Unified Tai Script for Unicode

Ngo Trung Viet

Institute of Information Technology
Vietnamese Academy of Science and Technology

&

Jim Brase

SIL International

January 30, 2006

Referred to as Viet Thai, row AA in the latest Roadmap
See also: http://www.evertype.com/standards/tai/viet-thai-comparisons.pdf

Sociolinguistic Background

The Tai script is used by four Tai languages spoken primarily in northwestern Vietnam, northern Laos, and central Thailand—Tai Daeng (also Red Tai or Tai Rouge), Tai Dam (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* estimates the total population of the four languages, across all countries, at 1.5 million. (Tai Daeng 165,000, Tai Dam 764,000, Tai Dón 490,000, Thai Song 32,000.)

The degree of usage of the script varies from community to community. It has been widely used by the Tai Dam community in the United States. There is a desire to introduce it into formal education in Vietnam (Cam Trong 2005). On the other hand, it is not known whether it is in current use by the Thai Song, and the only dated document available from Laos is a 30-year old Tai Daeng manuscript.

The Traditional Script vs. the Unified Alphabet

Anyone attempting to establish a standard for writing the Tai script must cope with the great diversity between communities in the traditional form of the script. Cam Trong (2005) lists eight dialects of the script for Vietnam alone. Other dialects exist in Laos and Thailand.

The Vietnamese government has attempted to establish a standard for the Tai script, which was called the *Thống Nhất*, or Unified Alphabet in an anonymous 1961 paper (*Các Mẫu Tự Thái Ở*

Miền Tây Bắc Viêt Nam). This standard will be referred to as the Unified Alphabet in the remainder of this paper. The most recent revision of the Unified Alphabet is found in Cam Trong (2005).

Not everyone has had the opportunity to learn the Unified Alphabet. This would include the elderly, who learned to read and write before the Unified Alphabet was introduced, and those Tai communities outside of Vietnam, including the Tai Daeng of Laos, the Thai Song of Thailand, Tai Dam communities in Laos, the United States, and France, and many smaller communities. Thus, it is my desire to include the traditional forms of the script as well as the Unified Alphabet in this proposal.

The Tai Writing System

Basic Features

The Tai scripts share many features common to most Thai alphabets:

- They are written left to right. (One variation of the script, Tai Do, is written vertically, but is beyond the scope of this study.)
- There is a double set of initial consonants, one for high tone class and one for low tone class.
- In the traditional form, vowel marks can be placed before, after, above, or below the syllable's initial consonant, depending on the vowel. Vowel digraphs are common. In the Unified Alphabet, the diacritic vowels have been replaced with spacing vowels.

Tone Classes and Tone Marks

In the Tai scripts each consonant has two forms. The high form of the initial consonant indicates that the syllable uses tone 1, 2, or 3. The low form of the initial consonant indicates that the syllable uses tone 4, 5, or 6. (Tai Daeng has only five tones, but the practice is similar.)

Traditionally, these scripts did not use any further marking for tone, and the reader had to determine the tone from the context. In recent times, however, several groups have introduced tone marks into Tai writing. The Tai Heritage font (Tai Dam) borrowed tone marks from Lao, and these are now widely used by the Tai Dam community in the U.S. The Song Petburi font (Thai Song) includes Thai style tone marks, which are identical to the Lao. The Unified Tai Alphabet invented a new set of spacing tone marks which are placed at the end of the syllable. Aam and Aanu (1974) present a unique set of diacritic tone marks for Tai Daeng.

When combined with the consonant class, two tone marks are sufficient to unambiguously mark the tone. Thus, some authors mark tone in Tai Dam as follows:

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

Note, however, that checked syllables (those ending /p/, /t/, /k/, or /ʔ/) are restricted to tones 2 and 5, and that no marking other than the consonant class is necessary for those syllables.

The practice for the other languages would be similar to that for Tai Dam.

Final Consonants

In written form, the high-tone class symbols for 'b' (ᦸ)[1] and 'd' (ᦷ) are used for syllable final /p/ and /t/, as is the practice in all Thai scripts. This usage should not mislead one into thinking that oral /b/ and /d/ occur syllable final.

The high-tone class symbol for 'k' (ᦺ) is used for both final /k/ and final /ʔ/.

The low-tone class symbols are used for writing final /j/ (ᦻ) and the final nasals, /m/ (ᦵ), /n/ (ᦶ), and /ŋ/ (ᦴ). Low-tone /v/ (ᦼ) is used for final /w/.

There are a number of exceptions to the above rules, in the form of Vowel + Final Consonant ligatures. These vary from region to region, but the ones with the broadest usage are the ligatures for /-aj/ (ᦋ◌), /-am/ (◌᧶), /-an/ (◌ᦶ), and /-əw/ (ᦎ◌). The ligature /-at/ (◌ᦾ) is limited to some dialects of Tai Dón.

Word Spacing and Baseline

Traditional Tai writing does not use space between words. Thus, a line breaking algorithm will have to be developed to accommodate the oldest forms of the script.

In the last 20 years the Tai Dam community in the U.S. has adopted the practice of using word spacing, although the spaces are usually narrower than for Latin alphabets. A trilingual pamphlet published by the Hanoi National University in 1999, *Giới Thiệu Chương Trình Thái Học Việt Nam,* shows spacing between words in the Tai script. (See Figure 1 in Script Samples, below.)

The Tai Daeng sample (Figure 2, Script Samples) has clear spacing between words. This is a surprise, as the manuscript appears to be rather old.

Tai scripts usually use a bottom baseline. But the Tai Daeng manuscript in Figure 2, written on lined paper, again surprises us with a center baseline.

Sort Order

The Tai scripts do not have an established standard for sorting. Sequences have sometimes been borrowed from neighboring languages. Baccam, et. al. (1989) use an order borrowed from Lao. On the other hand, Cam Trong (2005) preferred an order based on the Vietnamese alphabet (the Quốc Ngữ). These will be discussed further, below.

Key Issues

Is it sufficient to encode only the Unified Alphabet?

This is the most crucial question to be answered. My conclusion is "No, it is not sufficient," for the following reasons.

1. As noted above, not everyone one can read the Unified Alphabet. Some communities will try to continue using their traditional form of the script.
2. One possible solution is to encode only the Unified Alphabet, and then to make language-specific fonts for each of the languages which reflect their traditional form. Thus, a Tai Dón person would use a Tai Dón font, and a Tai Dam person would use a

[1] The samples of the symbols shown in this section, except for the /-at/ ligature, are from the Tai Heritage font (Tai Dam). Forms may vary across script dialects.

Tai Dam font, but they would have the same encoding. However, this would result in an encoding that is interpreted by the font, which defeats the purpose of Unicode.

Therefore, I have concluded that it is necessary to encode the Unified Alphabet plus any characters that are required by the traditional forms of the script.

Should Tai Daeng be included?

The Tai Daeng character set has only about a 50% correlation to those of the other languages. Should it be included as part of the Unified Tai Script, or should it be encoded as its own script?

Although it has many unique characters, the basic form and mechanics of the script are similar to that of the other languages. Furthermore, encoding Tai Daeng as a separate script would require the duplication of those characters which are similar. Consequently, Tai Daeng should be considered part of the Unified Tai Script.

Should Thai Song be included?

The contrast between the styles of Thai Song writing and Tai Dam writing is quite stark. Yet when the stylistic differences are set aside, the underlying form of many of the characters are similar. Therefore, at this time Thai Song should be considered part of the Unified Tai Script.

Unfortunately, only a limited amount of data was available for evaluation for this proposal. Some new data has recently become available, but has not yet been analyzed. It is possible that additional analysis will lead to different conclusions about the Thai Song writing.

Character Order and Sort Order

As noted above under **The Tai Writing System**, the Tai script does not have an established sort order. The two best options are an order derived from Lao or one derived from the Quốc Ngữ.

The advantage of an order based on the Quốc Ngữ is that the majority of the users of the Tai script live in an area where Vietnamese is the language in influence. Communication between the Tai dialects and Vietnamese would thus be enhanced. This would help to encourage the teaching of the Tai script in the schools—an important consideration.

The advantage of a Lao based order is that Tai and Lao scripts are from the same family.

This matter will require additional discussion. The order of the currently suggested character chart is based on the Lao.

Bibliography

___. "Các Mẫu Tư Thái Ở Miền Tây Bắc Viêt Nam." Internet: http://www.evertype.com/standards/tai/viet-thai-samples.pdf

___. *Conflict and Relocation.* http://www.seasite.niu.edu/tai/TaiDam/article/a4.htm.

___. 1999. *Giới Thiệu Chương Trình Thái Học Việt Nam.* ([illegible]; Introduction for Vietnam Programme of Thai Studies). Hanoi National University, Hanoi.

___. Khhãm Kháo Đi Chảu Dê-su Seo Lũng Ók Mác Tẻm.

___. *Sip Hoc Chau Thai.* (Journal de l'Association A.R.T.F.) Orly, France. Cited by http://www.evertype.com/standards/tai/viet-thai-samples.pdf.

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

___. Untitled, undated manuscript in Tai Daeng.

[illegible] and [illegible] (Aam and Aanu). 1974. [illegible] ("bɛ:p hiə:n4 a:n2 lɛʔ khiən1 naŋ1 sɨ:1 taj4 dɛŋ1", "Learning to Read and Write Tai Daeng"). Vientiane.

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

Cam Trong. 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.

Donaldson, Jean and Jerold A. Edmondson. 1997. "A Preliminary Examination of Tai Tac," in *Comparative Kadai, The Tai Branch,* edited by Jerold A. Edmondson and David B. Solnit. The Summer Institute of Linguistics and The University of Texas at Arlington. pp 235-266.

Everson, Michael. Proposal for the Universal Character Set. Internet: http://www.evertype.com/standards/tai/viet-thai-comparisons.pdf

Ferlus, Michel. 1988. "Langues et ecritures en asie du sud-est," The 21st International Conference on Sino-Tibetan Languages and Linguistics, University of Lund, Sweden.

Ferlus, Michel. 1999. "Les dialectes et les écritures des Tai (Thai) du Nghệ An (Vietnam)," *Treizièmes journées de linguistique d'asie orientale.* Centre de recherches linguistiques sur l'asie orientale (EHESS-CNRS), 105 bd raspail, 75006 Paris.

Ferlus, Michel. 2003 "L'intérêt linguistique du Hưng Hóa Ký Lược de Phạm Thậ Dụat," Dix-septièmes de Linguistiques d'Asie Orientale, Centre de Recherches Linguistiques sur l'Asie Orientale (EHESS-CNRS).

Finot, Louis, 1917. "Recherhes sur la literature laotienne," *Bulletin de l'Ecole Française d'Extrême-Orient.* 17(5).

Fippinger, Jay and Dorothy. 1970. "Black Tai Phonemes with reference to White Tai." *Anthropological Linguistics* 12.3: 83-97.

Lafont, Pierre-Bernard, 1962. "Les écritures 'Tay du Laos," Bulletin de l'Ecole Français d'Extrême-Orient.

Gedney, William J. 1989. “A Comparative Sketch of White, Black and Red Tai.” *Selected Papers on Comparative Tai Studies,* Michigan Papers on South and Southeast Asia no. 29, Center for South and Southeast Asia Studies, The University of Michigan.

Lo Văn Mươi (ꪩꪮ ꪫꪽꪙ ꪣꪳ꪿ꪒ). 1966. Ép Sư 'Táy Piên Peng. (ꪵꪮꪚ ꪎꪳ꪿ ꪼꪕ ꪜꪸ꪿ꪙ ꪵꪜꪉ, Learning the Revised Tai Alphabet.)

Marcus, Russell. 1970. *English-Lao, Lao-English Dictionary.* Charles E. Tuttle Company.

Martini, Francois. 1954. “Romanisation des parlers ‘Tay du Nord Vietnam.” *Bulletin de l’Ecole Française d’extrême-orient.*

Minot, Lieutenant. 1933. "Dictonnaire Français - Thay Blanc." Mường Té.

Minot, Georges. 1940. "Dictonnaire Tẫy Blanc-Français." *Bulletin de l'Ecole d'Extrême-Oriente*, t. XL.

Ngo Trung Viet. 2005. “ICT for Thai Ethnic Culture and Education Multilingual Projects,” in *Workshop on the Preservation and Digitization of Tai Scripts.* Hanoi, Vietnam.

Phan Anh Dung and Ngo Trung Viet. 2005. “Technical Design, Software for Inputting and Displaying Vietnam Thai Scripts,” in *Workshop on the Preservation and Digitization of Tai Scripts.* Hanoi, Vietnam.

Robert, R. 1941. *Notes sur les Tay Dèng de Lang Chánh (Thanh-Hoá, Annam).* Institut Indochinois pour l'Etude de L'Homme, mémoire n' 1. Hanoi: Imprimerie d'Extrême-Orient.

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

Whitehouse, Ruth. 1975. *Phonemic Write-up, Lao Song Language*. Unpublished paper.

ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646
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:** ***Unified Tai (referred to as Viet Thai, row AA in the latest Roadmap)***
see also: http://www.evertype.com/standards/tai/viet-thai-comparisons.pdf
2. Requester's name: *Ngo Trung Viet, Institute of Informatin Technology, VAST*
Jim Brase, SIL International
3. Requester type (Member body/Liaison/Individual contribution): *Individual contribution*
4. Submission date: *February 1, 2006*
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: *Unified Tai*
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: *124*
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? *shapes currently provided are a mixture of styles. We are waiting for a type designer to provide a uniform font.*
6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard? *SIL International*
If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:
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?
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)?

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.

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? *private individuals from Tai Dam community in United States Contact with community in Vietnam through Dr. Ngo Trung Viet*
   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: *Vietnam and United States. Uncertain about Laos and Thailand*
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?
   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:

UNIFIED TAI

	xx0	xx1	xx2	xx3	xx4	xx5	xx6	xx7
0								
1								
2								
3								
4								
5								
6								
7								
8								
9								
A								
B								
C								
D								
E								
F								

Names Table

Consonants & Symbols—Unified Tai Alphabet

xx00		UNIFIED TAI LETTER KO HIGH
xx01		UNIFIED TAI LETTER KO LOW
xx02		UNIFIED TAI LETTER KHO HIGH
xx03		UNIFIED TAI LETTER KHO LOW
xx04		UNIFIED TAI LETTER KHHO HIGH
xx05		UNIFIED TAI LETTER KHHO LOW
xx06		UNIFIED TAI LETTER GO HIGH
xx07		UNIFIED TAI LETTER GO LOW
xx08		UNIFIED TAI LETTER NGO HIGH
xx09		UNIFIED TAI LETTER NGO LOW
xx0A		UNIFIED TAI LETTER CO HIGH
xx0B		UNIFIED TAI LETTER CO LOW
xx0C		UNIFIED TAI LETTER CHO HIGH
xx0D		UNIFIED TAI LETTER CHO LOW
xx0E		UNIFIED TAI LETTER SO HIGH
xx0F		UNIFIED TAI LETTER SO LOW
xx10		UNIFIED TAI LETTER NHO HIGH
xx11		UNIFIED TAI LETTER NHO LOW
xx12		UNIFIED TAI LETTER DO HIGH
xx13		UNIFIED TAI LETTER DO LOW
xx14		UNIFIED TAI LETTER TO HIGH
xx15		UNIFIED TAI LETTER TO LOW
xx16		UNIFIED TAI LETTER THO HIGH
xx17		UNIFIED TAI LETTER THO LOW
xx18		UNIFIED TAI LETTER NO HIGH
xx19		UNIFIED TAI LETTER NO LOW
xx1A		UNIFIED TAI LETTER BO HIGH
xx1B		UNIFIED TAI LETTER BO LOW
xx1C		UNIFIED TAI LETTER PO HIGH
xx1D		UNIFIED TAI LETTER PO LOW
xx1E		UNIFIED TAI LETTER PHO HIGH
xx1F		UNIFIED TAI LETTER PHO LOW
xx20		UNIFIED TAI LETTER FO HIGH
xx21		UNIFIED TAI LETTER FO LOW
xx22		UNIFIED TAI LETTER MO HIGH

xx23		UNIFIED TAI LETTER MO LOW
xx24		UNIFIED TAI LETTER YO HIGH
xx25		UNIFIED TAI LETTER YO LOW
xx26		UNIFIED TAI LETTER RO HIGH
xx27		UNIFIED TAI LETTER RO LOW
xx28		UNIFIED TAI LETTER LO HIGH
xx29		UNIFIED TAI LETTER LO LOW
xx2A		UNIFIED TAI LETTER VO HIGH
xx2B		UNIFIED TAI LETTER VO LOW
xx2C		UNIFIED TAI LETTER HO HIGH
xx2D		UNIFIED TAI LETTER HO LOW
xx2E		UNIFIED TAI LETTER O HIGH
xx2F		UNIFIED TAI LETTER O LOW
xx30		UNIFIED TAI SYMBOL KON (Person)
xx31		UNIFIED TAI SYMBOL NEUNG (One)
xx32		UNIFIED TAI SYMBOL SAM (Repetition)

Consonants & Symbols—Additions used by two or more languages

xx33		UNIFIED TAI LETTER ALTERNATE KO LOW
xx34		UNIFIED TAI LETTER ALTERNATE CO LOW
xx35		UNIFIED TAI LETTER ALTERNATE SO HIGH
xx36		UNIFIED TAI LETTER ALTERNATE NHO LOW
xx37		UNIFIED TAI LETTER TAI DAENG NHO LOW
xx38		UNIFIED TAI LETTER ALTERNATE THO LOW
xx39		UNIFIED TAI LETTER ALTERNATE FO LOW
xx3A		UNIFIED TAI LETTER ALTERNATE YO HIGH
xx3B		UNIFIED TAI LETTER ALTERNATE YO LOW
xx3C		UNIFIED TAI LETTER ALTERNATE LO LOW

Consonants & Symbols—Tai Daeng additions

xx3D		UNIFIED TAI LETTER TAI DAENG KO LOW
xx3E		UNIFIED TAI LETTER TAI DAENG KO ALTERNATE
xx3F		UNIFIED TAI LIGATURE TAI DAENG KN
xx40		UNIFIED TAI LIGATURE TAI DAENG KW
xx41		UNIFIED TAI LETTER TAI DAENG KHO HIGH
xx42		UNIFIED TAI LETTER TAI DAENG NGO HIGH
xx43		UNIFIED TAI LETTER TAI DAENG NGO LOW
xx44		UNIFIED TAI LETTER TAI DAENG SO LOW
xx45		UNIFIED TAI LETTER TAI DAENG NHO HIGH
xx46		UNIFIED TAI LETTER TAI DAENG DO HIGH

xx47		UNIFIED TAI LETTER TAI DAENG PO LOW
xx48		UNIFIED TAI LETTER TAI DAENG FO ALTERNATE
xx49		UNIFIED TAI LETTER TAI DAENG YO
xx4A		UNIFIED TAI LIGATURE TAI DAENG HO YO
xx4B		UNIFIED TAI LETTER TAI DAENG VO HIGH
xx4C		UNIFIED TAI LETTER TAI DAENG VO LOW

Consonants & Symbols—Tai Dam additions

xx4D		UNIFIED TAI LETTER TAI DAM THO LOW

Consonants & Symbols—Tai Don additions

xx4E		UNIFIED TAI LETTER TAI DON KO LOW
xx4F		UNIFIED TAI LETTER TAI DON NGO HIGH
xx50		UNIFIED TAI LETTER TAI DON SO LOW
xx51		UNIFIED TAI LETTER TAI DON DO HIGH
xx52		UNIFIED TAI LETTER TAI DON FO LOW
xx53		UNIFIED TAI LETTER TAI DON MO HIGH

Consonants & Symbols—Thai Song additions

xx54		UNIFIED TAI LETTER THAI SONG KHO HIGH

Vowels & Tones—Unified Tai Alphabet

xx55		UNIFIED TAI VOWEL COMBINING A
xx56		UNIFIED TAI VOWEL SPACING A
xx57		UNIFIED TAI VOWEL AA
xx58		UNIFIED TAI VOWEL RAISED A
xx59		UNIFIED TAI VOWEL COMBINING I
xx5A		UNIFIED TAI VOWEL SPACING I
xx5B		UNIFIED TAI VOWEL COMBINING UE
xx5C		UNIFIED TAI VOWEL SPACING UE
xx5D		UNIFIED TAI VOWEL COMBINING U
xx5E		UNIFIED TAI VOWEL SPACING U
xx5F		UNIFIED TAI VOWEL SPACING E
xx60		UNIFIED TAI VOWEL EH
xx61		UNIFIED TAI VOWEL O
xx62		UNIFIED TAI VOWEL UH
xx63		UNIFIED TAI VOWEL COMBINING IA
xx64		UNIFIED TAI VOWEL SPACING IA
xx65		UNIFIED TAI VOWEL UEA
xx66		UNIFIED TAI VOWEL UA
xx67		UNIFIED TAI VOWEL UHW
xx68		UNIFIED TAI VOWEL AY

xx69		UNIFIED TAI VOWEL AN
xx6A		UNIFIED TAI VOWEL AM
xx6B		UNIFIED TAI TONE COMBINING MAI EK
xx6C		UNIFIED TAI TONE SPACING MAI EK
xx6D		UNIFIED TAI TONE COMBINING MAI THO
xx6E		UNIFIED TAI TONE SPACING MAI THO

Vowels & Tones—Tai Daeng additions

xx6F		UNIFIED TAI VOWEL TAI DAENG A
xx70		UNIFIED TAI VOWEL TAI DAENG II
xx71		UNIFIED TAI VOWEL TAI DAENG UE
xx72		UNIFIED TAI VOWEL TAI DAENG UUE
xx73		UNIFIED TAI VOWEL TAI DAENG U
xx74		UNIFIED TAI VOWEL TAI DAENG UU
xx75		UNIFIED TAI VOWEL TAI DAENG EE
xx76		UNIFIED TAI VOWEL TAI DAENG SHORT O
xx77		UNIFIED TAI VOWEL TAI DAENG UUA
xx78		UNIFIED TAI VOWEL TAI DAENG SHORT UH

Vowels & Tones—Tai Don additions

xx79		UNIFIED TAI VOWEL TAI DON A
xx7A		UNIFIED TAI VOWEL TAI DON AT
xx7B		UNIFIED TAI VOWEL LOW TONE AA

Character Properties

code value/ range	Rep Glyph	Unicode Character Name	Gen Cat	Can Comb Class	Bidi Cat	Char Decomp	Dec Dig Val	Dig Val	Num Val	Mirr'd	U 1.0 Name	10646 Com	Upper Case Equiv	Lwr Case Equiv	Title Case Equiv
xx00 ..xx2F			Lo	0	L					N					
xx30	ᩋ	UNIFIED TAI SYMBOL KON (Person)		0	L					N					
xx31	ᩍ	UNIFIED TAI SYMBOL NEUNG (One)		0	L				1	N					
xx32	ᩈ	UNIFIED TAI SYMBOL SAM (Repetition)		0	L					N					
xx33 ..xx54			Lo	0	L					N					
xx55	◌ᩴ	UNIFIED TAI VOWEL COMBINING A	MN	230	NSM					N					
xx56	◌ᩣ	UNIFIED TAI VOWEL SPACING A	Lo	0	L					N					
xx57	◌ᩱ	UNIFIED TAI VOWEL AA	Lo	0	L					N					
xx58	◌ᩤ	UNIFIED TAI VOWEL RAISED A	Lo	0	L					N					
xx59	◌ᩮ	UNIFIED TAI VOWEL COMBINING I	MN	230	NSM					N					
xx5A	◌ᩥ	UNIFIED TAI VOWEL SPACING I	Lo	0	L					N					
xx5B	◌ᩯ	UNIFIED TAI VOWEL COMBINING UE	MN	230	NSM					N					
xx5C	ᩉᩬ	UNIFIED TAI VOWEL SPACING UE	Lo	0	L					N					
xx5D	◌ᩩ	UNIFIED TAI VOWEL COMBINING U	MN	220	NSM					N					
xx5E	◌ᩧ	UNIFIED TAI VOWEL SPACING U	Lo	0	L					N					
xx5F	ᩉ◌	UNIFIED TAI VOWEL SPACING E	Lo	0	L					N					
xx60	ᩍ◌	UNIFIED TAI VOWEL EH	Lo	0	L					N					
xx61	ᩋ◌	UNIFIED TAI VOWEL O	Lo	0	L					N					
xx62	ᩥ◌	UNIFIED TAI VOWEL UH	Lo	0	L					N					
xx63	◌ᩰ	UNIFIED TAI VOWEL COMBINING IA	MN	230	NSM					N					
xx64	◌ᩢ	UNIFIED TAI VOWEL SPACING IA	Lo	0	L					N					
xx65	ᩁ◌	UNIFIED TAI VOWEL UEA	Lo	0	L					N					
xx66	◌ᩧ	UNIFIED TAI VOWEL UA	Lo	0	L					N					
xx67	ᩉ◌	UNIFIED TAI VOWEL UHW	Lo	0	L					N					
xx68	ᩋᩣ	UNIFIED TAI VOWEL AY	Lo	0	L					N					
xx69	◌ᩱᩴ	UNIFIED TAI VOWEL AN	Lo	0	L					N					

xx6A	◌	UNIFIED TAI VOWEL AM	MN	230	NSM					N					
xx6B	◌	UNIFIED TAI TONE COMBINING MAI EK	MN	230	NSM					N					
xx6C	e	UNIFIED TAI TONE SPACING MAI EK	Lo	0	L					N					
xx6D	◌	UNIFIED TAI TONE COMBINING MAI THO	MN	230	NSM					N					
xx6E	J	UNIFIED TAI TONE SPACING MAI THO	Lo	0	L					N					
xx6F	◌	UNIFIED TAI VOWEL TAI DAENG A	Lo	0	L					N					
xx70	◌	UNIFIED TAI VOWEL TAI DAENG II	MN	230	NSM					N					
xx71	◌	UNIFIED TAI VOWEL TAI DAENG UE	MN	230	NSM					N					
xx72	◌	UNIFIED TAI VOWEL TAI DAENG UUE	MN	230	NSM					N					
xx73	◌	UNIFIED TAI VOWEL TAI DAENG U	MN	220	NSM					N					
xx74	◌	UNIFIED TAI VOWEL TAI DAENG UU	MN	220	NSM					N					
xx75	◌	UNIFIED TAI VOWEL TAI DAENG EE	MN	230	NSM					N					
xx76	◌	UNIFIED TAI VOWEL TAI DAENG SHORT O	MN	230	NSM					N					
xx77	◌	UNIFIED TAI VOWEL TAI DAENG UUA	Lo	0	L					N					
xx78	◌	UNIFIED TAI VOWEL TAI DAENG SHORT UH	Lo	0	L					N					
xx79	◌	UNIFIED TAI VOWEL TAI DON A	MN	230	NSM					N					
xx7A	◌	UNIFIED TAI VOWEL TAI DON AT	Lo	0	L					N					
xx7B	◌	UNIFIED TAI VOWEL LOW TONE AA	MN	220	NSM					N					

Sort Order—Lao Based

The following description is an initial attempt to define a sort order based on Lao. It is adapted from the orders used by Baccam, et. al. (1989) for Tai Dam and Marcus (1970) for Lao. The primary difference between this description and the order used by Baccam is the addition of the aspirated stops from Tai Dón, the 'g' and 'r' characters, and the vowel length contrast from Tai Daeng. We will look to Marcus for guidance on how to make those adjustments.

Consideration of Word and Syllable Structure

The best information is available for Tai Dam. The information given here for Tai Dam is thought to be representative of the other languages, unless explicitly noted.

There are two syllable patterns in Tai Dam: CV and CVC. When sorting, the segments of the syllable are considered in their spoken order, not their written order. Thus, when comparing ꪁꪫꪖ 'moon, month', to another word, first compare the ꪫ (/b/) to the initial consonant of the other word. Next, compare ꪁ (/ɨə/) to the vowel of the other word. Third, compare ꪖ (/n/) to the final consonant of the other word. Compare the tones last of all.

In Thai Song, the syllable can have a very limited range of initial consonant clusters. It is not clear at this time how those clusters should be sorted.

Tai Dam is almost exclusively monosyllabic. A very small number of words have an unstressed initial syllable. The first uses a mid-central vowel even though it is written with an 'ꪱ' (/a:/). E.g. ꪣꪱꪣꪱ 'even if', ꪣꪱ 'eye'. For a set of words with any given initial consonant, those with two syllables sort before those with only one. Effectively, the unstressed vowel of the initial syllable is considered to precede all other vowels.

Consonant order

In general, the consonants in Thai languages are sorted according to the point of articulation, starting at the back of the mouth and moving to the front. A few residue characters are often tacked on at the end. This rule leads to the order shown for the Unified Alphabet in the code chart, from xx00 to xx2F. The symbols ꪎ (/kon^4/) and ꪚ (/nɨŋ^5/) are sorted as though the words were spelled out. The symbol ꪾ (Repetition) has no sort order value.

Two considerations arise as to the sort order for the traditional forms of the script. First, those consonants that are added between xx33 and xx54 for the traditional writing have the same sort order as the ones from the Unified Alphabet that they correspond to. E.g. UNIFIED TAI LETTER ALTERNATE SO HIGH (xx35, ꪵ) has the same sorting value as UNIFIED TAI LETTER SO HIGH (xx0E, ꪗ).

Second, if the sort order is always according to the point of articulation, then the order becomes language dependent. E.g. in Tai Dam, UNIFIED TAI LETTER PO LOW has the orthographic value /p/ low. Thus the sort order for it remains unchanged from the default of the Unified Alphabet. But in Tai Dón, UNIFIED TAI LETTER PO LOW has the orthographic value /m/ low. Therefore it would sort after the UNIFIED TAI LETTER TAI DON MO HIGH.

Labialized Consonants and Consonant Clusters

In Baccam (1989), words with a labialized consonant were sorted after all words with the corresponding unlabialized consonant. It may be best to handle the Thai Song consonant clusters in a similar fashion.

Vowel order

The order shown in the character chart is an approximation of the vowel order, but leaves out many digraph vowels. A more complete order is shown in this chart. The vowel + final consonant ligatures are treated as vowels for sorting. As with the consonants, the orthographic value assigned to the characters affects the sort order.

This chart shows the Unified Alphabet, Tai Dam, and Tai Daeng. More study is needed for Tai Dón and Thai Song.

Unified Alphabet (spacing vowels)	Tai Dam traditional form (combining vowels)	Tai Daeng traditional form (long & short vowels)	IPA representation
◌ǎ	◌ (closed syllables)	◌ᩮ	/a/
◌ᩣ̂			/ɐ/ (Vietnamese)
◌า	◌า	◌า	/a:/
◌ᩥ	◌ᩥ	◌ᩥ	/i/
		◌ᩦ	/i:/
◌ᩬ	◌ᩬ	◌ᩬ	/ɨ/
		◌ᩭ	/ɨ:/
◌ᩩ	◌ᩩ	◌ᩩ	/u/
		◌ᩪ	/u:/
◌ᩮ	◌ᩮ	◌ᩮᩴ	/e/
		◌ᩯ	/e:/
		◌ᩱᩴ	/ɛ/
◌ᩱ	◌ᩱ	◌ᩱ	/ɛ/ (/ɛ:/ in Tai Daeng)
		◌ᩮᩰ	/oʔ/
		◌ᩰ	/o/
◌ᩣ	◌ᩣ	◌ᩣ	/o/ (/o:/ in Tai Daeng)
		◌ᩮᩫ	/ɔ/
	◌ᩮᩢ		/ɔʔ/
	◌ (open syllable)		/ɔ#/
◌ᩮ	◌ᩮ	◌ᩮ	/ɔ/ (/ɔ:/ in Tai Daeng)
◌ᩤ	◌ᩤ	◌ᩤ	/ə/

		◌ᦷ	/əː/
		◌ᦺ	/iə/
◌᦮	◌ᦹ	◌ᦹ	/iə/ (/iəː/ in Tai Daeng)
◌ᦵ	◌ᦵ	◌ᦵ	/ɨə/
		◌ᦵ	/ɨəː/
◌ᦴ	◌ᦴ	◌ᦶ	/uə/
		◌ᦶ	/uəː/
◌ᦷ	◌ᦷ		/əw/
◌ᦸ	◌ᦸ	◌ᦸ	/aj/
◌ᦵᦷ	◌ᦵᦷ	◌ᦵᦷ	/aw/
	◌ᦻ		/an/
	◌ᦼ	◌ᦼ	/am/
	◌ᦼᦽ		/ap/

Line Breaking

This is an initial draft of the line breaking rules for the Unified Tai 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.

1. A line break can always occur before or after the characters:
   - UNIFIED TAI SYMBOL KON
   - UNIFIED TAI SYMBOL NEUNG
   - UNIFIED TAI SYMBOL SAM.
2. A break can always occur before a vowel which is written in front of the initial consonant. These vowels include:
   - UNIFIED TAI VOWEL SPACING E
   - UNIFIED TAI VOWEL EH
   - UNIFIED TAI VOWEL O
   - UNIFIED TAI VOWEL UH
   - UNIFIED TAI VOWEL UEA
   - UNIFIED TAI VOWEL UHW
   - UNIFIED TAI VOWEL AY
   - UNIFIED TAI VOWEL TAI DAENG SHORT UH
3. A break can always occur after a Vowel + Final Consonant ligature which is written after the initial consonant. These ligatures include:
   - UNIFIED TAI VOWEL AN
   - UNIFIED TAI VOWEL AM
   - UNIFIED TAI VOWEL TAI DON AT
   - UNIFIED TAI VOWEL LOW TONE AA (occurs only in open syllables)
4. a) A break can occur before a consonant providing:

(1) The break will not split a labialized velar consonant.

That is, if the consonant is a UNIFIED TAI LETTER VO LOW or UNIFIED TAI LETTER TAI DAENG VO LOW, it must not be preceded by a velar consonant:

- UNIFIED TAI LETTER KO HIGH
- UNIFIED TAI LETTER KO LOW
- UNIFIED TAI LETTER KHO HIGH
- UNIFIED TAI LETTER KHO LOW
- UNIFIED TAI LETTER KHHO HIGH
- UNIFIED TAI LETTER KHHO LOW
- UNIFIED TAI LETTER NGO HIGH
- UNIFIED TAI LETTER NGO LOW
- UNIFIED TAI LETTER ALTERNATE KO LOW
- UNIFIED TAI LETTER TAI DAENG KO LOW
- UNIFIED TAI LETTER TAI DAENG KO ALTERNATE
- UNIFIED TAI LETTER TAI DAENG KHO HIGH
- UNIFIED TAI LETTER TAI DAENG NGO HIGH
- UNIFIED TAI LETTER TAI DAENG NGO LOW
- UNIFIED TAI LETTER TAI DON KO LOW
- UNIFIED TAI LETTER TAI DON NGO HIGH
- UNIFIED TAI LETTER THAI SONG KHO HIGH

(2) None of the vowels listed in rule 2 occur before it.

b) and one of the following vowels or tones occurs after it:

- UNIFIED TAI VOWEL COMBINING A
- UNIFIED TAI VOWEL SPACING A
- UNIFIED TAI VOWEL AA
- UNIFIED TAI VOWEL RAISED A
- UNIFIED TAI VOWEL COMBINING I
- UNIFIED TAI VOWEL SPACING I
- UNIFIED TAI VOWEL COMBINING UE
- UNIFIED TAI VOWEL SPACING UE
- UNIFIED TAI VOWEL COMBINING U
- UNIFIED TAI VOWEL SPACING U
- UNIFIED TAI VOWEL COMBINING IA
- UNIFIED TAI VOWEL SPACING IA
- UNIFIED TAI VOWEL UA
- UNIFIED TAI VOWEL AN
- UNIFIED TAI VOWEL AM
- UNIFIED TAI TONE COMBINING MAI EK
- UNIFIED TAI TONE SPACING MAI EK
- UNIFIED TAI TONE COMBINING MAI THO
- UNIFIED TAI TONE SPACING MAI THO
- UNIFIED TAI VOWEL TAI DAENG A
- UNIFIED TAI VOWEL TAI DAENG II
- UNIFIED TAI VOWEL TAI DAENG UE

- UNIFIED TAI VOWEL TAI DAENG UUE
- UNIFIED TAI VOWEL TAI DAENG U
- UNIFIED TAI VOWEL TAI DAENG UU
- UNIFIED TAI VOWEL TAI DAENG EE
- UNIFIED TAI VOWEL TAI DAENG SHORT O
- UNIFIED TAI VOWEL TAI DAENG UUA
- UNIFIED TAI VOWEL TAI DON A
- UNIFIED TAI VOWEL TAI DON AT
- UNIFIED TAI VOWEL LOW TONE AA

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

Script Samples

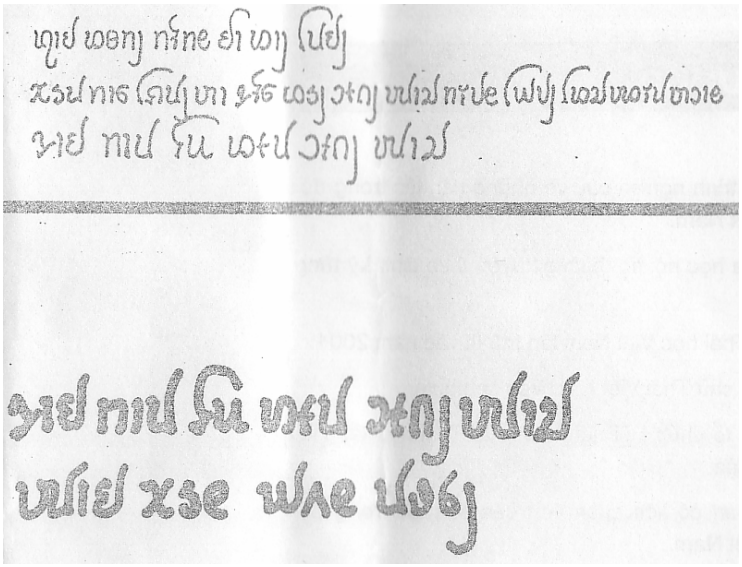

Figure 1—From *Giới Thiệu Chương Trình Thái Học Việt Nam*, 1999. Note the inter-word spacing.

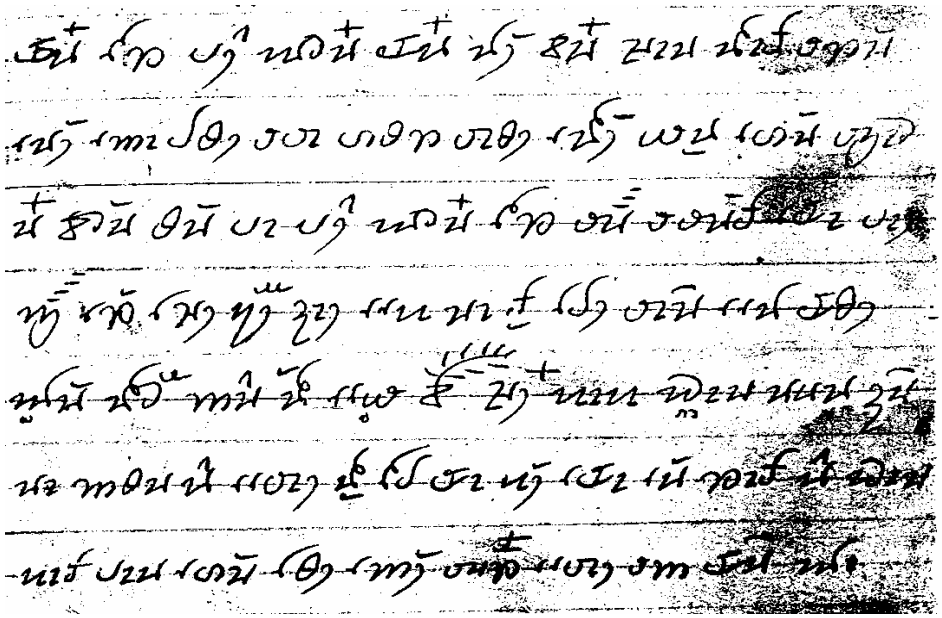

Figure 2—Untitled, undated manuscript in Tai Daeng. Note the word spacing and center baseline.

[illegible]	ກະຈົບ	ca-chốp	hoe
[illegible]	ກະຈອບ	ca-chóp	hoe
[illegible]	ກະດ້າງ	ca-đảng	stiffened, frozen
[illegible]	ກະເດັບ	ca-đếp	basket with strap
[illegible]	ກະຕາ	ca-ta	even if
[illegible]	ກະຕາກ	ca-tác	to cackle
[illegible]	ກະແບມ	ca-bem	coffer
[illegible]	ກັງ	căng	ape
[illegible]	ກັດ	cắt	to gnaw

Figure 3—From Baccam et. al., p 13.

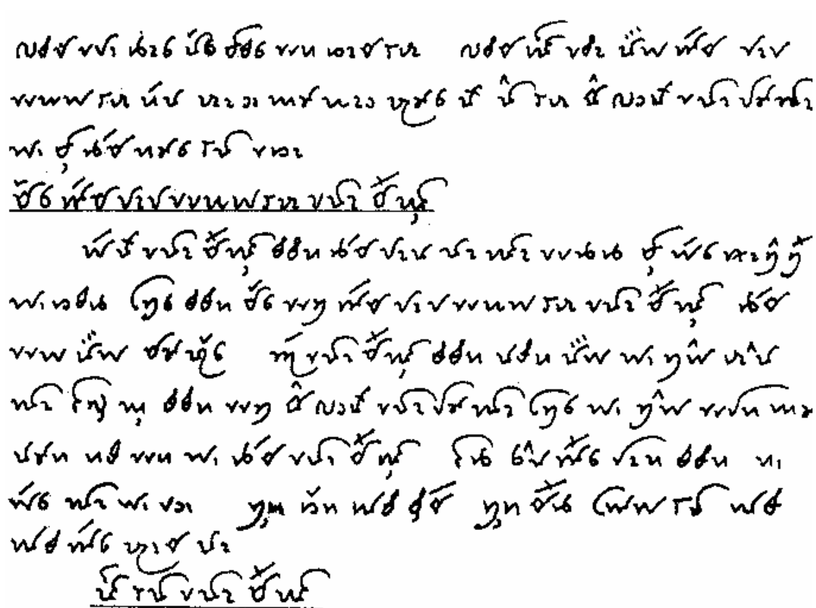

Figure 4—From *Khhăm Kháo Đi Chảu Dê-su Seo Lũng Ók Mác Têm*, 1983.