L2/05-166

Towards a Computerization of the Lao Tham System of Writing

Grégory KOURILSKY
gregory@kourilsky.com
(1)
(1) INALCO
2, rue de Lille
75343 Paris Cedex 7, France
http://www.inalco.fr

Vincent BERMENT
Vincent.Berment@imag.fr
(1) (2)
(2) GETA-CLIPS (IMAG)
BP 53
38041 Grenoble Cedex 9, France
http://www-clips.imag.fr/geta/

ABSTRACT — Although the subject of under-resourced *languages* is a problem often taken into consideration, one omits to make the distinction with under-resourced *scripts*. The Tham script of Laos, used to write Buddhist Texts, is one of the two official writing systems used in Lao P.D.R. (Laos). But if the other one — the so-called "laic" Lao script — is now quite well computerized (numerous fonts, input software, word processors, Unicode area), the Tham script seems to have always been forsaken by modern technologies (typewriters and computers). And this phenomenon does not seem to be being reversed soon since the Unicode Standard does not integrate it in any zone. Understanding the sociological and technical reasons of this neglect, we present an approach to mend it.

RÉSUMÉ — Bien que le sujet des *langues* peu dotées informatiquement soit un problème régulièrement abordé, on omet souvent de distinguer le cas des *écritures* peu dotées. L'écriture tham du Laos, employée pour noter les textes bouddhiques, est l'une des deux écritures offielles de la République Démocratique Populaire Lao (Laos). Si l'écriture lao proprement dite (appelée parfois par opposition *écriture lao "laïque"*) est aujourd'hui relativement bien dotée informatiquement (nombreuses polices de caractères, logiciels de traitement de texte, zone Unicode), l'écriture tham semble avoir toujours été délaissée des techniques modernes de saisie (mécanique aussi bien qu'informatique). Ce phénomène ne semble pas devoir s'inverser puisque le consortium Unicode ne l'a pas intégrée dans son standard. Tout en évoquant les raisons sociologiques et linguistiques de ce délaissement, nous proposons des éléments pour y remédier.

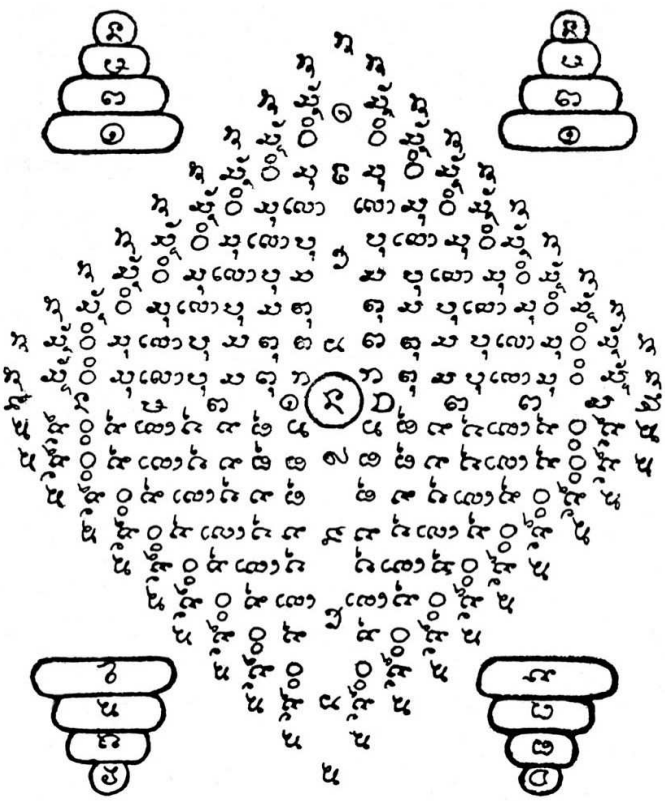
KEYWORDS — Tham script, ThamWord, Under-resourced scripts, Unicode.

1. An Under-Ressourced Script

According to the 10th Royal Order from 1949 that stills a reference although political changes, the Tham Script of Laos is the one of the two official writing systems used in Lao P.D.R. (Laos). The first one is the so-called "laic" Lao script used by population for administration and everyday life. The second one, Tham Script, is used to write Buddhist Texts, even if there are some examples of non-buddhist lapidary inscriptions like the famous stele of Dan Sai.
Contrary to the "laic" Lao Script which is now quite well computerized (numerous fonts, text input software, Unicode area), the Tham script seems to have been forsaken by modern technologies (typewriters and computers). We can find linguistical and sociological reasons to explain this phenomenon.

The linguistical reason is that the Tham Script is a complex script, never really codified, that allows many variant forms. Furthermore, it has the particularity to be used to write two different languages, Pali (language of Theravada Buddhism) and Lao, with different rules and specific characters for each of these languages (this aspect will be described further on).

The sociological reasons have several aspects. First of all, the Tham script is not well known among the Lao population: Principally high level Monks and few Scholars can read and write Tham. Secondly, the religious function of the Tham script makes it more than a simple language transcription system: as in Thailand, Cambodia and Myanmar, Pali letters in Laos are for an older time employed with a graphic consideration as disposition of the words, combinations, choice of graphic forms, that gives "an extra literary dimension to the reader consciousness" (BIZOT 1992:18). This aspect is well illustrated in the *yantra,* diagrams for protection made by grouping letters disposed in a non-linear way[1]. This utilization of these dispositions of letters is perceptible not only in texts but on inscriptions on diverse supports as sheet of metal or wood, woven and even tatoos on bodies[2]:



Lao *gāthā* [3] organized in a *yantra* with Tham letters[4]

It means that possibilities offered by text typing systems (mechanical typewriters and software) were a long time unsufficient for the use of the Tham Script in Laos. Another explanation is, last but not least, Tham has begun to be obsolete since the influence of Thai Buddhism reform[5] came deeper into Laos in the first decades of the twentieth century. Lao scholars like Maha Sila Viravongs who studied in pagodas in Thailand (especially a few years in Bangkok), were attached to Pali canonical texts from Ceylan and refractory to non orthodox practical as the use of letters evoked above. That ended in the Lao alphabet reform by the Buddhist Institute in adding missing

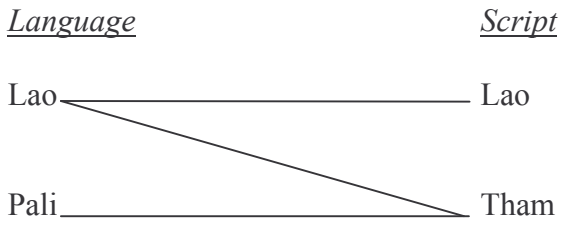
[1] See BIZOT 1996, 2001
[2] See BECCHETTI 1991.
[3] *gāthā* means "Stanza's heart", and can be considerated as magical formula (GABAUDE, 1988:246, note 6).
[4] This *yantra* illustrates the cover of (PHONE PHRA NAO MONASTERY 1965).
[5] King Mongkut (Rama IV), initiated in the first half of the 19th century a movement of reforms in Thai Buddhism that ended to the creation of a new *nikaya* called *dhammayut* (t. ธรรมยุตติกนิกาย) (GABAUDE 1988:31, note 73).

characters to write Pali in the Lao "laic" alphabet[6], so the Tham was not useful anymore considering Pali transcription. In fact the magical function of Tham continued to be used in the open country but quite badly considered by official Buddhism in the cities, place of technology and computer development. But today it seems that we can observe a resurgence of Tham Script even in the Capital, and old magical uses begin to get out their hiding-place[7].

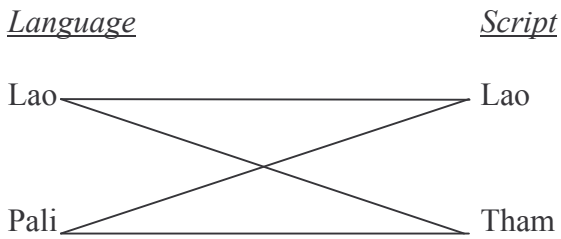
This phenomenon makes us think that the computerization of the Lao Tham system of writing is today a necessity more especially as modern technologies and computer softwares allow to render non-linear text and diagrams mentionned above. Far from analyzing the whole difficulties of computerization of the Tham Script system of writing, some important aspects will be reviewed.

2. Two "Sub-Writing Systems"

As we wrote, the Tham script has the particularity to transcribe two languages, Pali and Lao. Traditionally, relation between language and script in Laos can be illustrated in this schematic way:



Let's notice that during the existence of the Buddhist Institute (1933-1975), the Pali complementary letters added[8] in Lao "laic" alphabet made the relation completely symmetrical:



We can see that Tham is able (and employed) to write two languages, Pali and Lao, which are very different with regard to origin and phonology. Moreover, Phay Luang Maha Sena made a distinction in publishing two volumes of his manual to learn Tham Script: one volume concerns Tham to transcribe Lao, the second one teaches Tham transcribing Pali[9].

The first form of Tham Script was probably created towards the 15th century by the Mon in Lanna to transcribe all Pali letters included those that does not exist in Thai languages (FERLUS 1995), like

---

[6] In the 1930's, scholars from the Buddhist Institute introduced fourteen letters in the Lao alphabet to complete the Pali consonantic system (see VIRAVONGS 1937): ຫ *gh*, ຈ *ch*, ຊ *jh*, ຍ *ñ*, ຍ *ṭ*, ຖ *ṭh*, ຫ *ḍ*, ຝ *ḍh*, ຕ *ṇ*, ຣ *dh*, ກ *bh*, ຣ *ḷ*, and *niggahita* ◌ं *ṁ*. The two sanskrit letters ຣ *ś* and ຫ *ṣ*, unknown in Pali, were added too. All these letters falled completely into disuse after 1975 with a new reform lead by Phoumi Vongvichit.

[7] We can observe the same resurgence in Northern Thailand where Yuon (Lanna) script recently profited of new interest (but not necessarily for the same reasons).

[8] Cf. note 6.

[9] See SENA 1957, 1963.

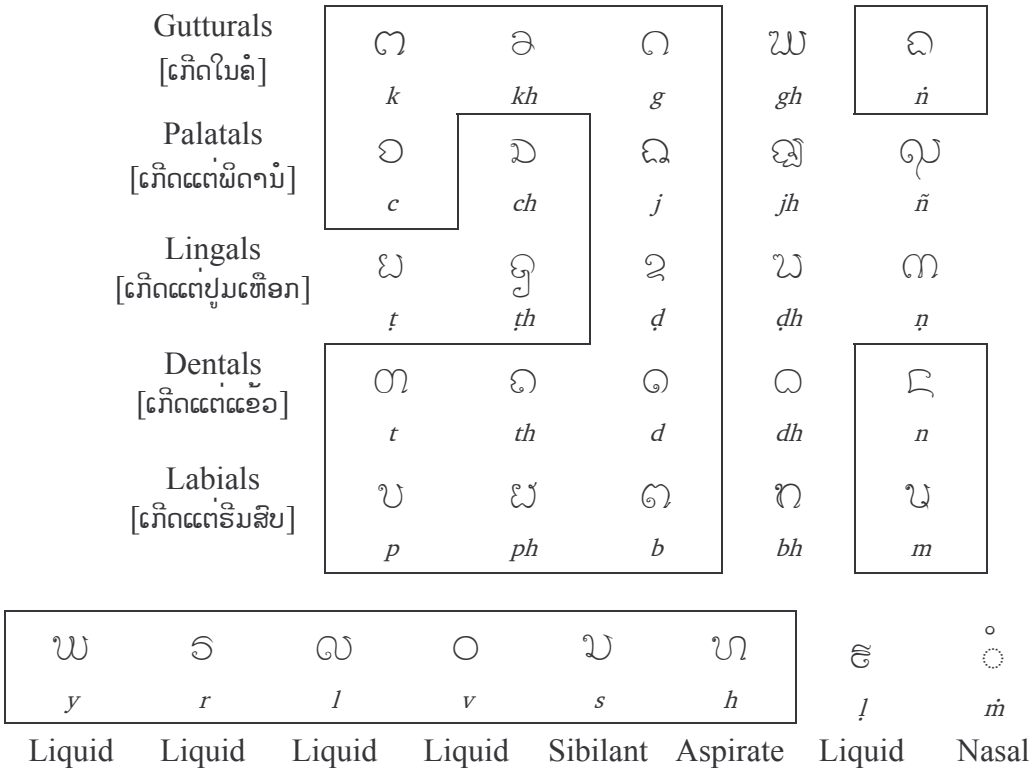
sounded or/and aspirate occlusives (*g*, *gh*, *j*, *jh*, *bh*, etc.), retroflexes (*ṭ*, *ṭh*, *ḍ*, *ḍh*, *ṇ*) and the *niggahita* sign *ṁ* (l. *ນິກຄະຫິດ* /nikk^{h}ahit/). Thus the Tham Script is related to Mon script and belongs to the vast family of Indic scripts derivated from Brahmi[10]. This writing system declinated later in the four scripts Yuon (ancient Lanna, actually part of the Thailand), Khün (region of Cheng Tung in Myanmar), Lü (Sipsongpanna Province in China) and Lao Tham (Laos). Each of those scripts was adapted according to the necessities of each language. That means that beside the Pali characters, populations added specific characters to transcribe their own languages. That is the reason why the assertion that the Tham Script is used to write two languages needs precisions because of differences between Tham transcribing Pali and Tham transcribing Lao.

The first difference is in the nature of the alphabet. The Tham transcribing Pali is a syllabic system (consonants have an inherent vowel) while Tham transcribing Lao is an alphabetic system. Thus the sign ຕ *k*[11] is read /ka/ in a Pali text, but read /k/ in a Lao text (to read /ka/, the sign ະ a must be added as well: ຕະ).

Secondly, some Tham characters are only employed to note Pali, some are specifically employed to write Lao, and some characters are used in both languages. If we just consider the Tham consonants traditionally used to transcribe Pali, only the characters with black border are common to Pali and Lao[12], the other ones are exclusive to Pali:

Gutturals [ເກີດໃນຄໍ]	ຕ *k*	ຂ *kh*	ດ *g*	ຝ *gh*	ຌ *ṅ*
Palatals [ເກີດແຕ່ພິດາບໍ]	ວ *c*	ສ *ch*	ຊ *j*	ຫ *jh*	ຍ *ñ*
Lingals [ເກີດແຕ່ປູມເຫືອກ]	ຍ *ṭ*	ຢ *ṭh*	ຊ *ḍ*	ຟ *ḍh*	ຕ *ṇ*
Dentals [ເກີດແຕ່ແຂ້ວ]	ຫ *t*	ຄ *th*	ຈ *d*	ດ *dh*	ຂ *n*
Labials [ເກີດແຕ່ຮີມສົບ]	ບ *p*	ຜ *ph*	ຕ *b*	ຫ *bh*	ຖ *m*

ຝ *y*	ຣ *r*	ລ *l*	ວ *v*	ມ *s*	ທ *h*	ຣ *ḷ*	ໍ *ṁ*
Liquid	Liquid	Liquid	Liquid	Sibilant	Aspirate	Liquid	Nasal



[10] Brahmi (p. *brāhmī*) script was known in the Empire of Maurya (India, IVth century B.C.). A tradition attributes its creation to god Brahma (PINAULT 2001:99).

[11] Pali and Sanskrit words are transliterated according to official transcription system established on 10th Congress of Orientalists (1894). Transcription of words, syllables or letters from other Asian languages is in IPA.

[12] Let's notice that as Tham letters are used in an etymological way, some characters have not the same phonetic value if writing Pali or Lao. For example ຊ has the value *j* in Pali but is prononced /s/ when transcribing Lao.

At the opposite, the following consonants have been added to write Lao and are not used in Pali:

ບ	ຝ	ຄ	ຢ	ຮ	ຫ
/b/	/fˇ/[13]	/f/	/j/	/h/	/ʔ/
ຫຼ	ຫຼ	ຫຼ	ຫຼ	ຫຼ	ຫຼ
/ŋˇ/	/nˇ/	/mˇ/	/ɲˇ/	/lˇ/	/vˇ/

We could make the same distinction for the vowels and other particular signs like tone marks only used in Lao. In Pali, only 8 vowels are used (a, ā, i, ī, u, ū, e, o) but 29 in Lao (sometimes written in several manners), plus particular signs as ໄ /aː/, ໄ /lɛ/, ຮ /rɯː/, etc[14]. But Pali uses independant vowels signs (ຫ,ຫາ,ຫ,ຫ,ຈ,ຈ,ຊ and ຫ) unknown in transcription of Lao.

Furthermore, rules and proprieties of writing differ when transcribing Pali or Lao. Thus it appears that a same word can be written in a different manner if written in a Pali text or a Lao text, for example a Lao word taken from Pali:

(1) ຫຄ້ຫາ, *taṇhā* , « desire » (Tham transcribing Pali),

(2) ຫ້ຼຫາ, [ຕັນຫາ], /tanhǎː/, « desire » (Tham transcribing Lao).

In a Pali transcription, a consonant following another consonant without inherent vowel is a subscript ((1) : ຫ *h*), but in a Lao transcription it is the final or medial consonant that can be (but not systematically) subscript ((2) : ຼ *n*).

Lastly, graphic forms of several subscript consonants differ if used in a Pali or in a Lao transcription:

value	Base Form	Pali subscript form	Lao subscript form
k	ຕ	ຕ	ຕ / ້
ṅ	ຂ	້	້ / ຂ
n	ຂ	ຼ	ຼ / ວ
l	ລ	ຼ	ຼ

For all these reasons, we've chosen to distinguish between two "sub-writing systems", each working in its own manner:

- P-Tham (Tham transcribing the Pali language),
- L-Tham (Tham transcribing the Lao language).

We precise that this terminology is made here for the purpose of computerization of the Tham Script and is not known in Laos.

---

[13] Lao high-class consonants are sounded, without tonemark, with a rising tone (INTHAMONE 1987:77) indicated with API sign ˇ placed on the right of the consonant. Sign ◌ represents any API vowel sign.

[14]For Tham Writing System (to write Pali and to write Lao), see (SENA 1957, 1963) (in Lao) and (KOURILSKY 2005) (in French).

3. Unicode and Input Methods

A simple way to mend the under-resourced problem of the Tham Script would be to create a Tham font which could be used on Windows system, the most popular operating system in the world. Considering the multi-utilization of several signs (for example the sign ᨠ is used to note Pali independant vowel /a/, Lao vowel /ɔː/ and Lao consonant /ʔ/; the sign ᨾ is employed to write vowel /eː/ and many non-connex vowel as ᨾᩣᩴ /ɔʔ/, ᨾᩣᩮ /ao/, ᨾᩣᩰᩴ /jaʔ/, etc.), about 132 elementary signs are needed to write all Tham letters. Thus we created a font called *ThamStandard*[15] that includes those 132 characters (plus a dozen of various forms). The problem is that much less than 132 characters are directly accessible on keyboards, so complex combinations are often necessary to type some letters (as *r* subscript ᩵ has to be typed with [Ctrl] + [4], [E] ; ligature *bb* ᩲ is typed with [CTRL] + [1], [o], etc.). Many Indic and derivated scripts fonts meet the same problem (Devanagari fonts, Khmer fonts[16], etc.) and sometimes several fonts are needed to type all the letters of those alphabets. One solution to face this problem is a normalization by an integration into the Unicode Standard[17]. The aim of the Unicode Standard is first to normalize every script in the world to permit a perfect compatibility between different fonts and platforms in exchanges of computerized texts (like Internet). In Unicode, each character has a single code and any "Unicode font" has to respect it. For example A LATIN CAPITAL LETTER A has the code U+0041, $ DOLLAR SIGN has the code U+0024, क DEVANAGARI LETTER KA has the code U+0915, ภ LAO LETTER KO has the code U+0E81, etc. But the useful point for our matter is a particularity of Unicode Standard that only *basic characters* are coded but not all *signs* (so-called *glyphs*). That means if क DEVANAGARI LETTER KA is coded U+0915, the various forms that can take this letter according to contexts (क, क्, क्, क्क, etc.) are not coded: it is the Windows Unicode Script Processor[18] (called *Uniscribe*[19]) jointly with an OpenType font that will take over to display the appropriate glyph.

The Tham script is not in Unicode yet but the next few examples will illustrate that the distinction between basic character and glyph is pertinent for this script:

Glyphs	*Basic Characters*
ᨠ ᩵ ᩰ	Tham letter *k*
ᨡ ᩰ ᩵	Tham letter *ṅ*
ᩴ ᨠ ᩵	Tham Lao vowel /ɔː/
ᨣ ᩱ	Tham vowel *ā*

[15] In free access on http://www.laosoftware.com.
[16] The Khmer font *Kdol*, created by ÉFEO (École Française d'Extrême-Orient) that includes numerous ligatures and various forms of letters regarding 17-19th centuries manuscripts, allows 438 signs (BIZOT 1992:19).
[17] For more information about the Unicode Standard, http://www.Unicode.org.
[18] Most of works concerning typing Unicode Asian fonts have been done for the Windows system. For more information about Unicode on Macintosh, see http://apple.com/macosx/features/international and http://www.alanworld.net/Unicode/fonts_macosx.html. For Linux, http://www.alanworld.net/Unicode/fonts_unix.html.
[19] See http://www.microsoft/typography/developers/uniscribe/default.htm.

By the way, the Unicode Standard and the associated input methods are able to solve the problem of the computerization of Indic and derivated scripts with numerous letters, because only a limited number of characters is required on the keyboard whereas the Windows Unicode Script Processor displays the adequate glyph according to the context (letter in a final position, ligature, etc.).
The matter to distinguish characters from glyphs is in practice not an easy task. The scripts implemented in Unicode solve their problems in different ways, not always conformable with general Unicode recommendations. A few cases taken among Indic and derivated Asian scripts (Devanagari, Tamil, Thai, Lao, Khmer, Burmese, Tibetan, etc.) will render illustrious the various choices available for an input method of the Tham script.

3.1 Consonants

Many Southeast Asian scripts use *subjoined*[20] forms for consonants. According to Unicode recommendations, subjoined forms have not to be coded and are supposed to be rendered by the system.

In Devanagari script, many various forms and ligatures occur, especially when a consonant lost its inherent vowel. The input method to render Unicode Devanagari introduces the special character 094D ◌् DEVANAGARI SIGN VIRAMA that points out that the preceding consonant has no inherent vowel. This character, according to the context, indicates to the system which form should be displayed for this preceding consonant:

0915 क + 094D ◌् + 092F य → क्य (क *k* displays as क्; the sequence is read /kja/),

0915 क + 094D ◌् + 0024 क → क्क (क्क is a ligature; the sequence is read /kka/),

In some contexts the *virama* is explicite and the sequence occurs as क्, etc.

In Khmer script, the presence of a subscript consonant in a syllable also indicates that the preceding consonant lost its inherent vowel. The difference with the Devanagari script is that the second consonant of the cluster will change its shape (and not the first one like in Devanagari) in becoming a subscript. In the word ក្សា /ksaː/, the subjoined position ◌្ of the consonant ស *s* indicates that we have to read /ksaː/ and not /kɑsaː/ (the consonant ក lost its inherent vowel /ɑː/). To inform rendering system that the consonant has to be rendered subjoined, an Unicode Khmer font has a character named *coeng*[21] (17D2 ◌្ KHMER SIGN COENG) that has to be typed before the consonant that has to be rendered subjoined:

1780 ក + 17D2 ◌្ + 179F ស (◌្ស) + 17B6 ◌ា → ក្សា /ksaː/

The Unicode Standard specify that character *coeng* does not exist in the Khmer alphabet and "does not have a conventional visual form[22] in Khmer as it is a control character to cause the formation of a subscript" (MICROSOFT CORPORATION 2004a:2).

[20] Terminology employed in Unicode's publication to call non-basic form of a consonant. For exemple Khmer subscripts consonants are the *subjoined* forms of the basic consonantic characters.
[21] That means "foot" in Khmer, and in a wider sense designs subscript consonants in the Khmer alphabet.
[22] The sign ◌្ which is the representation of the character U+17D2 SIGNE KHMER COENG in the Unicode's publications is an arbitrary form employed only for the needs of writing.

Two particular cases are :

- The subjoined ◌្រ *r* placed before consonant but coded (and typed) after it in "logical" order[23]:

  1785 ច + 17D2 ◌្ + 179A រ → ច្រ /craː/

  ◌្រ

- The *Robat* sign ◌៓ *r*, today not pronounced, is at the origin phonetically placed before the consonant above which it is placed but must be typed after it (KHMEROS 2004:13) in graphical order:

  1796 ព + 178E ណ + 17CC ◌៓ → ពណ៓ /poa/, etym. *poarna.*

  The case of the *Robat* illustrates that even a script so respectful of the Unicode recommendations as Khmer script can depart from the logical order.

The Tibetan Unicode works in a particular manner in coding twice the consonants: one basic form and one subjoined form (so in total contradiction with the Unicode recommendations). Example:

0F42 ག TIBETAN LETTER GA

0F92 ◌ྒ TIBETAN SUBJOINED LETTER GA

The reason put forward are there is no *virama* in Tibetan[24] and the necessity to reduce needs for memory while the use of subscripts is more than frequent. Whithout discussing this choice, this example shows that the Unicode's rules are not univocal.

3.2 Vowels

In most of the Asian scripts deriving from the Brahmi one, numerous vowels are formed by several graphemic signs (*non-connex* vowels) as in the few following examples:

◌ொ	Tamil vowel /o/,
◌ൊ	Malayalam vowel /ɔ/,
◌ັຍ	Lao vowel /jaʔ/,
◌ะ	Thai vowel /oʔ/,
◌ើ	Khmer vowel /ɯəː/,
◌ི	Tibetan vowel /iː/,
◌ᩱ	Tham vowel /ɯaː/.

We can observe that two families of Unicode input methods could be distinguished: one "Indic" model and one "Thai-Lao" model.

The "Indic" model strictly follows the Unicode Standard's recommendations and codes vowels as phonems (or known as phonems in the descriptive rules of those languages). Examples:

[23] The *logical order* agrees to phonetic order and in the most of case to typing order (UNICODE 2003:17).
[24] The letter *rok-me*, Tibetan equivalent to Indic *virama*, is employed only for transcribing Sanskrit.

0BCA ◌ொ TAMIL VOWEL SIGN O

0D4C ◌ൌ MALAYALAM VOWEL SIGN AU

0BCA ◌ျ KHMER VOWEL SIGN YA

0F73 ◌ི TIBETAN VOWEL SIGN II

We precise that from a general point of view the encoding method corresponds to the way of local spelling. That means that a single key has to be typed (after have typed the consonant) to display those vowels:

1780 ក + 0BCA ◌ျ → ក្យ /kɯəː/

This "Indic" vowel treatment implies to repositionning glyphs to avoid cross-over:

178F ត + 17B6 ◌ា +1780 ក + 0BCA ◌ျ → តាក្យ /taːkɯəː/

In this example the sylleme ក្យ has to be moved on the right to not cross-over with ◌ា. The reason put forward for this "logical" order is to facilitate alphasorting and speech rendering (SOLÁ 2004:3) in always having a consonant coded first[25]. There are exceptions as Khmer non-connex vowels not coded as ◌ាំ, ◌ុំ, ◌ុះ, ◌ោះ and ◌ាះ, which have to be rendered with two characters (◌ា + ◌ំ, etc.)[26].

In the opposite manner, Thai and Lao Unicode proceed differently in coding vocalic signs which are not systematically phonems but often only part of them. Examples:

- 0E42 โ THAI CHARACTER SARA O,

  0E30 ะ THAI CHARACTER SARA A,

  But the Thai vowel โ◌ะ /oʔ/ is not coded by itself but needs the typing of the two characters 0E42 โ THAI CHARACTER SARA O and 0E42 ะ THAI CHARACTER SARA A before and after the consonant to be displayed:

  0E42 โ + 0E15 ต + 0E30 ะ → โตะ /toʔ/

- 0EC0 เ LAO VOWEL SIGN E,

  0EB1 ◌ั LAO VOWEL SIGN MAI KAN,

  0E8D ຍ LAO SEMIVOWEL SIGN NYO,

  The Lao vowel เ◌ັຍ /jaʔ/ is not coded and has to be displayed in typing the three characters 0EC0 เ, 0EB1 ◌ັ and 0E8D ຍ, the first one before the consonant and the two others after it:

  0EC0 เ + 0E9A ບ + 0EB1 ◌ັ + 0E8D ຍ → ເບັຍ /bjaʔ/

Furthermore, scripts encoded following the "Indic" method have their vowels encoded in "logical order", which means that the vowel character always follows the consonant character, even if the first is shaped before the consonant sign in the writing system[27]. For example, the Devanagari

[25] Let's notice that in Tham script computerized alphasorting will need in all manner an electronic dictionary because of numerous *contractions* [l. ᨾᩣᩉᩮ᩠᩵] /kʰamɲɔ̀/ like: ᩋᩤᩥ /anvàː/ contraction of ᩋᩢᩅ᩠ᩅ, ᨴᩤᩦ /diːlǐː/ contraction of ᨴᩤᩃᩤᩥ, etc.

[26] Nevertheless those vowels are directly accessible on the keyboard with one single key (KHMEROS 2004:10). That means that keyboard does not correspond to encoding and more again special treatments have to be done on keyboard.

[27] This order correspond roughly to typing order (ANDRIES 2002:68). See also note 22.

vowel sign ि *i* is placed before the consonant sign. If one wants to type the syllabe पि *pi* with a Devanagari Unicode font, one should type the consonant character प *p* before the vowel character:

092A प + 093F ि → पि *pi*

It is the Windows Unicode Script Processor that will proceed to the reorganization of the elements and should ensure corresponding between the logical order (characters) and the visual order (glyphs) (UNICODE 2003:260).

The "Thai-Lao" model goes differently in coding characters in graphic order. Thus prevowels signs are coded and typed before the consonant (the same way as handwriting) following a graphic rule. Examples:

0E41 แ + 0E19 น → แน /mɛː/,

0EC4 ไ + 0E81 ก → ไก /kaj/.

The Myanmar Unicode system is between the two families in encoding prevowels in logical order but in not coding non-connex vowels:

1000 က + 1031 ေ → ေက /keː/,

1000 က + 1031 ေ + 102C ော → ေကာ /kɔ̀/.

Thus the Myanmar consonant ေော /ɔ/ is considered as an equivalent sequence of the two characters 1031 ေ MYANMAR VOWEL SIGN E + 102C ော MYANMAR VOWEL SIGN A.

This "in-between" is quite strange but we can suppose a will of simplification according to Khmer Unicode in not encoding non-connex vowels.

In conclusion we observe that the implementation of Unicode concerning Indic and derivated scripts presents differences in the consonants treatment (between Devanagari, Khmer, Tibetan) as well as in vowel treatment (between Thai-Lao, Khmer, Myanmar). Nevertheless it appears that the "Thai-Lao" input method is the most simple for implementation and utilization, especially for vowels. But because Thai nor Lao scripts use subjoined consonant, we have to choose a system inspired from another script that goes for Tham.

4. Proposal of an input method for Tham Script

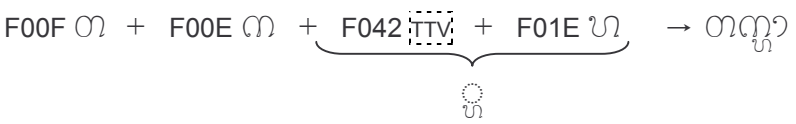
4.1 P-Tham consonants

The P-Tham has a common system with Khmer in the use of subscripts: a subscript consonant indicates that the preceding consonant lost its inherent vowel. In the word ᨲᩢᩉᩣ *taṇhā*, the subjoined form ᩢ of ᨻ *h* indicates that the consonant ᨶ *ṇ* has to be read without inherent vowel, so /n/ and not /na/[28]. Thus a system similar as Khmer Unicode can be applied in adding a rendering character which we could call "Tham theoric virama" (F042 TTV) that will indicate to the system that the following character has to be rendered subjoined. This character finds here its linguistic justification in "killing" the inherent vowel of the preceding consonant and goes as a *virama*. As this character does not exist in Tham alphabet, we call it "theoric".

[28] The phonetic value of *ṇ* in a final position of the syllable is /n/ (instead of /ŋ/).

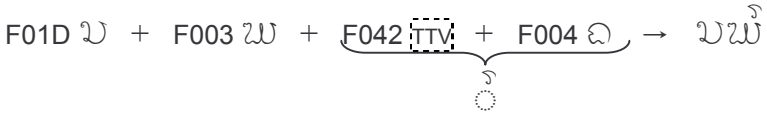
Example:

F00F ຫ + F00E ຕ + $\underbrace{\text{F042 TTV + F01E ທ}}_{\text{ຫ້ຫ}}$ → ຫຕ້ຫ



Two particular cases are:

- ້ , subjoined form of ຂ *ṅ* occurs above the consonant and is a final (i.e. without inherent vowel). The rendering method is the same but has no linguistic justification because the preceding consonant hasn't loose its inherent vowel. To type the word ນພ້ *saṅgha* (with ນ *s* and ພ *gh*):

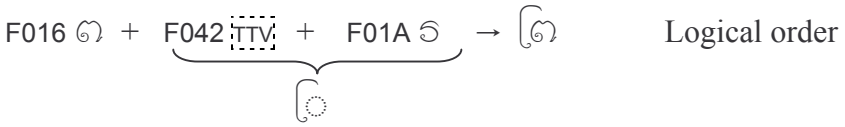


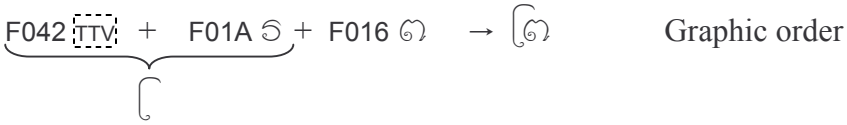
  F01D ນ + F003 ພ + $\underbrace{\text{F042 TTV + F004 ຂ}}_{\text{້ ຂ}}$ → ນພ້

- ເ subjoined form of ຣ *r*, occurs before the consonant. To type the word ເຕ *bra* we have to choose between "logical order" or "graphic order":

  F016 ຕ + $\underbrace{\text{F042 TTV + F01A ຣ}}_{\text{ເ ຂ}}$ → ເຕ Logical order

  $\underbrace{\text{F042 TTV + F01A ຣ}}_{\text{ເ}}$ + F016 ຕ → ເຕ Graphic order





  The implementation in logical order has a linguistic justification (lost of preceding consonant's inherent vowel) and is conform to Unicode recommendations (and to other Unicode Southeast Asian scripts). The graphic order has the advantage to not need any special treatment for reordering and may be easier for users. We think better to choose the graphic order because of the similarity with Tham prevowels (see *infra*, 4.4).

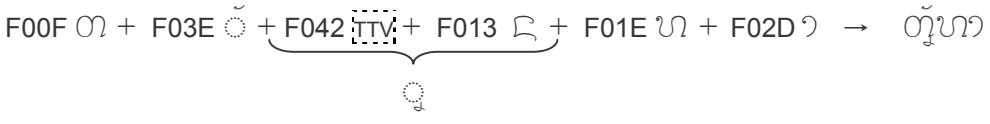
4.2 L-Tham consonants

Two aspects differentiate L-Tham subjoined consonants from P-Tham's.
The first one is the context where subjoined occur. In L-Tham, subscripts occur:

- In a final position of a syllabe: ຫ້ຫາ [ຕັນຫາ], /tanhǎː/.

  Note that subscript is not systematic in a final position but depends on consonant and vowel: in fact it is graphical criterions that will determine to use or not a subjoined. For example in the word ທ້ຢ້ຕ [ໝົກ] /mŏk/ the final /k/ is written in its basic form because of an unaesthetic appearence of the word if written with a subscript as ທ້ຢ້ຕ.

  This first aspect doesn't rule out the subjoined rendering system used in p-Tham:



  F00F ຫ + F03E ັ + $\underbrace{\text{F042 TTV + F013 ຂ}}_{\text{ຊ}}$ + F01E ທ + F02D າ → ຫ້ຫາ

Here the P-Tham linguistic justification (lost of preceding consonant's inherent vowel) disappears because the character F042 ᩵ follows a vowel (in all manner there is no inherent vowel in Lao). But the most important is to be able to type L-Tham and users will not take offense for linguistical rules in computer typing.

- In a medial position of a syllabe: ᩃ᩠ᩅ [ลอ] /lwaː/

  One more time, the P-Tham rendering system can be applied:

  F01B ᩃ + F042 ᩵ + F01C ᩋ + F02D ᩅ → ᩃ᩠ᩅ

The second aspect is differences between several subjoined forms used in P-Tham and L-Tham as we evoked in chapter 1.

For example if we take the Lao word ᨹ᩠ᩃᩢ [ผลัน] /phlan/, a problem appears because the glyph ᩃ᩠ needed for that Lao word is not the glyph ᩤ used in P-Tham. Thus if we use the character F042 ᩵, we'll have:

F016 ᨿ + F042 ᩵ + F01B ᩃ + F03E ᩮ + F01A ᩢ → ᨹᩤᩢ

The same problem happens with glyphs ᩮ (subjoined ᨿ /k/) and ᩯ (subjoined ᩣ /ŋ/), that occur only in L-Tham.

Two solutions have been considered:

- Display glyphs automatically according to context.

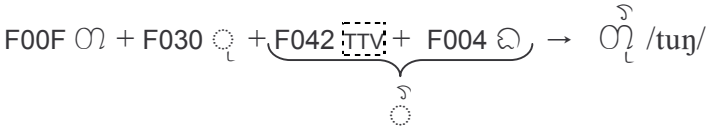
It is not possible in Tham because of words written in different manners like ᨴᩥᩫᩤ /diliː/ that can be written ᨴᩤᩫᩤ or ᨴᩥᩫᩥ, ᩁᩣᨠ /nǎːk/ that can be written ᩉᩣᩴ or ᩉᩣᩰ, etc. The orthography is often let to the choice of the copyist. The use of contractions (l. ᨾᩣᩁᩢ /kʰamɲǎ/) is frequent but also not obligatory. For example Lao word ᩍᩣᩁ /ao/ can be written ᩍᩣ᩠ᩅ or ᩉᩮᩣ᩠ᩅ, word ᩋᩢᩁᩣ /anvàː/ can be written ᩋᩥ or ᩋᩢᩴᩅᩣ, etc.

- Display glyphs with a second subjoiner.

As character F042 ᩵ THAM THEORIC VIRAMA is employed for all P-Tham subjoined and all common L-Tham/ P-Tham subjoined, one complementary character could be used to display specific L-Tham subjoined forms, called F043 ᩶ THAM VARIANT SUBJOINER. Thus a user will be allowed to choose between two rendering characters depending on the glyph he wants to display.

+	F000 ᨿ	F004 ᩃ	F01B ᩃ	F01D ᩅ
F042 ᩵	ᩤ	ᩯ	ᩤ	ᩥ
F043 ᩶	ᩮ	ᩯ	ᩤ	ᩥ

Example with ᩋ /ŋ/ subjoined rendered ◌᩵ or ◌᩶:



4.3 Various forms for subjoined consonants

A rendering method with two subjoiner characters allows an advantage in respect of the non-codified aspect of the Tham Script. In manuscripts many various forms occur especially for the subjoined consonants. It is because the Tham Script stills archaic and was never really strictly codified by any national nor religious institution[29]. Sometimes three or four forms are known for one subscript consonant[30]. The non-canonic aspect evoked (*supra*, 1) gives importance to graphic forms for physical appearance of texts. For example subscript form of ᩈ *m* can occur as ◌᩠ᨾ, ◌᩠ᩃ or ◌᩠ᩁ, ᩋ *th* can occur as ◌᩠ᨳ or ◌᩠ᨵ, etc.

As the character THAM THEORIC VIRAMA will be used to render canonical Pali subjoined forms according to Maha Sena's manuals, the THAM VARIANT SUBJOINER will allow to display L-Tham particular subjoined forms *and* variants (P-Tham and L-Tham). In that manner every consonant has two possibilities for subjoined by using one or the other of those two characters. Of course a system able to render all variants is not possible here (the user who wants to dispose numerous variants should have recourse to several fonts), but having two variants in a single font offers enough possibilities to be quite conform with many texts.

Exemples:

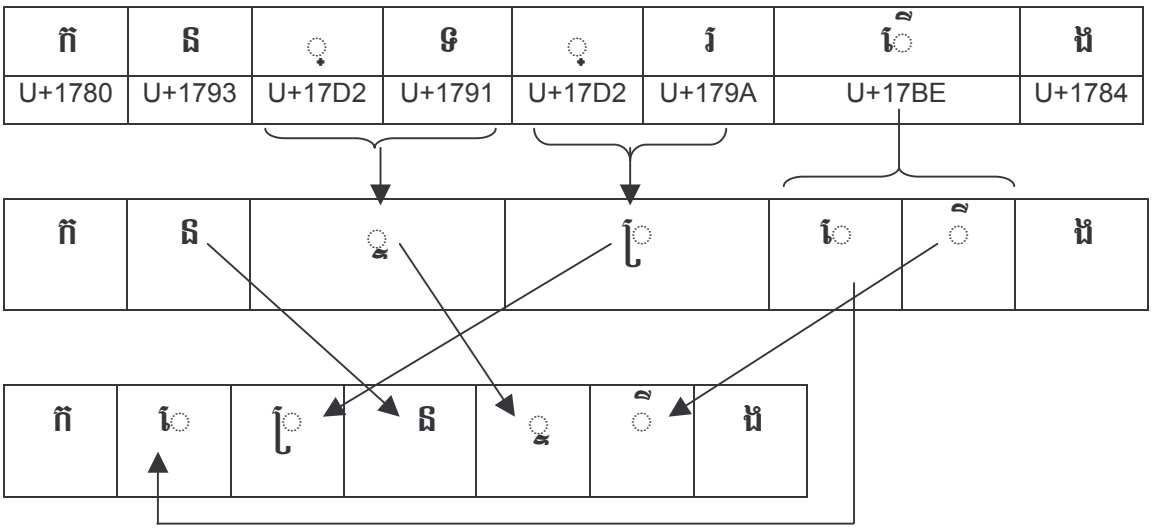
+	F00C ᨬ	F018 ᨸ	F012 ᨒ	F01C ᨜	F01C ᨝	F01C ᨞	F01C ᨡ
F042 TTV	◌᩵ᨬ	◌᩵ᨾ	◌᩵ᨒ	◌᩵᨜	◌᩵᨝	◌᩵ᨠ	◌᩵ᨡ
F043 TVS	◌᩶ᨬ	◌᩶ᨾ	◌᩶ᨒ	◌᩶᨜	◌᩶᨝	◌᩶ᨠ	◌᩶ᨡ

We precise that the fact that two forms of subjoined are avalaible does not go against Unicode recommendations because only base form consonants and the two subjoiners are coded while the particular glyphs will be included in the font (OpenType or other).

[29] The two volumes of Phay Luang Maha Sena respect graphic forms from a lapidary inscription dated from 1819 (GAGNEUX 1983:77). But those graphics were not really layed down as a canonical form by the Buddhist Institute.
[30] See (GABAUDE 1979:83-90). The author counted not less than three hundred graphic variations between two ten pages manuscripts.

4.4 Vowels (P-Tham and L-Tham)

We've seen that two ways of encoding are possible for Unicode, one "Indic" input method and one "Thai-Lao" input method. If the "Indic" method could be in a first approach tempting in a reason of a linguistic logic (a vowel is before all a sound without consideration of the way it is written), some problems appear very quickly when one try to implement with complex words, in particular those contain at one and the same time a subjoined consonant and a non-connex vowel. For example to render the word កន្ត្រើង /kɑːntrəːŋ/, one has to type (example from (BAUHANH 2002:15-16)):



Result: កន្ត្រើង /kɑːntrəːŋ/

We can see the complexity of the operations that have to be done by the system:

- Transformation of consonants រ and ត in their subjoined form ◌្រ[31] and ◌្ត,
- Reordering of the glyph ◌្រ,
- Split up of ◌ើ in ◌ and ◌ើ,
- Reordering of ◌ (just before ◌្រ).

We can observe the misleading operation that consists in spliting up ◌ើ in ◌ and ◌ើ, although Unicode insist to code the character ◌ើ separately (U+17BE). This ends to important difficulties to make a Khmer Unicode Font: After years and work of numerous computer scientists and linguists, only one set of Khmer Unicode font is now avalaible[32] (to our knowledge) and it is not yet sure it gives entire satisfaction. The point of view of the users does not convince much more: Unicode admits that the typing is not intuitive and an education has to be delivered to users (UNICODE 2004:277).

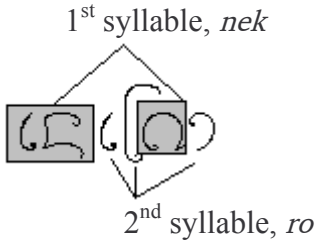
With the example of a word like កន្ត្រើង /kɑːntrəːŋ/, which could have equivalent in Tham, we are able to conclude that if we can retain the Khmer Unicode's rendering process for subjoined

[31] In this context the consonant រ subscript has to be rendered as ◌្រ, more elongated as normal glyph ◌្រ.

[32] The font KhmerOS for Windows is available on http://www.Khmeros.info/Khmeros_download.html. It seems that no Unicode font are avalaible for Macintosh yet.

consonants, the input method for the « Indic » model vowels does not convince. Neither does the "in-between" method of the Myanmar Unicode which does not seem to be more successfull than the Khmer Unicode regarding to local users [33].

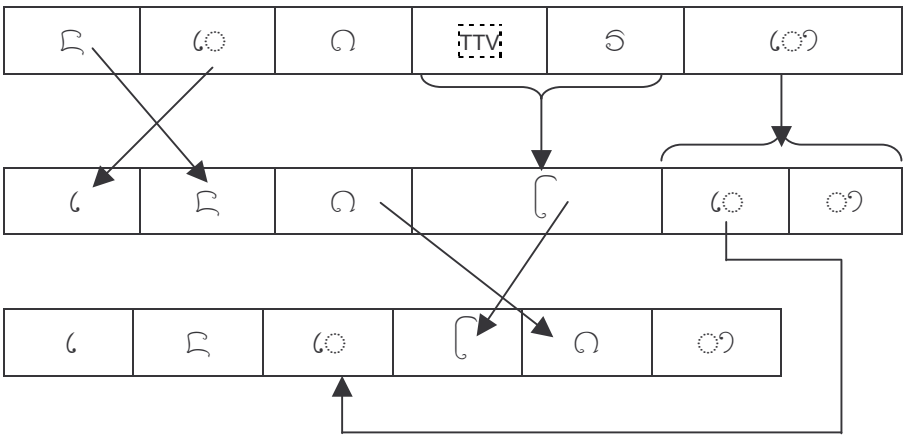
Let's take an example in Tham. The Pali disyllable ᨶᨦᨶ᩵ᨣᩮᩣ *nekro*, includs a pre-vowel (ᨦᩮ *e*), a subjoined consonant (᩵ᨣ *r*) and a non-connex vowel sign (ᨦᩮᩣ *o*). The ordering of the disposition of letters does not correspond to the ordering of the reading:



Representation with latin letters: *e n o r k*[34] *o*

We can see that the first vocalic graphem (ᨦ, here part of vowel ᨦᩮᩣ *o*) of the second syllabe is placed before the final consonant (ᨣ *g*) of the first syllabe. That means that visually speaking, the two syllabes cross-over (as in the latter Khmer exemple) and each does not form a monolithic group. To render the word in the Unicode manner, we have to determine how to display the good positionning for the non-connex vowel ᨦᩮᩣ *o* regarding to the subjoined consonant ᩵ᨣ *r* (for example not to write ᨶᨦ᩵(ᨦᩮᩣ)). For this reason, "Indic" method seems to be too hard to implement non-connex vowels as single characters.

In "logical" order, one must type:



Result: ᨶᨦ᩵(ᨦᩮᩣ)

[33] "Although character codes for Myanmar Languages has been allocated in UCS/Unicode (U+1000-U+109F), lack of implementation makes unavailable to local end users. Much effort had been made to develop Myanmar character codes and fonts by many international experts and local experts." (KO KO, YOSHIKI 2005:9).

[34] In a final position, ᨣ is sounded /k/.

The operations executed by the system are:

- Transformation of TTV + ᨿ in ᨿ,
- Permutation of ᨾᨿ in ᨾᨾ,
- Permutation of ᨿᨿ in ᨿᨿ,
- Split of ᨿᨾ in ᨿ and ᨾ,
- Positionning of the part ᨿ before ᨿ.

Besides of the complexity of implementation and font making which is an evidence[35], one can point out the non-intuitive typing method, in actual fact the user has to think how the word is sounded rather than it is written. Furthermore, users that do not know Tham Script or not well (students, novices, or even copyists for any publication) will not be able even to type Tham correcty without a headache.

Indeed some argue that users do not have to know how the encoding is done and typewriting could be simplified with other program operations (for example to *type* prevowels *before* the consonant). So let's imagine what this argument infer: first, implementation has to be done to allow system to reorganize (as seen above) characters in a "logical" order in displaying correct glyphs in a graphic order. Secondly, one will have to implement more treatments to bring in a more intuitive typing, and to finally come back at the starting-point:

1. Typing	ᨾᨾ	ᨿᨿ
2. Reorganization for correct coding	ᨾᨿ	ᨿᨿ
3. Reorganization for correct display	ᨾᨾ	ᨿᨿ

Could we not disclose little absurdity in this ?

Finaly, too much automatism in rendering impeeds freedom in orthographic variation. In Khmer Unicode for example, if one types a word in a non canonical way, the system will considerate that "an invalid combination is found; more than one vowel character in a syllabe, more than two subscripts on the same base character, or the incorrect ordering of subscripts, a new cluster will be formed that has the dotted circle [◌][36] as the base glyph" (MICROSOFT CORPORATION 2004b:2). That means if one wants to type, for whatever reason, ᨠᨶᨿᨿᨾ, which has no sense and is unreadable, it will not be possible and one will get something as ᨠᨶ◌◌ᨿᨿᨾ◌. Of course this is an extreme example because it is rare to want to type non-sense succession of letters. But what could happen is one types words in a manner unknown from canonical manners, as diagrams, combinations of letters or syllabes as evoked in the beginning of this article ? To avoid this, a space must be added[37] between each character, and Unicode recomands to add also characters U+200D ZERO WIDTH NON-JOINER "to explicitly suppress a ligature" (UNICODE 2004:282), U+200C ZERO WIDTH JOINER and U+200B ZERO WIDTH SPACE "which can be used for word boundaries" (MICROSOFT CORPORATION

[35] Especially as Tham script will certainly not benefit such human and financial resources as Khmer Unicode did.

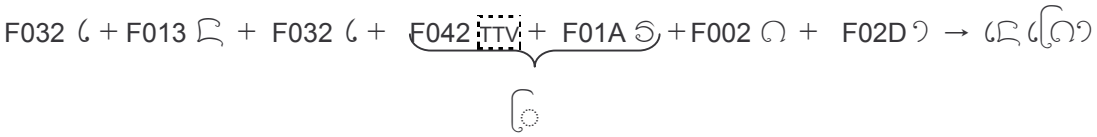
[36] At the condition that dotted circle is included in the font. If not, error sign □ will appear instead.

[37] To "render a sign standalone (in apparent isolation from any base) one should apply it on a space" (MICROSOFT CORPORATION 2004b:1).

2004b:2). Thus to use freely the whole Khmer character set, those three characters (that do not belong to Khmer alphabet) must be added in a Khmer Unicode font. This adds in the complexity on typing Khmer Unicode.

Thus we think that in respect to "Thai-Lao" input method, only vowel single signs should be coded and words like ᨠᩕᨠᩕᨣᨾ *nekro* should be typed in a graphic order:

F032 ᨠ + F013 ᩕ + F032 ᨠ + F042 TTV + F01A ᩚ + F002 ᨣ + F02D ᨾ → ᨠᩕᨠᩕᨣᨾ



ᩕᩦ

Beside the advantages due to the simplicity of implementation, we have to consider the point of view of national Lao users and their needs to dispose a single system for the two system of writing, Lao "laic" and Tham. It would be contradictory if an Unicode Tham font would not goes the same way as Lao "laic" font especially as conceptors and computer scientists are not so numerous and won't be so enthousiastic in having two different rules of implementation for each system of writing. Notice that Tham script is also used in Northeast Thailand (Region called "Isan") since it is an ancient region of Laos, and Thai who would make Unicode Tham fonts would certainly prefer to not have to change their habits regarding Thai Unicode.

To sum up, Input method for an Unicode Tham Script should goes in:

- Encoding vocalic unit signs, as graphem and not as phonem, according to the "Thai-Lao" method,
- Encoding signs in a graphic order (differently than the "Indic" logic order),
- Using two subjoiners: THAM THEORIC VIRAMA, based on a linguistical logic (lost of preceding consonant's inherent vowel) to render P-Tham subjoined, and THAM VARIANT SUBJOINER used to display variants and specific L-Tham subjoined consonants.

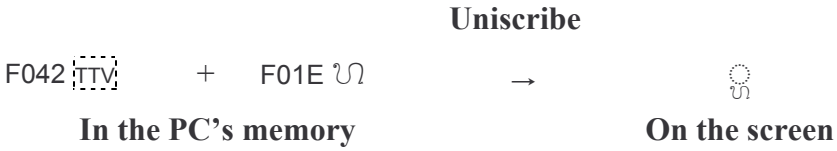
The advantages of those choices are:

- No special treatment for ordering is required, so as early as the Tham Script is included into the Unicode Standard, only a few time will be needed to end to a good Tham typing system (and not several years like for Khmer or Myanmar),
- Easiness for Unicode Tham font makers from Laos who will recognize the way to make Lao Unicode fonts,
- Easiness for Unicode Tham font makers from Thailand who will recognize the way to make Thai Unicode fonts,
- Simplicity for users in typing,
- Utilization possible for copyists who don't know Tham (or not well),
- Flexibility in utilization in allowing to type non-canonical forms, according to Tham handwriting tradition,
- Possibility with a single font to render two forms of subjoined consonants in a respectfully approach of multi-forms Tham letters like in the Lao palm-leaves manuscripts tradition.

5. ThamWord: An Interim Tool for Tham Unicode Input and Rendering

5.1 Uniscribe: an obstacle to the implementation of a Tham Unicode system

Though the "Thai-Lao" input method we propose is much simpler to implement than the "Indic" one, some processing will need to be applied to render the text properly due to some of the options we took (e.g. to have the same key for the normal and for the subjoined forms, only differing by the ◌ྭ prefix). We mentioned *Uniscribe*[38] several times and said that it has this role of transforming character chains into its desired graphical form:



Unfortunately, the Tham script is not handled by Unicode yet so ◌ྵ will not appear on the screen when we will type the two characters or have them in the PC's memory. In a word, Uniscribe is not working for Tham yet.

5.2 ThamWord: An interim tool to bypass the obstacle

Until the days where Uniscribe will handle Tham, we propose ThamWord, an interim tool that will simulate its behavior. To do that, we first have to design a font in the Unicode private area. We selected the range F000-F0FF as suggested in (BERMENT 2004:153) for this font that we called *ThamUnicode*. In addition to the code points described in (KOURILSKY 2005:174-194), this font includes other points for the subjoined characters. It actually covers all the 132 characters (plus variants) as does the *ThamStandard* font so all the Tham glyphs can be displayed through a code. The extra codes are taken from the code points of the range that remained unused.

Then, when typing for example F042 ◌ྭ and then F01E ྵ, the tool will replace the two codes by a third one which represents ◌ྵ. Note that this processing is different from what Uniscribe would do, as Uniscribe would only compute the graphical form associated to the two codes and ThamWord actually replaces the two codes themselves in the memory.

5.3 Virtual keyboards and technical solution

If the rendering of the text is probably the most complex issue in the absence of a Tham-capable Uniscribe, an interim tool has also to provide a Tham virtual keyboard. Several techniques have been described for that purpose in (BERMENT 2004). We chose to implement this function in a hook procedure that we derived from LaoWord, a Word add-in adapted to the Lao language that contains such a hook for its virtual keyboards (BERMENT 2003).

[38] The Unicode Script Processor embedded inside Windows.

The main features of ThamWord are:

- Providing an *ad-hoc* virtual keyboard for Tham,
- Transforming the newly typed characters and their environment according to the contextual rules described before.

The two features are included in the hook procedure which is launched by a Word add-in during its initialization. This add-in is a dynamic library (DLL) that is mapped in the memory space of Word when this application starts. It provides a human-machine interface through a toolbar added to the Word's toolbars.

The virtual keyboard is modifying the keycodes sent to Word following a table of correspondence defined both by the keyboard mapping and by the Tham Unicode zone[39]. At this level, the font used doesn't need to have any glyph but the ones defined by a code in the Tham Unicode zone.

After this modification of the code, the hook procedure analyses the environment around the insertion point. If a rule has to be applied for the entered code in the current context, another modification is done, possibly with a deletion of some characters around the new one as for F042 ᩅ + F01E ᩍ → ◌ᩍ where ᩅ is removed from the memory when ᩍ appears in the hook queue and is processed. In order to explore the environment around the insertion point, the hook procedures has to dialog with Word. This is done thanks to the Component Object Model (COM) interface.

5.4 Preparing the future

We can hope that the Tham writing system will soon be part of the Universal standard that Unicode aims to become and that the operating systems will be able to natively process complex Tham fonts (OpenType...). When this will happen? Who knows. At that time, the texts made in between with ThamWord or other equivalent software will constitute an existing corpus that will need to be slightly modified for fully complying the standard that was defined. Reverse rules like ◌ᩍ → F042 ᩅ + F01E ᩍ will have to be applied to go back to the original input. This mechanism will soon be added to ThamWord.

[39] Actually it is still in the private area, as the submission to Unicode of a Tham zone is ongoing. See (KOURILSKY 2005).

Sources

ANDRIES, P. (2002), *Introduction à Unicode et à l'ISO 10646,* Québec, Document numérique, Volume 6, n° 3-4, p. 51-88.

BAUHANH, M. (2002), "Rendering the World's Complex Scripts: A Case Study in Khmer", 21th International Unicode Conference, Dublin, 20 p.

BECCHETTI, C. (1991), *Le mystère dans les lettres, ปริศนาธรรมในอักษร*, Édition des Cahiers de France, Bangkok, 116 p.

BERMENT, V. (2003), *LaoWord's Word Processing Functions*, Pan Asia Networking All Partners 2003, March 3 - March 10, 2003, Vientiane Novotel, Vientiane, Laos, 10 p.

BERMENT, V. (2004), *Méthodes pour informatiser des langues et des groupes de langues « peu dotés »*, Grenoble, P.H.D. Thesis (Université Joseph Fourier), 277 p.

BIZOT, F. (1992), *Le Chemin de Laṅkā*, Paris, ÉFEO, Textes bouddhiques du Cambodge n° I, Paris, 352 p.

BIZOT, F. (1996), *ปฏิบต, La pureté par les mots*, Paris, Phnom Penh, Chiang Mai, Vientiane, ÉFEO, Textes bouddhiques du Laos, 275 p.

FERLUS, M. (1995), "Les circonstances de l'introduction de l'alphabet Tham lanna", *in*: *La Thaïlande des débuts de son histoire jusqu'au XVe siècle*, Premier Symposium Franco-Thai, 18-24 juillet 1988, Silpakorn University, p. 101-109.

GABAUDE, L. (1979), *Les cetiya de sable au Laos et en Thaïlande. Les textes*, Paris, ÉFEO, PÉFEO 118, 96 + 338 p.

GABAUDE, L. (1988), *Une herméneutique bouddhique contemporaine de Thaïlande :* buddhadasa bhikkhu, Paris, ÉFEO, PÉFEO CL, 690 p.

GAGNEUX, P.M. (1983), "Les écritures lao et leur évolution du XVe au XIXe siècles", ASEMI, vol. XIV, 1-2, p.75-95.

INTHAMONE, L. (1987), *Je lis et j'écris lao*, Paris, INALCO, 207 p.

KO KO, W. YOSHIKI, M. (2005), "Languages of Myanmar in Cyberspace", Dourdan (France), TALN Conference, June 2005, 10 p.

KOURILSKY G. (2005), *Éléments pour un traitement informatique de l'écriture tham du Laos*, Master Thesis (INALCO), 330 p.

PELTIER, A.R. (2000), *ມ໋ຍ໋ຫ, บางผึมหอม, La fille aux cheveux parfumés*, Vientiane, Institut de Recherches sur la Culture, 490 p.

PINAULT, G.J. (2001), "Les écritures de l'Inde continentale", in *Histoire de l'écriture, de l'idéogramme au multimédia* (dir. Anne-Marie Christin), Paris, Flammarion, p.93-121.

MAHA SENA, P. (1943), *ແບບ ธรม ไอ ธรม อ่าน หัังสืทัม ธรมเป้นพาสาลาอ [Learn fast to read Tham Characters in Lao Texts]*, Bangkok, Kramol Tirannasur, 80 p.

MAHA SENA, P. (1943), *ແบบ ธรม ไอ ธรม อ่าน หัังสืทัม ธรมเป้นพาสาบาลิ [Learn fast to read Tham Characters in Pali Texts]*, Bangkok, Kramol Tirannasur, 51 p.

SOLÁ, J. (2004), "Displaying Khmer", KhmerOS Initiative, http://www.khmeros.info/download/DisplayingKhmerScript.pdf, 4 p.

VIRAVONG, S. (MAHA SILA) (1937), *ບາລິເຍຍາກຣຕ ອັກຂຣະວິສີກາຄທີ່ນຶງ ສັມຄຸຄາກິສານແລະສນຸສິ, [Pali Grammar, Akkharavidhi, Part one, Samaññabhidāna - Sandhi]*, Vientiane, Institut Bouddhique, 68 p.

Collectives :

KHMEROS (2004), "Creating and supporting OpenType fonts for the Khmer script", http://www.khmeros/download/KhmerUnicodeTyping.pdf, 18 p.

MICROSOFT CORPORATION (2004a), "Creating and supporting OpenType fonts for the Khmer script", http://www.microsoft.com/typography/otfntdev/khmerot/default.htm.

MICROSOFT CORPORATION (2004b), "Other encoding issues: Khmer OpenType specification", http://www.microsoft.com/typography/otfntdev/khmerot/other.htm.

UNICODE (2003), Unicode 3.1 annoté, http://iquebec.hapax.com.

UNICODE (2004), The Unicode Standard 4.0, http://www.Unicode.org.

PHONE PHRA NAO MONASTERY [ວັດໂພນພຣະເນົາ] (1965), *ທັມມະຄາຖາ* [*dhammagāthā*], Vientiane, 88 p.