L2/07-099
replaces L2/07-039

ISO/IEC JTC1/SC2/WG2 **N3220**

Date: 2007-03-20

ISO/IEC JTC1/SC2/WG2

Universal Multiple-Octet Coded Character Set (UCS) – ISO/IEC 10646

Secretariat: ANSI

Title	Proposal to encode the Tai Viet script in the UCS
Source	Jim Brase, SIL International
Action	For review at the next SC2/WG2
Distribution	SC2/WG2 Experts and Liaison Organizations

1. Sociolinguistic background

The Tai Viet script is used by three Tai languages spoken primarily in northwestern Vietnam, northern Laos, and central Thailand—Tai Dam (also Black Tai or Tai Noir), Tai Dón (White Tai or Tai Blanc), and Thai Song (Lao Song or Lao Song Dam). The Thai Song of Thailand are geographically removed from, but linguistically related to the Tai people of Vietnam and Laos. There are also populations in Australia, China, France, and the United States. The script is related to other Thai scripts used throughout Southeast Asia.

The *Ethnologue* (Gordon 2005) estimates the total population of the three languages, across all countries, at 1.3 million. (Tai Dam 764,000, Tai Dón 490,000, Thai Song 32,000.)

The script is still used by the Tai people in Vietnam, and there is a desire to introduce it into formal education there (Cầm Trọng 2005). On the other hand, it is not known whether it is in current use in Laos, Thailand, or China.

A fourth language, Tai Daeng (Red Tai or Tai Rouge, 165,000), uses a very closely related script. But the differences in the vowel structure of Tai Daeng are significant enough that it will probably require encoding as a separate script.

2. Script name

Several different spellings have been employed for the name of the language and script. In linguistic circles, it is common to use “Thai” to indicate the language of central Thailand, and “Tai” to indicate the language family. But even that usage is followed not consistently.

Some segments of the language community prefer the spelling “Tay”, first because it more closely reflects their own pronunciation for the name of their language, secondly because the spelling “Tai” resembles their word for “death”, and thirdly because of some negative connotations associated with the spelling “Tai” in Vietnamese. But these feelings are by no means universal. At least one major group in the Tai community in Des Moines, Iowa, has indicated to the author that they will continue to use the spelling “Tai”.

After some debate and experimenting with other names, we have settled on the name “Tai Viet”. The spelling “Tai” appears to be less confusing to the IT community which must implement the script. “Viet” distinguishes this script from other Tai scripts, while recognizing the fact that 90% of the user community is in Vietnam. The order “Tai Viet” is used, parallel to names like “Tai Le”, to make it easy to find when searching for “Tai”.

3. Basic features

The Tai Viet script shares many features with other Tai alphabets:

- It is written left to right.
- There is a double set of initial consonants, one for low tone class and one for high tone class.
- Vowels marks are positioned before, after, above, or below the syllable's initial consonant, depending on the vowel. Some vowels are written with digraphs.
- The consonants *do not* carry an implicit vowel. The vowel must always be written explicitly.

4. Storage order

Characters will be stored in visual order, for the following reasons:

- There are several ambiguities that occur in the script involving velar consonants plus the TAI VIET LETTER HIGH VO. In some instances, these interact with the encoding order. Visual order will involve fewer and less severe problems with these ambiguities than logical order, and will enable a more usable computing solution for the Tai Viet script than the alternatives. Please see Appendix 1 for a complete analysis of these ambiguities and how they interact with the encoding order.
- Established keyboarding practices use visual order, as do established handwriting practices. While input methods can be developed to support reordering of the input stream, those currently available are not advanced enough to provide a transparent editing environment after the input stream has been reordered. When the typist wants to edit text that has already been reordered, he finds that it is not stored in the order he sees and expects.
- Experience gained by SIL with this script in the 1990s on a Macintosh based system revealed that the user experiences considerable confusion when both the input and output streams are reordered. Thus, visual order will result in a much improved user experience.
- The Tai make up a relatively small user community. Consequently, vendors may not produce the necessary software to render phonetic order, while there is a strong and urgent need to get the script into users' hands as soon as possible. Complex systems will be difficult and expensive to maintain for a small community. Simplifying the input model will help to ensure that they have a system that works well.
- Visual order is also used by the Lao script, to which the Tai Viet script is closely related, and which some Tai users already read and type.

5. Word and Syllable Structure

The Tai languages are almost exclusively monosyllabic. A very small number of words have an unstressed initial syllable, and loan words may be polysyllabic. The practice followed in Baccam, et al. (1989) was to write polysyllabic words without space between the syllables. No tone is written on loan words or on the unstressed initial syllable of a native word.

There are two different systems of tone marks in use, one using combining marks written over the initial consonant, the other using spacing marks written on the baseline at the end of the syllable. See **Tone classes and tone marks**, below, for a discussion of the two different tone systems.

Depending on the tone system that is used, the written syllable may have any of the following structures:

Using combining tone marks	**Using spacing tone marks**
V_1 C W? T? F?	V_1 C W? F? T?
C W? V_2 T? F?	C W? V_2 F? T?
C W? T? V_3 F?	C W? V_3 F? T?
V_1 C W? V_2 T? F?	V_1 C W? V_2 F? T?
V_1 C W? T? V_3 F?	V_1 C W? V_3 F? T?
C W? F V_2	C W? F V_2

An initial consonant C is always written. Even when the initial consonant is null (phonetically a glottal stop), it is written with the symbol ꪮ or ꪯ.

Initial velar consonants may be labialized, indicated by W?. The labialization is marked by the high-series letter ‘v’, ꪫ, following the consonant.

V_1 indicates a pre-vowel that is rendered before the consonant. V_2 is a combining vowel rendered above or below the consonant, and V_3 is a post-vowel rendered after the consonant. Vowel digraphs can be formed from sequences $V_1 + V_2$ or $V_1 + V_3$.

T indicates an optional tone mark which, as already noted, may be either a combining tone mark over the initial consonant, or a spacing mark at the end of the syllable.

F indicates an optional final consonant.

The last syllable pattern is unusual. It only occurs for writing the vowel-final consonant combination /-ap/, which is written with the /am/ vowel placed over the final low-series /b/, rather than over the initial consonant.

In handwriting, styles vary as to where combining marks are placed. They are typically placed over (or under) the initial consonant, but when there is a final consonant present, they often drift to the right, sometimes being in the gap between the two consonants, and sometimes being over the final consonant.

When there is a labialized consonant, the placement of a combining mark can be important in resolving ambiguity. Combining marks are written over the second part of the labialized consonant. Thus:

ꪀꪲꪫ /kiw/	vs.	ꪀꪫꪲ /k^{w}i/, and
ꪵꪀꪰꪫ /kɛw^3/	vs.	ꪵꪀꪫꪰ /k^{w}ɛ^3/

6. Vowels

Vowel symbols can be classified according to where they are written relative to the initial consonant. Some of the vowels carry an inherent final consonant.

Vowels written before the consonant:

ꪵ /ɛ/	ꪵꪀ꪿ꪙ /kɛn^2/ ‘seed’
ꪶ /o/	ꪶꪄꪙ /xon^1/ ‘fur, feather/
ꪹ /ɨə/	ꪹꪎ /sɨa^1/ ‘tiger
ꪻ /əw/	ꪻꪑ꪿ /ɲəw^2/ ‘large, big’
ꪼ /aj/	ꪼꪒ꫁ /daj^3/ ‘to attain’

Vowels written above the consonant:

◌ꪰ /a/	ꪕꪰꪒ /tat^2/ ‘to cut’
◌ꪲ /i/	ꪖꪲꪫ /t^{h}iw^1/ ‘to whistle’
◌ꪸ /iə/	ꪣꪸ /miə^4/ ‘wife’
◌ꪳ /ɨ/	ꪜꪳꪙ /pɨn^1/ ‘arrow’
◌ꪷ /ɔ/	ꪜꪷ /pɔ^4/ ‘enough, sufficient’ (only used in open syllables)

◌ꪾ /am/ ꪀꪾ /kam^4/ ‘gold’

Vowels written below the consonant:

◌ꪴ /u/ ꪄꪴ꪿ꪙ /xun^2/ ‘dust’

Vowels written after the consonant:

ꪱ /aː/ ꪮꪱ꪿ꪉ /ʔaːŋ^2/ ‘basin, tub’

ꪮ /ɔ/ ꪮꪮꪀ /ʔɔʔ^2/ ‘to go out’ (used in closed syllables)

Note the double use of the TAI VIET LETTER LOW O in this example: in the first instance it functions as a consonant; in the second instance it functions as a vowel.

ꪺ /uə/ ꪎ꫁ꪺꪉ /suəŋ^3/ ‘trousers’

ꪽ /an/ ꪜ꫁ꪽ /pan^3/ ‘to squeeze’

In some dialects, the TAI VIET VOWEL AN can be used in isolation to represent the word /nan^6/, ‘that’.

Digraph vowels and other special sequences:

These are combining sequences which do not need to be encoded as separate units.

ꪹ◌ꪸ = ꪹ+ C$_i$ + ◌ꪸ /e/ ꪹꪔꪸꪣ /tem^1/ ‘full’

TAI VIET VOWEL UEA + initial consonant + TAI VIET VOWEL IA

ꪹ◌ꪷ = ꪹ+ C$_i$ + ◌ꪷ /ə/ ꪹꪬꪷꪉ /həŋ^1/ ‘long’

TAI VIET VOWEL UEA + initial consonant + TAI VIET MAI KHIT

ꪹ◌ꪱ = ꪹ + C$_i$ + ꪱ /aːw/ ꪹꪉꪱ /ŋaːw^1/ ‘reflection’

TAI VIET VOWEL UEA + initial consonant + TAI VIET VOWEL AA

ꪵ◌ꪫꪥ = ꪵ + C$_{\text{velar}}$ + ꪫ = C$_{\text{labialized-velar}}$ + /ɛ/ ꪵꪁꪫꪥ /k^{w}ɛ^2/ ‘cinnamon’

TAI VIET VOWEL E + initial consonant + TAI VIET LETTER HIGH VO + TAI VIET LETTER HIGH YO

The traditional sequence ꪵ + C$_{\text{velar}}$ + ꪫ is ambiguous in open syllables. It can be interpreted as either C$_{\text{velar}}$ + /ɛw/, or as C$_{\text{labialize-velar}}$ + /ɛ/. To eliminate this ambiguity, the character TAI VIET LETTER HIGH YO is sometimes appended to the end of the sequence to indicate the second pronunciation. Since /j/ never occurs after /ɛ/, this can be done without creating a new ambiguity. This spelling is only used in some dialects of the traditional script. However, it has been adopted as a standard spelling in a project sponsored by the Son La Department of Science and Technology.

◌ꪚꪾ = C_i + ꪚ + ◌ꪾ /ap/ ꪀꪚꪾ /kap^2/ ‘with, and’

initial consonant + TAI VIET LETTER LOW BO + TAI VIET VOWEL AM

The above combination is only used in some dialects of the script.

Vowels not encoded at this time:

ꪱ̂ There is a general consensus in the Tai community of the need to write certain sounds which are not part of the Tai languages, but which occur in words borrowed from Vietnamese. Among these is the sound represented in the Vietnamese alphabet by the LATIN CAPITAL/SMALL LETTER A WITH CIRCUMFLEX. A Tai Viet character which corresponded to this, TAI VIET VOWEL AA WITH CIRCUMFLEX, was included in the proposal presented to the UTC on 6 February 2007. Since then, we have learned that there is some disagreement over the properties of this proposed character. Consequently, the TAI VIET VOWEL AA WITH CIRCUMFLEX has been withdrawn from the proposed character repertoire. It may be brought as a separate proposal later.

7. Tone classes and tone marks

In the Tai Viet script each consonant has two forms. The low form of the initial consonant indicates that the syllable uses tone 1, 2, or 3. The high form of the initial consonant indicates that the syllable uses tone 4, 5, or 6. This is sufficient by itself to define the tone of checked syllables (those ending /p/, /t/, /k/, or /ʔ/), in that these syllables are restricted to tones 2 and 5.

Traditionally, the Tai Viet script did not use any further marking for tone. The reader had to determine the tone of unchecked syllables from the context. In recent times, however, several groups have introduced tone marks into Tai Viet writing. Tai Dam speakers in the United States begin using Lao tone marks with their script about 30 years ago, and those marks are included in SIL’s Tai Heritage font. These symbols are written as combining marks above the initial consonant, or above a combining vowel, and are identified by their Laotian names, *mai ek* and *mai tho*. These marks are also used by the Song Petburi font (developed for the Thai Song language), although they were probably borrowed from the Thai alphabet rather than the Lao.

The Tai community in Vietnam, however, invented their own tone marks written on the base line at the end of the syllable, which they name *mai nueng* and *mai song*.

When combined with the consonant class, two tone marks are sufficient to unambiguously mark the tone. Thus, depending on which system one uses, tones may be written as follows on unchecked syllables:

	no mark	◌꪿	◌꫁
low class consonant	tone 1	tone 2	tone 3
high class consonant	tone 4	tone 5	tone 6

Marking tones with symbols *mai ek* and *mai tho* in unchecked syllables

	no mark	◌..ꫀ	◌..ꫂ
low class consonant	tone 1	tone 2	tone 3
high class consonant	tone 4	tone 5	tone 6

Marking tones with symbols *mai nueng* and *mai song* in unchecked syllables

It is recognized that the existence of two distinct sets of tone marks is a disadvantage to the script. However, they cannot be unified, because both their combining classes and their storage order are different. For example:

ꪄ꫁ꪮꪉ = ꪄ + ◌꫁ + ꪮ + ꪉ

ꪄꪮꪉꫂ = ꪄ + ꪮ + ꪉ + ꫂ

(/xɔŋ^{3}/, ‘to trip over)

Perhaps in time, one system will become dominant and the other will die out. But for now, both are in use.

8. Final consonants

In written form, the low-tone class symbols for ‘b’ (ꪚ) and ‘d’ (ꪒ) are used for syllable final /p/ and /t/, respectively, as is the practice in many Thai scripts.

The low-tone class symbol for ‘k’ (ꪀ) is used for both final /k/ and final /ʔ/.

The high-tone class symbols are used for writing final /j/ (ꪥ) and the final nasals, /m/ (ꪣ), /n/ (ꪙ), and /ŋ/ (ꪉ). High-tone /v/ (ꪫ) is used for final /w/.

There are a number of exceptions to the above rules in the form of vowels which carry an inherent final consonant. These vary from region to region. The ones included in this proposal are the ones with the broadest usage: /-aj/ (ꪼ◌), /-am/ (◌ꪾ), /-an/ (◌ꪽ), and /-əw/ (ꪻ◌). Another vowel-final consonant form, /-ap/ (◌ꪾꪚ), is composed of the –am ligature plus ‘b’, but does not need to be encoded.

9. Symbols

There are five non-alphabetic symbols:

Symbol	Name	Tai name/ pronunciation	meaning
ꫛ	TAI VIET SYMBOL KON	/kon^4/	‘person’
ꫜ	TAI VIET SYMBOL NUENG	/nɨŋ^5/	‘one’
ꫝ	TAI VIET SYMBOL SAM	*sam*	signals repetition of the previous word
꫞	TAI VIET SYMBOL HO HOI	*ho hoi*	beginning of text (used in songs and poems)
꫟	TAI VIET SYMBOL KOI KOI	*koi koi*	end of text (used in songs and poems)

10. Symbols as ligatures

Two of the symbols listed above, TAI VIET SYMBOL KON and TAI VIET SYMBOL NUENG, may be regarded as ligatures of entire words. They should nevertheless be encoded as separate characters.

In the case of TAI VIET SYMBOL KON, the use or non-use of the ligature is used to distinguish between homophonous words:

ꫛ = /kon^4/ ‘person’ ꪶꪀꪙ = /kon^4/ ‘to stir’

It is not known if there is any homonym for TAI VIET SYMBOL NUENG. But it is necessary to encode TAI VIET SYMBOL KON, and for the sake of consistency it is felt that they should both encoded.

11. Word spacing and line breaks

Traditionally, the Tai Viet script was written without spaces between words. In the last 30 years, users in both Vietnam and the United States have started writing spaces between words, in both hand written and machine produced texts. Most users now use interword spacing. However, considering the scripts historical usage, it was felt that a set of line breaking rules should be provided for instances where interword spacing is not desired. These can be found in Appendix 2.

12. Code chart order and sort order

The Tai Viet scripts does not have an established standard for sorting. Sequences have sometimes been borrowed from neighboring languages. Baccam, et al. (1989) is a Tai Dam-English dictionary that uses the Lao order, adjusted for differences between the Tai Dam and Lao character sets. Cầm Trọng (2005) prefers an order based on the Vietnamese alphabet (Quốc ngữ). More discussion with the Tai community is needed on this matter, but it is possible that communities in different countries will want to use different orders.

Given the Tai Viet's script similarity to Lao, we chose the order from Baccam for our code chart. However, a number of characters had to be added to the set used in Baccam. These mostly consisted of characters required for the larger consonant inventory of Tai Don. However, two pairs of consonants, the LOW and HIGH GO and the LOW and HIGH RO, are innovations of the last 50 years by the Tai Viet community in Vietnam for writing Vietnamese loan words. The placement of the LOW and HIGH RO between the YO and LO was straightforward—it follows the order in the Lao alphabet.

The placement of the LOW and HIGH GO is more problematic. These characters do not generally occur in Thai scripts, so there were no other alphabets to use as a pattern. Thai scripts do follow a pattern, however. For characters at any given point of articulation, they usually place voiced consonants first, then voiceless unaspirated stops, aspirated stops, fricatives, and then nasals. Placing the GO immediately before the NGO violates this pattern. It was done because we did not wish to place an innovative character at the beginning of the alphabet.

Work is underway on two different collating orders. The default collating order will be based on Baccam. As in other Thai scripts, it will be based on pronunciation.

As noted above, Cầm Trọng and others in Vietnam prefer an order based on Vietnamese. However, they do not use a true Latin-based order. They use a pronunciation-based collating order, similar to the practice of other Thai scripts, but with the consonants sorted in Latin order, and the vowels sorted in Latin order. We will provide an alternative collating sequence based on this practice.

13. Reserved characters

It is recommended that character codes AAC3..AADA be reserved for future expansion of the Tai Viet character set.

The current proposal is focused on the use of the script by the Tai Dam of Son La province, Vietnam. It contains the traditional Son La character set, plus three pairs of aspirated consonants required by the Tai Don language of Lai Chau province.

TAI VIET LETTER LOW KHO	TAI VIET LETTER HIGH KHO
TAI VIET LETTER LOW CHO	TAI VIET LETTER HIGH CHO
TAI VIET LETTER LOW PHO	TAI VIET LETTER HIGH PHO

Tai Don can be written with the resulting character set, but only if one uses the orthographic conventions of Son La. If one wishes to write Tai Don in one of their traditional styles, some additional characters will probably be needed. The author has identified four consonants and two vowels which have definite or probable contrast with characters from the Son La tradition, and 12-14 characters which use significantly different glyphs, although not in a contrastive way. However, these require additional study, and hopefully the input of someone who is an expert in Tai Don, before they can be proposed for the character set.

14. Character Properties

```
AA80;TAI VIET LETTER LOW KO;Lo;0;L;;;;;N;;;;;
AA81;TAI VIET LETTER HIGH KO;Lo;0;L;;;;;N;;;;;
AA82;TAI VIET LETTER LOW KHO;Lo;0;L;;;;;N;;;;;
AA83;TAI VIET LETTER HIGH KHO;Lo;0;L;;;;;N;;;;;
AA84;TAI VIET LETTER LOW KHHO;Lo;0;L;;;;;N;;;;;
AA85;TAI VIET LETTER HIGH KHHO;Lo;0;L;;;;;N;;;;;
AA86;TAI VIET LETTER LOW GO;Lo;0;L;;;;;N;;;;;
AA87;TAI VIET LETTER HIGH GO;Lo;0;L;;;;;N;;;;;
AA88;TAI VIET LETTER LOW NGO;Lo;0;L;;;;;N;;;;;
AA89;TAI VIET LETTER HIGH NGO;Lo;0;L;;;;;N;;;;;
AA8A;TAI VIET LETTER LOW CO;Lo;0;L;;;;;N;;;;;
AA8B;TAI VIET LETTER HIGH CO;Lo;0;L;;;;;N;;;;;
AA8C;TAI VIET LETTER LOW CHO;Lo;0;L;;;;;N;;;;;
AA8D;TAI VIET LETTER HIGH CHO;Lo;0;L;;;;;N;;;;;
AA8E;TAI VIET LETTER LOW SO;Lo;0;L;;;;;N;;;;;
```

```
AA8F;TAI VIET LETTER HIGH SO;Lo;0;L;;;;;N;;;;;
AA90;TAI VIET LETTER LOW NYO;Lo;0;L;;;;;N;;;;;
AA91;TAI VIET LETTER HIGH NYO;Lo;0;L;;;;;N;;;;;
AA92;TAI VIET LETTER LOW DO;Lo;0;L;;;;;N;;;;;
AA93;TAI VIET LETTER HIGH DO;Lo;0;L;;;;;N;;;;;
AA94;TAI VIET LETTER LOW TO;Lo;0;L;;;;;N;;;;;
AA95;TAI VIET LETTER HIGH TO;Lo;0;L;;;;;N;;;;;
AA96;TAI VIET LETTER LOW THO;Lo;0;L;;;;;N;;;;;
AA97;TAI VIET LETTER HIGH THO;Lo;0;L;;;;;N;;;;;
AA98;TAI VIET LETTER LOW NO;Lo;0;L;;;;;N;;;;;
AA99;TAI VIET LETTER HIGH NO;Lo;0;L;;;;;N;;;;;
AA9A;TAI VIET LETTER LOW BO;Lo;0;L;;;;;N;;;;;
AA9B;TAI VIET LETTER HIGH BO;Lo;0;L;;;;;N;;;;;
AA9C;TAI VIET LETTER LOW PO;Lo;0;L;;;;;N;;;;;
AA9D;TAI VIET LETTER HIGH PO;Lo;0;L;;;;;N;;;;;
AA9E;TAI VIET LETTER LOW PHO;Lo;0;L;;;;;N;;;;;
AA9F;TAI VIET LETTER HIGH PHO;Lo;0;L;;;;;N;;;;;
AAA0;TAI VIET LETTER LOW FO;Lo;0;L;;;;;N;;;;;
AAA1;TAI VIET LETTER HIGH FO;Lo;0;L;;;;;N;;;;;
AAA2;TAI VIET LETTER LOW MO;Lo;0;L;;;;;N;;;;;
AAA3;TAI VIET LETTER HIGH MO;Lo;0;L;;;;;N;;;;;
AAA4;TAI VIET LETTER LOW YO;Lo;0;L;;;;;N;;;;;
AAA5;TAI VIET LETTER HIGH YO;Lo;0;L;;;;;N;;;;;
AAA6;TAI VIET LETTER LOW RO;Lo;0;L;;;;;N;;;;;
AAA7;TAI VIET LETTER HIGH RO;Lo;0;L;;;;;N;;;;;
AAA8;TAI VIET LETTER LOW LO;Lo;0;L;;;;;N;;;;;
AAA9;TAI VIET LETTER HIGH LO;Lo;0;L;;;;;N;;;;;
AAAA;TAI VIET LETTER LOW VO;Lo;0;L;;;;;N;;;;;
AAAB;TAI VIET LETTER HIGH VO;Lo;0;L;;;;;N;;;;;
AAAC;TAI VIET LETTER LOW HO;Lo;0;L;;;;;N;;;;;
AAAD;TAI VIET LETTER HIGH HO;Lo;0;L;;;;;N;;;;;
AAAE;TAI VIET LETTER LOW O;Lo;0;L;;;;;N;;;;;
AAAF;TAI VIET LETTER HIGH O;Lo;0;L;;;;;N;;;;;
AAB0;TAI VIET MAI KANG;Mn;230;NSM;;;;;N;;;;;
AAB1;TAI VIET VOWEL AA;Lo;0;L;;;;;N;;;;;
AAB2;TAI VIET VOWEL I;Mn;230;NSM;;;;;N;;;;;
AAB3;TAI VIET VOWEL UE;Mn;230;NSM;;;;;N;;;;;
AAB4;TAI VIET VOWEL U;Mn;220;NSM;;;;;N;;;;;
AAB5;TAI VIET VOWEL E;Lo;0;L;;;;;N;;;;;
AAB6;TAI VIET VOWEL O;Lo;0;L;;;;;N;;;;;
AAB7;TAI VIET MAI KHIT;Mn;230;NSM;;;;;N;;;;;
AAB8;TAI VIET VOWEL IA;Mn;230;NSM;;;;;N;;;;;
AAB9;TAI VIET VOWEL UEA;Lo;0;L;;;;;N;;;;;
AABA;TAI VIET VOWEL UA;Lo;0;L;;;;;N;;;;;
AABB;TAI VIET VOWEL AUE;Lo;0;L;;;;;N;;;;;
AABC;TAI VIET VOWEL AY;Lo;0;L;;;;;N;;;;;
AABD;TAI VIET VOWEL AN;Lo;0;L;;;;;N;;;;;
AABE;TAI VIET VOWEL AM;Mn;230;NSM;;;;;N;;;;;
AABF;TAI VIET TONE MAI EK;Mn;230;NSM;;;;;N;;;;;
AAC0;TAI VIET TONE MAI NUENG;Lo;0;L;;;;;N;;;;;
AAC1;TAI VIET TONE MAI THO;Mn;230;NSM;;;;;N;;;;;
AAC2;TAI VIET TONE MAI SONG;Lo;0;L;;;;;N;;;;;
AADB;TAI VIET SYMBOL KON;So;0;L;;;;;N;;;;;
AADC;TAI VIET SYMBOL NUENG;So;0;L;;;;;N;;;;;
AADD;TAI VIET SYMBOL SAM;So;0;L;;;;;N;;;;;
AADE;TAI VIET SYMBOL HO HOI;So;0;L;;;;;N;;;;;
AADF;TAI VIET SYMBOL KOI KOI;So;0;L;;;;;N;;;;;
```

The Logical Order Exception flag should be set for the following characters:

```
AAB5;TAI VIET VOWEL E
AAB6;TAI VIET VOWEL O
AAB9;TAI VIET VOWEL UEA
AABA;TAI VIET VOWEL UA
AABB;TAI VIET VOWEL AUE
AABC;TAI VIET VOWEL AY
```

iISO/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://www.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://www.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html**.**

See also http://www.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html **for latest *Roadmaps*.**

A. Administrative

1. **Title:** ***Proposal to encode the Tai Viet script in the UCS***

2. Requester's name: *Jim Brase, SIL International*

3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*

4. Submission date: *2006-01-30*

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: *Tai Viet*
b. The proposal is for addition of character(s) to an existing block: *no*
Name of the existing block:

2. Number of characters in proposal: *72*

3. Proposed category (select one from below - see section 2.2 of P&P document):

A-Contemporary	*X*	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. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document): *3*
Is a rationale provided for the choice?
If Yes, reference:

5. 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*

6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *TayVN Working Group (TrueType format)*
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used: *Ngô Trung Việt vietnt@trprog.gov.vn, James Đỗ jdo@pacificlinks.org*
FontLab

7. 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? *yes*

8. 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)? *yes*
line-breaking algorithm; included as appendix

9. 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 http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

[1] Form number: N3002-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)

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.)? *yes*

If YES, with whom? *Viet Nam National Body (Ngô Trung Việt), Sơn La Province (Lò Mai Cương)*
Private individuals in the Tai Dam community, Des Moines, IA (Bacthi Siang)

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? *yes*

Reference:

4. The context of use for the proposed characters (type of use; common or rare) *common*

Reference:

5. Are the proposed characters in current use by the user community? *yes*

If YES, where? Reference: *Viet Nam and United States; possibly also Laos, Thailand, and China*

6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP? *yes*

If YES, is a rationale provided? *strong user community*

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? *yes--ligatures*

If YES, is a rationale for its inclusion provided? *yes*

If YES, reference:

10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to 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? *combining characters are an inherent part of the writing system*

If YES, reference:

Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided? *none*

If YES, reference:

12. Does the proposal contain characters with any special properties such as control function or similar semantics? *no*

If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility character(s)? *no*

If YES, is the equivalent corresponding unified ideographic character(s) identified?

If YES, reference:

	AA8	AA9	AAA	AAB	AAC	AAD
0	ꪀ AA80	ꪐ AA90	ꪠ AAA0	◌ꪰ AAB0	ꫀ AAC0	
1	ꪁ AA81	ꪑ AA91	ꪡ AAA1	ꪱ AAB1	◌꫁ AAC1	
2	ꪂ AA82	ꪒ AA92	ꪢ AAA2	◌ꪲ AAB2	ꫂ AAC2	
3	ꪃ AA83	ꪓ AA93	ꪣ AAA3	◌ꪳ AAB3		
4	ꪄ AA84	ꪔ AA94	ꪤ AAA4	◌ꪴ AAB4		
5	ꪅ AA85	ꪕ AA95	ꪥ AAA5	ꪵ AAB5		
6	ꪆ AA86	ꪖ AA96	ꪦ AAA6	ꪶ AAB6		
7	ꪇ AA87	ꪗ AA97	ꪧ AAA7	◌ꪷ AAB7		
8	ꪈ AA88	ꪘ AA98	ꪨ AAA8	◌ꪸ AAB8		
9	ꪉ AA89	ꪙ AA99	ꪩ AAA9	ꪹ AAB9		
A	ꪊ AA8A	ꪚ AA9A	ꪪ AAAA	ꪺ AABA		
B	ꪋ AA8B	ꪛ AA9B	ꪫ AAAB	ꪻ AABB		ꫛ AADB
C	ꪌ AA8C	ꪜ AA9C	ꪬ AAAC	ꪼ AABC		ꫜ AADC
D	ꪍ AA8D	ꪝ AA9D	ꪭ AAAD	ꪽ AABD		ꫝ AADD
E	ꪎ AA8E	ꪞ AA9E	ꪮ AAAE	◌ꪾ AABE		꫞ AADE
F	ꪏ AA8F	ꪟ AA9F	ꪯ AAAF	◌꪿ AABF		꫟ AADF

Consonants

AA80	ꪀ	TAI VIET LETTER LOW KO
AA81	ꪁ	TAI VIET LETTER HIGH KO
AA82	ꪂ	TAI VIET LETTER LOW KHO
AA83	ꪃ	TAI VIET LETTER HIGH KHO
AA84	ꪄ	TAI VIET LETTER LOW KHHO
AA85	ꪅ	TAI VIET LETTER HIGH KHHO
AA86	ꪆ	TAI VIET LETTER LOW GO
AA87	ꪇ	TAI VIET LETTER HIGH GO
AA88	ꪈ	TAI VIET LETTER LOW NGO
AA89	ꪉ	TAI VIET LETTER HIGH NGO
AA8A	ꪊ	TAI VIET LETTER LOW CO
AA8B	ꪋ	TAI VIET LETTER HIGH CO
AA8C	ꪌ	TAI VIET LETTER LOW CHO
AA8D	ꪍ	TAI VIET LETTER HIGH CHO
AA8E	ꪎ	TAI VIET LETTER LOW SO
AA8F	ꪏ	TAI VIET LETTER HIGH SO
AA90	ꪐ	TAI VIET LETTER LOW NYO
AA91	ꪑ	TAI VIET LETTER HIGH NYO
AA92	ꪒ	TAI VIET LETTER LOW DO
AA93	ꪓ	TAI VIET LETTER HIGH DO
AA94	ꪔ	TAI VIET LETTER LOW TO
AA95	ꪕ	TAI VIET LETTER HIGH TO
AA96	ꪖ	TAI VIET LETTER LOW THO
AA97	ꪗ	TAI VIET LETTER HIGH THO
AA98	ꪘ	TAI VIET LETTER LOW NO
AA99	ꪙ	TAI VIET LETTER HIGH NO
AA9A	ꪚ	TAI VIET LETTER LOW BO
AA9B	ꪛ	TAI VIET LETTER HIGH BO
AA9C	ꪜ	TAI VIET LETTER LOW PO
AA9D	ꪝ	TAI VIET LETTER HIGH PO
AA9E	ꪞ	TAI VIET LETTER LOW PHO
AA9F	ꪟ	TAI VIET LETTER HIGH PHO
AAA0	ꪠ	TAI VIET LETTER LOW FO
AAA1	ꪡ	TAI VIET LETTER HIGH FO
AAA2	ꪢ	TAI VIET LETTER LOW MO
AAA3	ꪣ	TAI VIET LETTER HIGH MO
AAA4	ꪤ	TAI VIET LETTER LOW YO
AAA5	ꪥ	TAI VIET LETTER HIGH YO
AAA6	ꪦ	TAI VIET LETTER LOW RO
AAA7	ꪧ	TAI VIET LETTER HIGH RO
AAA8	ꪨ	TAI VIET LETTER LOW LO
AAA9	ꪩ	TAI VIET LETTER HIGH LO
AAAA	ꪪ	TAI VIET LETTER LOW VO
AAAB	ꪫ	TAI VIET LETTER HIGH VO
AAAC	ꪬ	TAI VIET LETTER LOW HO
AAAD	ꪭ	TAI VIET LETTER HIGH HO
AAAE	ꪮ	TAI VIET LETTER LOW O
AAAF	ꪯ	TAI VIET LETTER HIGH O

Vowels and Finals

AAB0	◌ꪰ	TAI VIET MAI KANG
AAB1	ꪱ	TAI VIET VOWEL AA
AAB2	◌ꪲ	TAI VIET VOWEL I
AAB3	◌ꪳ	TAI VIET VOWEL UE
AAB4	◌ꪴ	TAI VIET VOWEL U
AAB5	ꪵ	TAI VIET VOWEL E
AAB6	ꪶ	TAI VIET VOWEL O
AAB7	◌ꪷ	TAI VIET MAI KHIT
AAB8	◌ꪸ	TAI VIET VOWEL IA
AAB9	ꪹ	TAI VIET VOWEL UEA
AABA	ꪺ	TAI VIET VOWEL UA
AABB	ꪻ	TAI VIET VOWEL AUE
AABC	ꪼ	TAI VIET VOWEL AY
AABD	ꪽ	TAI VIET VOWEL AN
AABE	◌ꪾ	TAI VIET VOWEL AM

Tones

AABF	◌꪿	TAI VIET TONE MAI EK
AAC0	ꫀ	TAI VIET TONE MAI NUENG
AAC1	◌꫁	TAI VIET TONE MAI THO
AAC2	ꫂ	TAI VIET TONE MAI SONG

Symbols

AADB	ꫛ	TAI VIET SYMBOL KON
AADC	ꫜ	TAI VIET SYMBOL NUENG
AADD	ꫝ	TAI VIET SYMBOL SAM
AADE	꫞	TAI VIET SYMBOL HO HOI
AADF	꫟	TAI VIET SYMBOL KOI KOI

Appendix 1—Visual order vs. logical order

The purpose of this section is to explore several of the ambiguities that exist in the Tai Viet Script and the manner in which they interact with the encoding. It concludes that visual order is preferable to logical order for this script.

The ambiguities with which we are concerned revolve around the interpretation of the TAI VIET LETTER HIGH VO[1]. In addition to the normal function of representing the initial or final consonant of a syllable, this character can be used to mark labialization of a velar consonant. That is, labialized consonants are represented by the digraphs

ꪀꪫ	TAI VIET LETTER LOW KO + TAI VIET LETTER HIGH VO
ꪁꪫ	TAI VIET LETTER HIGH KO + TAI VIET LETTER HIGH VO
ꪂꪫ	TAI VIET LETTER LOW KHO + TAI VIET LETTER HIGH VO
ꪃꪫ	TAI VIET LETTER HIGH KHO + TAI VIET LETTER HIGH VO
ꪄꪫ	TAI VIET LETTER LOW KHHO + TAI VIET LETTER HIGH VO
ꪅꪫ	TAI VIET LETTER HIGH KHHO + TAI VIET LETTER HIGH VO
ꪈꪫ	TAI VIET LETTER LOW NGO + TAI VIET LETTER HIGH VO
ꪉꪫ	TAI VIET LETTER HIGH NGO + TAI VIET LETTER HIGH VO

These digraphs can interact with syllable boundaries and/or with left-side vowels in three ways. I will start by identifying each of these interactions. I will examine how each is affected by the use of visual order vs. the use of logical order, and how each interacts with the processes of rendering, sorting, and line/word breaking.

[1] There is an additional ambiguity involving the TAI VIET LETTER LOW KO, which can represent either a /k/ or /ʔ/ in the syllable final position. But this is solely an orthographic issue. The encoding does not affect it in any way, so we will not deal with it here.

1. Ambiguous syllable boundary in two syllable sequences involving LOW KO/HIGH NGO + HIGH VO and a left-side vowel.

	Visual Order	Logical Order
Encoding Sequence:	$\text{C} + \text{V} + \text{V}_{\text{(left)}} + \text{C}_{\text{(LowKo/HighNgo)}} + \text{HighVo}$ or $\text{C} + \text{V} + \text{C}_{\text{(LowKo/HighNgo)}} + \text{V}_{\text{(left)}} + \text{HighVo}$	$\text{C} + \text{V} + \text{C}_{\text{(LowKo/HighNgo)}} + \text{HighVo} + \text{V}_{\text{(left)}}$
Ambiguous?	no	yes
Possible interpretations[2]:	Determined by the encoding. The syllable break is before the $\text{V}_{\text{(left)}}$.	$\text{C} + \text{V}\ .\ \text{C}_{\text{(LowKo/HighNgo)}} + \text{HighVo} + \text{V}_{\text{(left)}}$ or $\text{C} + \text{V} + \text{C}_{\text{(LowKo/HighNgo)}}\ .\ \text{HighVo} + \text{V}_{\text{(left)}}$
Affects rendering?	n/a[3]	yes
Affects sorting?	n/a	yes
Affects line/word break?	n/a	yes
Affects reading?	n/a	not if rendered correctly
Ambiguity resolved by interword spacing?	n/a	yes
Ambiguity resolved by adding tone mark or final consonant?	n/a	no

The ambiguity that exists in the logical-order encoding for this sequence arises from the dual function of two of the consonants. The characters TAI VIET LETTER LOW KO and TAI VIET LETTER HIGH NGO can function as either initial or final consonants. At the same time, the character TAI VIET LETTER HIGH VO can function either as an initial consonant or to mark the labialization of an initial velar consonant. This prevents the rendering engine from identifying where the syllable boundary is. But the rendering engine must be able to identify which is the initial consonant of the second syllable in order to place the left-side vowel correctly.

Note that once the text has been rendered correctly, there is no ambiguity for the reader.

Following are a number of examples which illustrate the ambiguity.

[2] I have used the IPA convention of a dot (.) to indicate the syllable boundary.

[3] “n/a” = not applicable

Examples with contrasting minimal pairs

1) /suk vɛn di/, ‘cooked is better’ vs. /su kʷɛn kan/ ‘you (pl) acquainted with each other’

data in logical order:

LOW SO +	VOWEL U +	LOW KO +	HIGH VO +	VOWEL E +	HIGH NO	+ ꪒꪲ / ꪀꪰꪙ
ꪎ	◌ꪴ	ꪀ	ꪫ	ꪵ	ꪙ	+ ꪒꪲ / ꪀꪰꪙ

data in visual order for /suk vɛn di/:

LOW SO +	VOWEL U +	LOW KO +	VOWEL E +	HIGH VO +	HIGH NO	+ ꪒꪲ
ꪎ	◌ꪴ	ꪀ	ꪵ	ꪫ	ꪙ	+ ꪒꪲ

data in visual order for /su kʷɛn kan/:

LOW SO +	VOWEL U +	VOWEL E +	LOW KO +	HIGH VO +	HIGH NO	+ ꪀꪰꪙ
ꪎ	◌ꪴ	ꪵ	ꪀ	ꪫ	ꪙ	+ ꪀꪰꪙ

rendering of /suk vɛn di/: rendering of /su kʷɛn kan/:

ꪎꪴꪀ + ꪵꪫꪙ + ꪒꪲ ꪎꪴ + ꪵꪀꪫꪙ + ꪀꪰꪙ

2) /fuʔ vɛn di/ ‘tied better’ vs. /fu kʷɛŋ huə/ ‘person/who shakes (his) head’

data in logical order:

LOW FO +	VOWEL U +	LOW KO +	HIGH VO +	VOWEL E +	HIGH NO +	LOW DO +	VOWEL I
ꪠ	◌ꪴ	ꪀ	ꪫ	ꪵ	ꪙ	ꪒ	◌ꪲ

data in visual order for /fuʔ vɛn di/:

LOW FO +	VOWEL U +	LOW KO +	VOWEL E +	HIGH VO +	HIGH NO +	LOW DO +	VOWEL I
ꪠ	◌ꪴ	ꪀ	ꪵ	ꪫ	ꪙ	ꪒ	◌ꪲ

data in visual order for /fu kʷɛŋ huə/:

LOW FO +	VOWEL U +	VOWEL E +	LOW KO +	HIGH VO +	HIGH NGO +	LOW HO +	VOWEL UA
ꪠ	◌ꪴ	ꪵ	ꪀ	ꪫ	ꪉ	ꪬ	ꪺ

rendering for /fuʔ vɛn di/: rendering for /fu kʷɛŋ huə/:

ꪠꪴꪀ + ꪵꪫꪙ + ꪒꪲ ꪠꪴ + ꪵꪀꪫꪉ + ꪬꪺ

Examples in which the syllable boundary should be between the velar consonant and the /v/:

3) /coŋ vaj/, ‘to reserve’

data in logical order:

LOW CO +	VOWEL O +	HIGH NGO +	HIGH VO +	VOWEL AY
ꪊ	ꪶ	ꪉ	ꪫ	ꪼ

data in visual order:

VOWEL O +	LOW CO +	HIGH NGO +	VOWEL AY +	HIGH VO
ꪶ	ꪊ	ꪉ	ꪼ	ꪫ

correct rendering:	incorrect rendering:
ꪶꪊꪉ + ꪼꪫ	ꪶꪊ + ꪼꪉꪫ
/coŋ vaj/	/co ŋ^{w}aj/

4) /teŋ ven/, ‘daytime’

data in logical order:

HIGH TO +	VOWEL EUA +	VOWEL IA +	HIGH NGO +	HIGH VO +	VOWEL EUA +	VOWEL IA +	HIGH NO
ꪕ	ꪹ	◌ꪸ	ꪉ	ꪫ	ꪹ	◌ꪸ	ꪙ

data in visual order:

VOWEL EUA +	HIGH TO +	VOWEL IA +	HIGH NGO +	VOWEL EUA +	HIGH VO +	VOWEL IA +	HIGH NO
ꪹ	ꪕ	◌ꪸ	ꪉ	ꪹ	ꪫ	◌ꪸ	ꪙ

correct rendering:	incorrect rendering:
ꪹꪕꪸꪉ + ꪹꪫꪸꪙ	ꪹꪕꪸ + ꪹꪉꪫꪸꪙ
/teŋ ven/	/te ŋ^{w}en/

5) /tok vɛn taː/ ‘drop eyeglasses’

data in logical order:

LOW TO +	VOWEL O +	LOW KO +	HIGH VO +	VOWEL E +	HIGH NO +	LOW TO +	VOWEL AA
ꪔ	ꪶ	ꪀ	ꪫ	ꪵ	ꪙ	ꪔ	ꪱ

data in visual order:

VOWEL O +	LOW TO +	LOW KO +	VOWEL E +	HIGH VO +	HIGH NO +	LOW TO +	VOWEL AA
ꪶ	ꪔ	ꪀ	ꪵ	ꪫ	ꪙ	ꪔ	ꪱ

correct rendering:	incorrect rendering:
ꪶꪔꪀ + ꪵꪫꪙ + ꪔꪱ	ꪶꪔ + ꪵꪀꪫꪙ + ꪔꪱ
/tok vɛn taː/	/to kʷɛn taː/

Examples in which the syllable boundary should be before the labialized velar consonant:

6) /ci kʷaj/ ‘will clear (a ditch)’

data in logical order:

HIGH CO +	VOWEL I +	LOW KO +	HIGH VO +	VOWEL AY
ꪋ	◌ꪲ	ꪀ	ꪫ	ꪼ

data in visual order:

HIGH CO +	VOWEL I +	VOWEL AY +	LOW KO +	HIGH VO
ꪋ	◌ꪲ	ꪼ	ꪀ	ꪫ

correct rendering:	incorrect rendering:
ꪋꪲ + ꪼꪀꪫ	ꪋꪲꪀ + ꪼꪫ
/ci kʷaj/	/cik vaj/

7) /sɔŋ fu kʷɛn kan/ ‘two persons acquainted with each other’

data in logical order:

ꪏꪮꪉ +	LOW FO +	VOWEL U +	LOW KO +	HIGH VO +	VOWEL E +	HIGH NO	+ ꪀꪰꪙ
ꪏꪮꪉ +	ꪠ	◌ꪴ	ꪀ	ꪫ	ꪵ	ꪙ	+ ꪀꪰꪙ

data in visual order:

ꪏꪮꪉ +	LOW FO +	VOWEL U +	VOWEL E +	LOW KO +	HIGH VO +	HIGH NO	+ ꪀꪰꪙ
ꪏꪮꪉ +	ꪠ	◌ꪴ	ꪵ	ꪀ	ꪫ	ꪙ	+ ꪀꪰꪙ

correct rendering:	incorrect rendering:
ꪏꪮꪉ + ꪠꪴ + ꪵꪀꪫꪙ + ꪀꪰꪙ	ꪏꪮꪉ + ꪠꪴꪀ + ꪵꪫꪙ + ꪀꪰꪙ
/sɔŋ fu kʷɛn kan/	/sɔŋ fuk vɛn kan/

2. Ambiguity regarding the location of the syllable boundary in two syllable sequences involving LOW KO/HIGH NGO + HIGH VO and a combining or right-side vowel.

	Visual Order	Logical Order
Encoding Sequence:	$C + V + C_{(LowKo/HighNgo)} + HighVo + V_{(comb/right)}$	$C + V + C_{(LowKo/HighNgo)} + HighVo + V_{(comb/right)}$
Ambiguous?	yes	yes
Possible interpretations:	$C + V \;.\; C_{(LowKo/HighNgo)} + HighVo + V_{(comb/right)}$ or $C + V + C_{(LowKo/HighNgo)} \;.\; HighVo + V_{(comb/right)}$	$C + V \;.\; C_{(LowKo/HighNgo)} + HighVo + V_{(comb/right)}$ or $C + V + C_{(LowKo/HighNgo)} \;.\; HighVo + V_{(comb/right)}$
Affects rendering?	no (except for line breaking)	no (except for line breaking)
Affects sorting?	yes	yes
Affects line/word break?	yes	yes
Affects reading?	yes	yes
Ambiguity resolved by interword spacing?	yes	yes
Ambiguity resolved by adding tone mark or final consonant?	no	no

This sequence is similar to the first, but changes the last vowel to a combining mark or a right-side vowel.

Rendering is not a problem for this sequence. Even though the interpretation of the sequence remains ambiguous, the vowel will always be written above, below, or to the right of the HIGH VO. However, sorting and line/word breaking algorithms may have problems. One potential solution to the line breaking problem is to disallow line breaking wherever an ambiguous syllable boundary exists. But that still leaves the sorting problem unresolved.

The choice of storage order does not resolve the potential problem, because visual and logical order are the same.

Unlike ambiguities of Type 1, correct rendering does not resolve the ambiguity for the reader. The reader must rely on the context to determine the interpretation of the sequence.

Examples:

8) /haːʔ vaː/ ‘but’

data in visual or logical order:

LOW HO +	VOWEL AA +	LOW KO +	HIGH VO +	VOWEL AA
ꪪ	ꪱ	ꪀ	ꪩ	ꪱ

correct line breaking:	incorrect line breaking:
ꪭꪱꪀ + NewLine + ꪫꪱ	ꪭꪱ + NewLine + ꪀꪫꪱ
/haːʔ vaː/	/haː kʷaː/

9) /to kʷaːŋ/ ‘deer’

data in logical order:

LOW TO +	VOWEL O +	LOW KO +	HIGH VO +	VOWEL AA +	HIGH NGO
ꪔ	ꪶ	ꪀ	ꪫ	ꪱ	ꪉ

data in visual order:

VOWEL O +	LOW TO +	LOW KO +	HIGH VO +	VOWEL AA +	HIGH NGO
ꪶ	ꪔ	ꪀ	ꪫ	ꪱ	ꪉ

correct line breaking:	incorrect line breaking:
ꪶꪔ + NewLine + ꪀꪫꪱꪉ	ꪶꪔꪀ + NewLine + ꪫꪱꪉ
/to kʷaːŋ/	/tok vaːŋ/

3. Ambiguous interpretation of the one syllable visual sequence $V_{(Vowel\ E)}$ + $C_{(Velar)}$ + HIGH VO in some dialects.

	Visual Order	Logical Order
Encoding Sequence:	$V_{(Vowel\ E)} + C_{(Velar)} + HighVo$	$C_{(Velar)} + HighVo + V_{(Vowel\ E)}$ or $C_{(Velar)} + V_{(Vowel\ E)} + HighVo$
Ambiguous?	yes	no
Possible interpretations:	$C_{(Labialized)} + V_{(Vowel\ E)}$ (e.g. /kʷɛ/) or $C_{(Velar)} + V_{(Vowel\ E)} + HighVo$ (e.g. /kɛw/)	determined by the encoding
Affects rendering?	no	n/a
Affects sorting?	yes	n/a
Affects line/word break?	no	n/a
Affects reading?	yes, in some dialects	n/a
Ambiguity resolved by interword spacing?	no	n/a
Ambiguity resolved by adding tone mark or final consonant?	yes	n/a

The character TAI VIET LETTER HIGH VO has three functions: (i) to write the syllable initial high series /v/, (ii) to mark labialization of a velar consonant, and (iii) to write a syllable final /w/. Ambiguities 1 and 2, above, arise from functions (i) and (ii). Ambiguity 3 arises from functions (ii) and (iii). In the sequence in question, the HIGH VO character can be interpreted as marking labialization of the velar consonant, or as a final /w/.

Note that the context in which this ambiguity can occur is extremely limited.

- First, the vowel has to be the TAI VIET VOWEL E because:
  - There is no ambiguity when a combining vowel, right-side vowel, or digraph vowel is involved. The position of a combining mark or right-side vowel relative to the velar consonant and the HIGH VO will reveal the meaning of the HIGH VO.
  - Of the five left-side vowels:
    - Collocation restrictions in the language prevent TAI VIET VOWEL O and TAI VIET VOWEL EUA from occurring either before /w/ or after a labialized consonant.
    - TAI VIET VOWEL AY and TAI VIET VOWEL AUE already carry an inherent final consonant. Therefore, if these vowels occur in this context, the HIGH VO must be interpreted as marking labialization.
  - Thus we are left with an ambiguity only when the vowel is TAI VIET VOWEL E.
- Second, there cannot be any other final consonant. If there is, it reveals the HIGH VO to be a mark for labialization.
- Third, there cannot be a combining tone mark. If there is, its position relative to the consonants will reveal the meaning of the HIGH VO.

When all of these restrictions are taken into account, we are left with eight possible ambiguous spellings:

ꪹꪀꪝ	/kɛw/ or /k^{w}ɛ/ (low series)
ꪹꪁꪝ	/kɛw/ or /k^{w}ɛ/ (high series)
ꪹꪂꪝ	/k^{h}ɛw/ or /k^{hw}ɛ/ (low series)
ꪹꪃꪝ	/k^{h}ɛw/ or /k^{hw}ɛ/ (high series)
ꪹꪄꪝ	/xɛw/ or /x^{w}ɛ/ (low series)
ꪹꪅꪝ	/xɛw/ or /x^{w}ɛ/ (high series)
ꪹꪈꪝ	/ŋɛw/ or /$ŋ^{w}$ɛ/ (low series)

ꪵꪉꪫ /ŋɛw/ or /ŋʷɛ/ (high series)

If the author writes tone with the combining tone marks, the ambiguity is limited to syllables with tones 1 and 4. If tone is written with the spacing tone marks, or if tone is not written, the ambiguity can exist on syllables with any tone.

Using Điêu and Donaldson for Tai Don and Baccam, et al. for Tai Dam, I have identified the following words which use these spellings. Note that Điêu and Donaldson do not give the Tai spellings for the Tai Don. I have derived the Tai spelling from their Latin spelling. Baccam, et al., however, do give the indicated Tai spellings for the Tai Dam.

ꪵꪁꪫ	(T.Dam & T.Don)	/kɛw^1/ ‘Vietnamese’	
	(Tai Don)	/kɛw^2/ ‘a magpie’	
	(T.Dam & T.Don)	/kɛw^3/ ‘scissors’	
ꪵꪀꪫ	(Tai Don)		/k^{w}ɛ^4/ ‘crippled’
	(T.Dam & T.Don)		/k^{w}ɛ^5/ ‘cinnamon’
	(T.Dam & T.Don)	/kɛw^6/ ‘to chew’	
ꪵꪃꪫ	(Tai Don)		/k^{hw}ɛ^1/ ‘to brag or boast’
ꪵꪂꪫ	(Tai Don)	/k^{h}ɛw^6/ ‘to beg earnestly’	/k^{hw}ɛ^6/ ‘a young partridge’
ꪵꪅꪫ	(T.Dam & T.Don)	/xɛw^1/ ‘green or blue’	
	(T.Dam & T.Don)	/xɛw^2/ ‘section’	
	(T.Dam & T.Don)	/xɛw^3/ ‘tooth’	
ꪵꪄꪫ	(T.Dam & T.Don)	/xɛw^6/ ‘to grasp’	
ꪵꪉꪫꪵꪉꪫ	(Tai Don)	/ŋɛw^5 ŋɛw^5/ ‘(to work) feebly and slowly’	

In most of these words, the HIGH VO represents a final /w/, but there are a few words where it does mark labialization. So the potential ambiguity is realized in real life, and presents a problem for sorting processes. However, this problem exists in only some dialects of the script.

In other dialects, a device has been adopted to resolve this ambiguity. It consists of adding the character TAI VIET LETTER HIGH YO to the end of syllables that are pronounced $C_{\text{labialized-velar}}$ + /ɛ/. The result is that /k^{w}ɛ^5/, ‘cinnamon’, would be spelled ꪵꪀꪫꪥ. This convention was not used in Baccam, et al., but it has been adopted as a standard spelling in a project sponsored by the Son La Department of Science and Technology.

For those dialects which have not adopted the use of the HIGH YO as described above, the ambiguity could be resolved by using logical order. But that would introduce a potential data entry problem that is described in the next section.

Disambiguating logical order

A device might be introduced to the logical-order encoding to disambiguate the interpretation of the HIGH VO. The most likely approach would be to use a virama-like character to bind the HIGH VO to the velar consonant when the VO is used to indicate labialization. For example:

LOW SO + VOWEL U + LOW KO + virama + HIGH VO + VOWEL E + HIGH NO → ꪎꪴꪀ + ꪵꪫꪙ

would indicate the HIGH VO is bound to the LOW KO, so that the syllable boundary and the VOWEL E would be placed before the LOW KO. On the other hand, this sequence without the virama

LOW SO + VOWEL U + LOW KO + HIGH VO + VOWEL E + HIGH NO → ꪎꪴ + ꪵꪀꪫꪙ

would allow the syllable boundary to fall between the LOW KO and the HIGH VO.

The chief argument against employing such a device is that, since the effect of the virama is invisible in many instances, it will lead to a very high rate of keyboarding errors and inconsistency in the text. It is not a question of technical capability, nor of the order in which users prefer to type the text, but of how the user community views the text.

Although it is secondary to my argument here, I need to point out that the user community prefers to enter text in visual order. Typical of the comments that they make to me are:

1) Handwriting is done in visual order.
2) When telling someone how to spell a word, they describe it in visual order (specifying the left side vowel first).
3) Previous implementations of the Tai Viet script, on both computer and typewriter, have used visual order.
4) Among the portion of the community that is familiar with Lao, they are accustomed to typing in visual order.
5) They would definitely want to keyboard in the order that they see on the screen. To go by sound rather than sight would cause a lot of confusion.

The important thing here is not that they prefer to keyboard in visual order, but that they view their written language from a visual rather than a phonetic perspective. The result is that when a Tai author keyboards a string of text such as

ꪶꪔꪀꪫꪱꪉ

he would not expect to make any distinction in how he keys it based on pronunciation. So even though an input method may provide distinctive key sequences for

LOW TO + VOWEL O + LOW KO + virama + HIGH VO + VOWEL AA + HIGH NGO → ꪶꪔ + ꪀꪫꪱꪉ → ꪶꪔꪀꪫꪱꪉ

vs.

LOW TO + VOWEL O + LOW KO + HIGH VO + VOWEL AA + HIGH NGO → ꪶꪔꪀ + ꪫꪱꪉ → ꪶꪔꪀꪫꪱꪉ

that distinction is foreign to the script. The user will often fail to employ the distinctive sequences. He is likely to choose whichever key sequence is easiest to use, and use it for both syllables. Unless a line break occurs between these syllables, the effect of the virama will be invisible. Thus, the error will often go unnoticed and uncorrected.

Furthermore, the use of a hidden character as a flag is undesirable. Past experience has shown that these flags can be deleted without the typist knowing it, or they can be left behind in a text when the base character is deleted, resulting in inconsistent text and unpredictable behavior.

Summary

It might be said that visual order is faithful to the script in that it leaves the ambiguities that exist in the orthography in the encoding as well. These ambiguities (numbers 2 and 3) are orthographic in nature. After the text is rendered, these ambiguities will still exist on the printed page.

The use of logical order by itself only resolves ambiguity 3 while it introduces the much more serious ambiguity 1. Another device such as a virama must be added to the encoding to resolve ambiguities 1 and 2.

The use of logical order with a virama forces the resolution of all ambiguities in the encoding, but ambiguities 2 and 3 are then reintroduced to the output by the rendering process. Logical order with a virama also forces the typist to make distinctions in his key selection which he will not normally make in writing the script. This will likely result in many data entry errors which will not be visible to the user.

On balance, I believe that visual order will leave us with fewer and less severe problems than either plain logical order or logical order with a virama, and will enable a more usable computing solution for the Tai Viet script than the alternatives.

Appendix 2—Line breaking rules

This is an initial draft of the line breaking rules for the Tai Viet Script. These rules apply when a text does not have inter-word spacing, which would be the case with the oldest tradition of the script. However, these rules should not be applied to text that has inter-word spacing, as this may result in undesirable line breaks within the small number of two-syllable words that occur in the language, or of polysyllabic loan words.

Additional study is needed to determine whether these rules are accurate and adequate.

1. Punctuation rule.
   Any initial or final punctuation clusters with the syllable or symbol.
2. Symbol rule.
   A line break can always occur before or after the characters:
   TAI VIET SYMBOL KON
   TAI VIET SYMBOL NUENG
   TAI VIET SYMBOL SAM.
   TAI VIET SYMBOL HO HOI
   TAI VIET SYMBOL KOI KOI
3. Syllable rules.
   A break can occur at the beginning of any syllable. Use the following rules to identify the first character of a syllable.
   3.1. Identify the vowel and the initial consonant cluster.
   The vowel characters can be grouped into the following classes:
   - V_1 — left-side vowels
     TAI VIET VOWEL E
     TAI VIET VOWEL O
     TAI VIET VOWEL UEA
     TAI VIET VOWEL AUE
     TAI VIET VOWEL AY
   - V_2 — combining vowels
     TAI VIET MAI KANG
     TAI VIET VOWEL I
     TAI VIET VOWEL UE
     TAI VIET VOWEL U

TAI VIET MAI KHIT
TAI VIET VOWEL IA
TAI VIET VOWEL AM

- V_3 — right-side vowels

  TAI VIET VOWEL AA
  TAI VIET VOWEL UA
  TAI VIET VOWEL AN
  TAI VIET LETTER LOW O

The initial consonant cluster, C_i, can be:

- A labialized consonant consisting of any velar consonant + TAI VIET LETTER HIGH VO. The velar consonants include:

  TAI VIET LETTER LOW KO
  TAI VIET LETTER HIGH KO
  TAI VIET LETTER LOW KHO
  TAI VIET LETTER HIGH KHO
  TAI VIET LETTER LOW KHHO
  TAI VIET LETTER HIGH KHHO
  TAI VIET LETTER LOW NGO
  TAI VIET LETTER HIGH NGO

  Always form a labialized cluster if possible. But a labialized cluster can never be formed if vowel or tone mark can occurs between the velar consonant and the HIGH VO.

- Any single character in the range AA00..AA2F.

These may occur in any one of the following patterns:

- $V_1 + C_i + V_2?$
- $V_1 + C_i + T? + V_3?$
- $C_i + V_2$
- $C_i + T? + V_3$
- $C_i + T? +$ TAI VIET LETTER BO LOW + TAI VIET VOWEL AM

where

T = one of the following combining tone marks

TAI VIET TONE MAI EK
TAI VIET TONE MAI THO

3.2. Left-side vowel rule.

If the vowel includes one of the left-side vowel characters, a break can occur before it. (This rule takes precedence over the Initial consonant rule.)

3.3. Initial consonant rule.

3.3.1. Do not break before initial clusters with LOW KO and HIGH NGO (TAI VIET LETTER LOW KO or TAI VIET LETTER HIGH NGO may in fact be the final consonant of a preceding syllable. Additional rules can be added here to determine the syllable boundary.)

3.3.2. A break may be made before any other initial cluster.

Appendix 3—Script samples

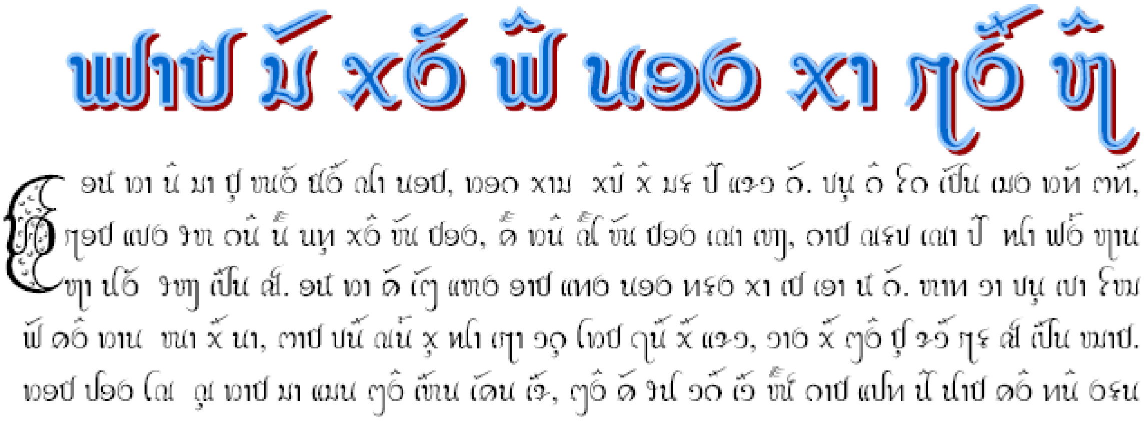

Figure 1—A modern text from Son La.

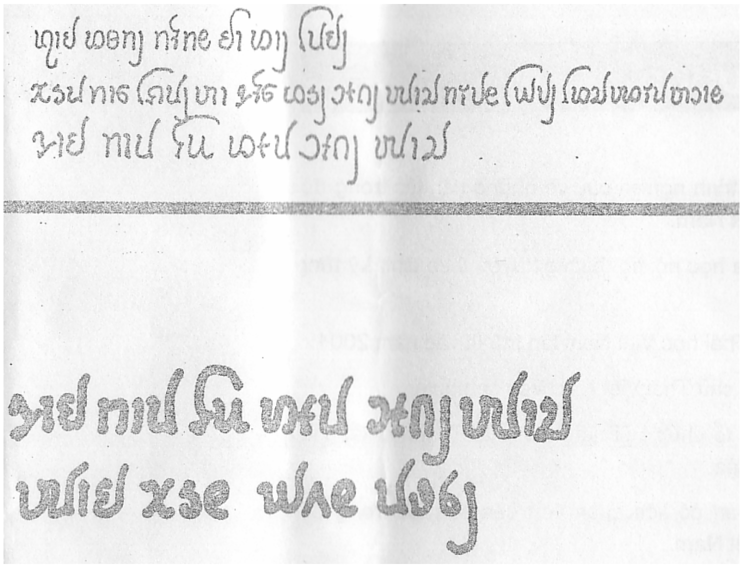

Figure 2—From *Giới Thiệu Chương Trình Thái Học Việt Nam*, 1999.

ꪀꪱꪶꪊ꪿ꪚ	ກະຈົບ	ca-chốp	hoe
ꪀꪱꪊ꪿ꪮꪚ	ກະຈອບ	ca-chóp	hoe
ꪀꪱꪒ꫁ꪱꪉ	ກະດ້າງ	ca-dảng	stiffened, frozen
ꪀꪱꪵꪒ꫁ꪚ	ກະເດັບ	ca-đếp	basket with strap
ꪀꪱꪔꪱ	ກະຕາ	ca-ta	even if
ꪀꪱꪔꪱꪀ	ກະຕາກ	ca-tác	to cackle
ꪀꪱꪵꪚꪣ	ກະແບມ	ca-bem	coffer
ꪀꪰꪉ	ກັງ	căng	ape
ꪀꪰꪒ	ກັດ	cắt	to gnaw

Figure 3—From Baccam et al., p 13. The left hand column is Tai Dam.

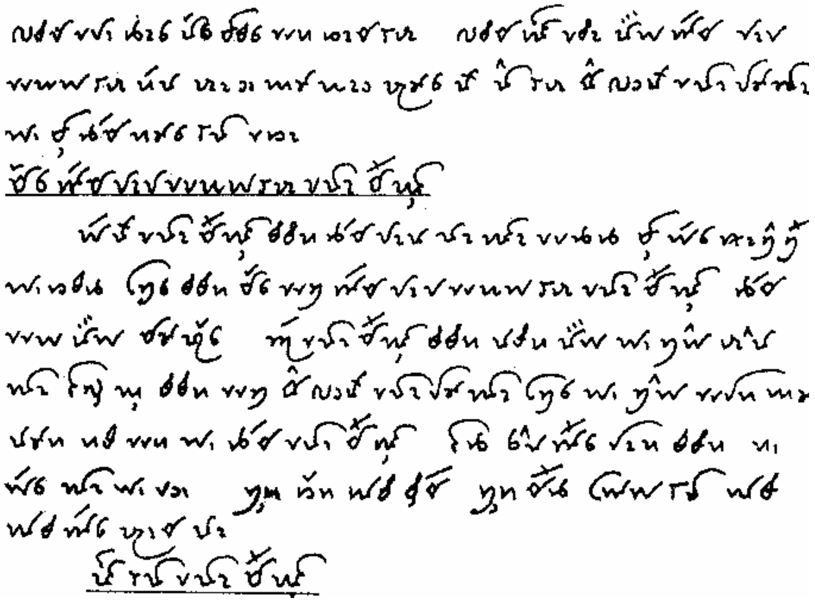

Figure 4—From *Khhăm Kháo Đi Chảu Dê-su Seo Lũng Ók Mác Tẻm*, 1983. A handwritten text in Tai Don.

Bibliography

___. "Các mẫu tự Thái ở miền Tây Bắc Việt Nam." Internet: http://www.evertype.com/standards/tai/viet-thai-samples.pdf

___. Song Petburi font. Internet: http://www.seasite.niu.edu/tai/TaiDam/index.htm.

Baccam Don, Baccam Faluang, Baccam Hung, Dorothy Fippinger. 1989. ꪜꪰꪚ꫁ ꪵꪚ꫁ꪙꪮꪀ ꪁꪫꪱꪣꪼꪕ – ꪮꪲꪉꪁꪲꪒ, *Tai Dam – English, English – Tai Dam Vocabulary Book*. Summer Institute of Linguistics.

Cầm Trọng. 2005. "Thai Scripts in Vietnam," in *Workshop on the Preservation and Digitization of Tai Scripts*. Hanoi, Vietnam.

Điêu Chính Nhìm and Jean Donaldson. 1970. *Păp San Khhãm Pák Tãy-Keo-Eng* (Ngũ-Vụng Thái-Việt-Anh, Tai-Vietnamese-English Vocabulary). Saigon.

Gordon, Raymond G., Jr. (ed.), 2005. Ethnologue: Languages of the World, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/

Lò Văn Mười (ꪨꪮ ꪫꪽ꫁ ꪹꪣꪷ꫁). 1966. ꪵꪮꪚ ꪏꪳ ꪼꪕ ꪚꪸꪙ ꪵꪚꪉ. (Ép Sư Táy Piên Peng, Learning the Revised Tai Alphabet.)

SIL International. Tai Heritage font. Internet: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&item_id=SILTD_home

Workshop on the Preservation and Digitization of Tai Scripts. Hanoi, Vietnam. 2005.

Workshop on Encoding and Digitalizing Thai Scripts. Điện Biên, Vietnam. 2006.

Acknowledgements

Many people and organizations contributed to this project. UNESCO sponsored workshops on the Tai Viet script in Vietnam in 2005 and 2006. Japan's Advanced Institute of Science and Technology (AIST) sponsored Tai Viet research for open source software. The Scripts Encoding Initiative helped to fund my travel to the 2006 workshop.

The Tai Viet font we used in this proposal was the work of Lò Mai Cương.

Ngô Trung Việt and [James] Đỗ Bá Phước worked hard to organize the workshops and the TayVN Working Group, and have served as an indispensable liaison between the Tai community in Vietnam and me. Siang Bacthi and Dorothy Fippinger have likewise served as my primary contacts with the Tai Dam community in Des Moines, Iowa.