2/22-098. Almost all the table is credible and acceptable, but there are some issues I need to mention here.

1.1. HIGH XO (0xXX0A) and HIGH KHO (0xXX0C)

0xXX0A ꪊ glyph looks very different from U+AA84 ꪄ glyph, but Việt Nam had unified them in 1955 according to *Bảng chữ Thái các vùng và chữ thống nhất*.

Fig. 1.1 Entry 25 of the consonants list in Bảng chữ Thái các vùng và chữ thống nhất

For 0xXX0C ⟨ and U+AA82 ⟨, *Bảng chữ Thái các vùng và chữ thống nhất* didn't provide more useful information, but we can know the main body of 0xXX0C ⟨ is related to 0xXX0A ⟨, and the main body of U+AA82 ⟨ is also related to U+AA84 ⟨.

1.2. LOW XO (0xXX0B)

I need to mention one related issue first. The current code chart includes U+AA86 ⟨, U+AA87 ⟨, U+AAA6 ⟨, and U+AAA7 ⟨, but they are not used in the common words in the Vietnamese Tai textbook, and I can't find the similar letters in Tai Tam and New Tai Lue scripts. Are they only used for the loanword? The following table shows the glyphs for these four code points in L2/06-041, current code chart, the font Oswald, the font Tai Heritage Pro and the font Noto Sans Tai Viet. Oswald is a common font used by the Tai people in Việt Nam (I guess Oswald here means the Tai people in Việt Nam developed a new font for Tai Viet script, not the font designed by [Vernon Adams](#)); Tai Heritage Pro is developed by SIL and encapsulated by Graphite; Noto Sans Tai Viet is a part of the great Noto project, which the font is designed by [Monotype](#).

Table 1.2.1 Glyph comparison

	Submitted	Current	Oswald	SIL	Noto
U+AA86					
U+AA87					
U+AAA6					
U+AAA7					

It looks the submitted U+AA86 and U+AA87 glyphs are derived from U+0047 G, the ones in the current code chart and Tai Heritage Pro and Noto Sans Tai Viet are derived from U+0067 g, U+AA86 in Oswald is the same as U+AA85 in Phong Thổ (please see Fig. 1.2), U+AA87 in Oswald is derived from U+AA85 ⟨.

If the glyphs for U+AA86 and U+AA87 could be confirmed by the Tai people in Việt Nam, is it better to map 0xXX0B ⟨ to U+AA86? For overall consideration, 0xXX0B ⟨ glyph is unified to U+AA85 in 1955, so it is also acceptable for the current mapping.
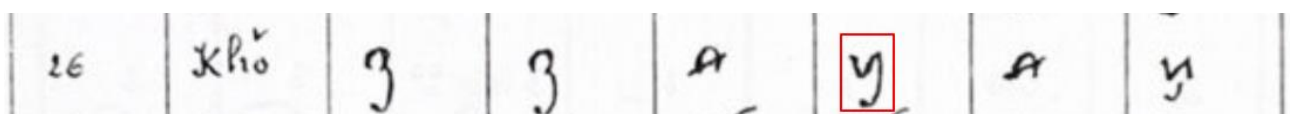

Fig. 1.2 Entry 26 of the consonants list in Bảng chữ Thái các vùng và chữ thống nhất

For U+AAA6 and U+AAA7, it looks the submitted glyphs are derived from U+0053 S or U+0073 s, and the ones in current code chart, Oswald, Tai Heritage Pro and Noto Sans Tai Viet are derived from U+0052 R. This pair is not used in Jinping at all.

Please also see Appendix B.

1.3. HIGH HO (0xXX0E) and related letters

In *Bảng chữ Thái các vùng và chữ thống nhất*, U+AAAC ꪬ glyph is common in Việt Nam, but it's a pair with U+AAAD ꪭ. 0xXX0E ꫛ and 0xXX0F ꪍ are also a pair, and 0xXX0F ꪍ is used in Phong Thổ and unified to U+AAAD ꪭ in 1955. The current mapping for 0xXX0E is acceptable.
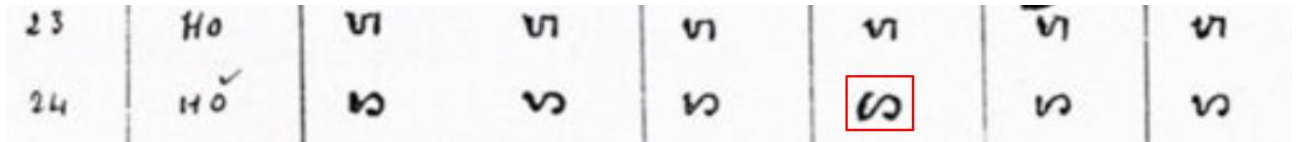
Fig 1.3.1 Entries 23 & 24 of the consonants list in Bảng chữ Thái các vùng và chữ thống nhất

The Tai textbook used in Jinping shows a picture on the relationship among HIGH HO ꫛ (0xXX0E) and HIGH NGO ꪂ (0xXX24), HIGH YO ꪑ (0xXX28), HIGH NO ꪁ (0xXX1E), HIGH MO ꪀ (0xXX20), HIGH LO ꪩ (0xXX22), HIGH VO ꪬ (0xXX26), HIGH NYO ꪒ (0xXX2A). Please see Fig. 1.3.2.
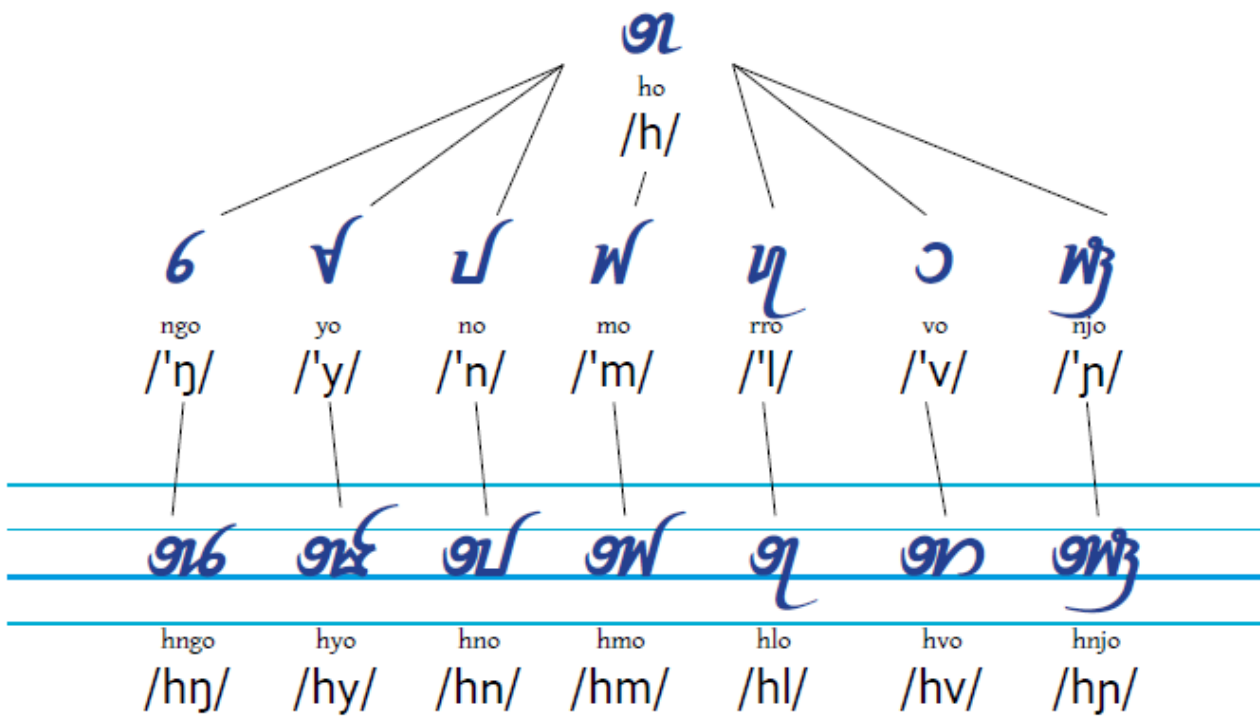
Fig. 1.3.2 HIGH HO and related letters

ꪵꪮꪆꪷꪚꪸꪮꪆꪚꪚꪚꪷ, *Lesson 3*

This relationship is also appropriate for Tai Viet script, so the mappings for the above ones are acceptable.

1.4. VOWEL AU (0xXX3A)

Tai Viet script uses a visual model, so 3 vowel letters are not encoded separately in the code chart. Please see Fig. 1.4.1.

17.　　⟨̃　　May kê (ê)

18.　　⟨̌　　May cơ (ơ)

19.　⟨　⟨̣　May cau (au)

Fig.1.4.1 May kê, May cơ & May cau in Tai Viet script

ขาก แอบ x̌ ใน - แอบ แก̌ม, *p. 14*

For May kê, the sequence should be <U+AAB9, ..., U+AAB8>; for May cơ, the sequence should be <U+AAB9, ..., U+AAB7>; for May cau, the sequence should be <U+AAB9, ..., U+AAB1>. Oswald added May kê and May cơ as U+AAC3 and U+AAC4.

VOWEL AU ◌̣ (0xXX3A) should correspond to May cau not only U+AAB9 ⟨. However, if Tai Don or Tai Khao uses the visual model and unified to Tai Viet, the sequence should be <U+AAB9, ..., U+AAB9>.

2. Syllable Structure

Kushim proposed the syllable structure as C[P]VT. Jim said "we did not include any rule of this nature in Tai Viet, but we could have. In hand written text, which served as a model for SIL's Tai Heritage font, there is great variation in where combining vowels are placed." in his feedback.

In the real running text for Tai Don or Tai Khao in Jinping, the position of the upper and bottom combining marks are not stable. For example, *nẃ* and *ḱw* both means the same value.


/kenm/

Fig. 2.1 nẃ
แอ√ห̌วปีทว√ปปึว, *Lesson 1*

以上 18 个为基础元音字母，需要说明的是字母 ⟨̃ (en) 在傣文中很少使用，使用时必须与字母 "ẃ" (m) 搭配使用，"ńẃ" 读音为 "kenm"；字母 ⟨̃ (anm) 在使用时与字母 "ẃ" 搭配使用，"ńẃ" 读音为" kanm"。字母 "⟨̌" 在傣文中需要与字母 "6" 搭配使用，"6" 读音为：/ang/，

Fig. 2.2 ḱw
แอ√ห̌วปีทว√ปปึว, *Lesson 1*

4

We can meet the same situation at the Tai Viet running text as well. Please also see Fig. 3.1 and 3.2.

In general, there is no consonant clusters in the current Jinping Dai language, so if two consonant letters appear side by side, the first one means the initial consonant and the last one means the final consonant. The vowel combining marks and the tone combining marks are placed in which consonant or right-hand side vowel, and the actual value is still clear for the readers.

Prof. Dao Jie and her colleagues pointed out the labialization syllables used in Jinping Dai language, so the syllable structure for the language level should be C[C]V[C] not CV[C].

声母	金平傣文	文字音	口语音	汉义
kw	ꪻꪫꪶ	kwa：ŋ³³	ka：ŋ³³	鹿
	ꪻꪫꪶ	kwa：ŋ²⁵	ka：ŋ²⁵	宽
xw	ꪫꪩꪬ	xwa：n³³	xa：n³³	斧头
	ꪬꪩꪫ	xwa：i⁴³	xa：i⁴³	水牛

Fig. 2.3 Labialization syllables in Jinping Dai language

刀洁，周焱，王玉清：《云南金平傣文语音的规范化问题》

In the paper, Prof. Dao didn't mention the labialization syllable with other vowels. I find the following examples in ꪖꪱꪒ ꪵꪮꪜ ꪲx̂ ꪶꪙ - ꪵꪮꪜ ꪵꪀꪜ which is a Vietnamese Tai language textbook. They are ꪬꪒꪮꪜ, ꪵꪀꪒꪮꪜ, ꪵꪀꪒꪮꪜ, ꪹꪀꪒꪮꪜ and ꪬꪒꪮꪜ.

| - | Quý | Quē | Que | Huệ | Quý |
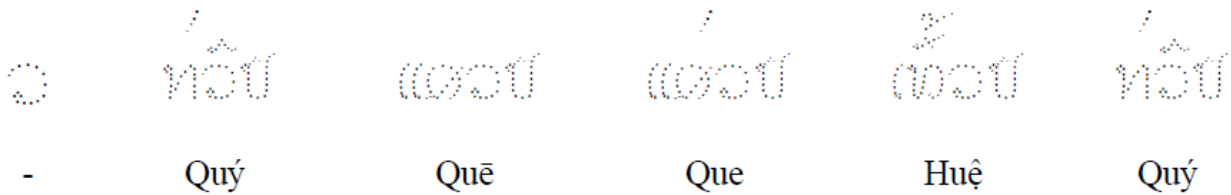
Fig. 2.4 ꪮ related syllables

ꪖꪱꪒ ꪵꪮꪜ x̂ ꪶꪙ - ꪵꪮꪜ ꪵꪀꪜ, *p. 27*

Jim once mentioned the labialization syllable issue in N3220 as below.

ꪬꪮ̂ /kiw/   vs.   ꪬꪮ̂ /kʷi/, and

ꪵꪬꪮ̌ /kɛw³/   vs.   ꪵꪬꪮ̌ /kʷɛ³/

Fig. 2.5 Labialization syllable examples in WG2 N3220

However, I checked the Tai Viet materials, ꪬꪮ̌ and ꪬꪮ̂ both means /kiw/, /kʷi/ should be ꪬꪮꪒꪜ or ꪬꪮꪒꪜ; ꪵꪬꪮ̌ and ꪵꪬꪮ̂ both means /kɛw³/, /kʷɛ³/ should be ꪵꪬꪮꪒꪜ and ꪵꪬꪮꪒꪜ. The upper or bottom vowels and the tone marks do not bear the function of distinguishing syllable elements, which they can drift to different positions. When ꪮ is the last one consonant in the syllable, it must be the final consonant not the second consonant of the initial consonant cluster. On the other hand, all the materials on the Vietnamese Tai languages and Jinping Dai language in my hand show the final part includes at most one consonant.

In a Tai Viet teaching video, more CCVC syllable examples with the consonant clusters are mentioned, so the -C- should not limit to ꪮ. Note that the glyph[s] for the -C- and -C part are still the same as the individual consonant glyph.

Cách đánh vần 꺶 꺖 꺩꺢꺺

- ꪆꪮꪣꪷ ꪆꪮꪣꪵ

    Kúc - ko - xo - mo = xmúc

    Cun - nŏ - xo - mo = xmun

- ꪁꪎꪸꪚ ꪁꪎꪸ

    Kap-bo-cŏ-lŏ = klạp

    Ka - cŏ - lŏ = cala

Fig. 2.6 A Tai Viet teaching video

And then, we need to meet an issue on the left-hand side vowels. In L2/22-098, Kushim proposes them as the combining marks. There is no virama used in Tai Don or Tai Khao and Tai Viet, and no medial consonant variations for any time. When the vowel is put after the second consonant, the glyph[s] for the vowel and the tone mark should be moved to the position before the second consonant, not the head most position of the whole cluster. It's also impossible to encode one other control function character like virama without any glyph or variation control in the cluster, that will make the end users confused for the visual spoofing.

Based on the above understandings, the visual model is the better choice for the Tai Don or Tai Khao and Tai Viet encoding.

On the other hand, we need some rules to make *nꪫ* and *hꪫ* as the same word for researching.

3. Punctuation

Kushim proposes 4 MAAI SAM symbols at the same time, and SAH asks why XX5E and XX5F are disunified in L2/22-068.

As Kushim shows, 0xXX5E 𖿢 should be mapped to U+AADE ꫞. In fact, there is other one symbol which should be mapped to 0xXX5F 𖿣 and Oswald added it to U+AADA ꫚. U+AADE ꫞ and U+AADA ꫚ could be used at the beginning of the text. Please see Fig. 3.1.

Fig. 3.1 ꪹꪁꪷ ꪚ꫁ ꪶꪙꪉ ꪁꪾ, p. 1

(Notice: The tone mark of the first syllable of the book name is placed at the last consonant in the main text, but at the first consonant in the cover, please also see Fig. 3.2.)

Fig. 3.2 Cover of ꪹꪁꪷ ꪚ꫁ ꪶꪙꪉ ꪁꪾ

In Tai Heritage Pro, the second symbol has been unified to U+AADE, and the end user should get it by the Graphite feature. Please see Fig. 3.3.

HO HOI Alternate (abbreviated)	hoia	0 *(default)*	Use full form of TAI VIET SYMBOL HO HOI	
		1	Use abbreviated form of TAI VIET SYMBOL HO HOI	

Fig. 3.3 `hoia` *feature in Tai Heritage Pro*

It looks it's better to encode them separately or add the second one to SVS, because the end users need them at the same time in the plain text.

4. Suggestion

1) Unify the Tai Don or Tai Khao letters to the corresponding letters in the Tai Viet block based on Kushim's mapping suggestions basically, and encode the unmapped letters used in Jinping separately.

2) Ask China NB to choose the proper script name in China, Tai Don or Tai Khao, and change the Tai Viet block name to "Tai Viet and Tai Don" or "Tai Viet and Tai Khao" like what we have done for the Bengali and Assamese block. I think Tai Khao will be better for Jinping.

3) Ask China NB to confirm the character name in Jinping and add the alternate character name as the annotations. The Tai Don or Tai Khao teaching video produced by the Jinping County shows the different letter names. Please see the video here.

4) Consider if it's suitable to use multi-columns form in the Tai Viet block, which is out of the CJKUI blocks. It will be very useful to include the Tai Don or Tai Khao glyphs and the Tai Viet glyphs at the same time under the unified code points.

Table 4 Possible muti-columns form example

	C	V
AA84		

5) Consider if it's suitable to add the SVS to U+AADE.

6) Consider a method on processing the combining marks drift issue.

7) Mark the Chinese conventions (Jinping) as `twh-Tavt-CN`, Vietnamese conventions as `blt-Tavt-VN` or `twh-Tavt-VN`. In an OS which includes the global languages, the languages for China (different dialects and minority languages in mainland, Hong Kong SAR, Macao SAR and Taiwan Province included), the glyphs for the Tai Viet block should use the Chinese conventions first; and the languages for Việt Nam and USA, the glyphs should use the Vietnamese conventions first.

5. Acknowledgement

(End of Document)

Appendix A *Bảng chữ Thái các vùng và chữ thống nhất*

Bảng chữ Thái các vùng và chữ thống nhất

I . Bảng phụ âm (Tô)

Số thứ tự	Tên chữ	Thống nhất	Thái đen	Mường Lay	Phong Thổ	Phù Yên	Mộc Châu
1	Bo						
2	Bỏ						
3	Po						
4	Pỏ						
5	Do						
6	Dỏ						
7	So						
8	Sỏ						
9	Pho						
10	Phỏ						
11	no						
12	nỏ						
13	mo						
14	mỏ						
15	ngo						
16	ngỏ						
17	vo						
18	vỏ						
19	lo						
20	lỏ						
21	Ko						
22	Kỏ						
23	Ho						
24	Hỏ						

Theo bản chữ Thái thống nhất năm 1955
của Khu tự trị Tây Bắc

9

Phụ âm (Tô) - tiếp

Số thứ tự	Tên chữ	Thống nhất	Thái đen	Mường Lay	Phong Thổ	Phù Yên	Mộc Châu
25	Kho	ก	ก	ก	ฆ	ฆ	ฆ
26	Khŏ	ჳ	ჳ	ฅ	ฆ	ฅ	ฆ
27	Do	ດ	ດ	ໄ	ໄ	ດ	ດ
28	Đŏ	ຯ	ຯ	ຯ	ຯ	ຯ	ຯ
29	Lo	ຶ	ຶ	ຶ	ຶ	ຶ	ຶ
30	Lŏ	ໄ	ໄ	ໄ	ໄ	ໄ	ໄ
31	o	໐	ໆ	ໆ	ໆ	໐	໐
32	ŏ	໐	໐	໐	໐	໐	໐
33	tho	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
34	thŏ	ໆ	ຍ	ໆ	ໆ	ໆ	ໆ
35	nho	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
36	nhŏ	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
37	Kh'	ໆ	—	ໆ	—	—	—
38	Kh'	ໆ	—	ໆ	ໆ	—	—
39	Pho	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
40	Phŏ	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
41	tro	ໆ	—	ໆ	ໆ	—	—
42	trŏ	ໆ	—	ໆ	ໆ	—	—
43	tho	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
44	thŏ	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ
45	mừng	1	ໆ	ໆ	ໆ	ໆ	ໆ
46	Kŏn	ໆ	ໆ	ໆ	ໆ	ໆ	ໆ

Ngoài ra còn có : ໆໆ hó hơi - ໆ <Kọi kọi> Lái sặm

Phong thổ: ໆ Mó mu - ໆ Bo bươn. Mộc Châu: út kịn

H . T . N

10

II - Bảng nguyên âm (May)

Số thứ tự	Tên may	Chữ Thống nhất	Vùng Thái đen	Lai Châu Phong thổ	Mộc Châu	Yên Châu
1	Khít			hay		
2	ca			hay		
3	cay					
4	Ke					
5	caư				(cở)	(cở)
6	căm					
7	caư					
8	căn					
9	cưa			(cở)		
10	cồng			(cung)		
11	cư					
12	cửa			(cở)		
13	cỏ					
14	cử					
15	Kia			(Kế)		
16	Ky					
17	cáng					
18	cắt					
19	quy					
20	que					
21	quồ					
22	quya					

H.T.N

11

Appendix B Tai Viet letters