Towards a comprehensive proposal for Thai Noi / Lao Buhan script

Compiled by Ben Mitchell, with contributions from Patrick Chew, Theppitak Karoonbanyanan and Prapanth Iamwiriakul.

This is a response to the ‘request to add Thai characters to ISO/IEC 10646/Unicode document L2/18-041’ [pdf]. This present document is not intended to be authoritative, but merely prompt further analysis and discussion of the script with a view to establishing the best approach for encoding the script. Further consultation with user communities and academics is recommended.

Context

Further information is needed to explain how and where the script is used, and for which languages. This will help settle which Unicode block the script should be encoded in.

Users of the script

Thai Noi script is also known in Laos, where it is known as Lao Buhan¹. Though not in widespread current use, Lao Buhan is the proximate source of modern Lao script. Analysis and comparisons should be made between Thai Noi and Lao Buhan, to ensure consistency in encoding characters and a robust mechanism for their composing behaviour(s) that accommodates both orthographies and any language differences.

The number of users of the Thai Noi / Lao Buhan script should be estimated; a large majority of Standard Thai users do not know Thai Noi.

These two factors indicate that the user groups for Thai Noi / Lao Buhan, Standard Lao and Standard Thai are quite different.

Languages and scripts

Hand in hand with the user groups, the proposal should state which languages the script is used for, and how the script may be used differently for each (for example, Thai Isaan language and Lao language written in the script may have different requirements). It should also be discussed the extent to which standard Thai language could be written in Thai Noi / Lao Buhan script. Use of the Thai Noi / Lao Buhan script for Pali-Sanskrit and any minority languages should also be documented, as different characters may be needed, they may combine in different ways, or have different sound values.

¹ Choice of which name to call the script conveys a cultural bias. The Thai name is ไทยน้อย /tʰâːj nɔːi/ meaning ‘lesser Thai’. The Lao name ແຫ້ວຽງ [pʰouèəŋ] /lāːw būːhâːn/ or Thai equivalent ลาวโบราณ means ‘ancient Lao’; in this document neither name is preferred, and the script is called by both. Buhan can alternatively be spelt ‘buhane’ or ‘bouhane’. Occasionally the term ‘Lao Buhan’ is also applied to the Tham Lao / Isaan script.
Equally, the relationships of Thai Noi / Lao Buhan to Lao, Khmer, Myanmar and Tham scripts should be explored to ensure a) difficulties that were encountered in encoding those scripts are learned from and b) consistency of encoding where appropriate.

To ensure the widest possible applicability, historical sub-genres of Thai Noi / Lao Buhan script should be identified, such as Lao Chiang (ลาวเฉียง). Study of the Nithet (อักษรนิเทศ) and Fakkham (อักษรฝักขาม) scripts may also reveal instructive overlaps. (As a parallel example, four main sub-genres are discernible for Tham script: Tham Lanna, Tham Isaan/Lao, Tham Lue and Tham Khuen, and their differing behaviours should all be accommodated by the Tham encoding). All this is to say: given the transnational and historic significance of the script, like with Tham script, efforts to encode Thai Noi / Lao Buhan should not be based around a single user-group, and need to be a non-partisan, language-agnostic international collaboration.

**Not an extension to the Thai alphabet**

Thai Noi / Lao Buhan letters are not part of the Thai alphabet, nor are they an extension to it. The two scripts do not interact: there are no words that contain letters of Thai and letters of Thai Noi / Lao Buhan, and they are used for different languages. This means Thai Noi / Lao Buhan is unlike the Khmu example in the Lao block, where the Khmu letters extend the Lao alphabet to produce sounds not found in Standard Lao, and they would be used as part of the Lao alphabet. This is not to say that Thai and Thai Noi / Lao Buhan could not occur side by side: a dictionary of Isaan language using Thai Noi / Lao Buhan script may very well need the two scripts typeset together.

Instead, Thai Noi / Lao Buhan seems to require a different character set from Thai. Indeed, the proposal lists many characters that are found already in the Lao Unicode block, and some that are glyph variants of Thai characters. The cleanest way to avoid this muddle would be by encoding separately. If encoding in the Thai or Lao blocks, it should be explained why these characters should be encoded again.

Two further factors that suggest separate encoding are that the behaviour of the script is different from Thai, which does not have any mechanism for stacking consonants, ligatures and other consonant forms as described below; and that the proposal does not use Thai letter names: instead of NO NU or MO MA, it calls the letters NA and MA.

---

2 As in footnote 1, care should be taken not to inadvertently reinforce cultural bias by allocating Thai Noi / Lao Buhan letters to the existing Thai or Lao Unicode blocks; or by using a partisan pronunciation scheme or nomenclature.

3 Though differing morpho-syntactic behaviour alone is not enough to warrant separate encoding, revising the current Thai (or Lao) implementation to accommodate stacking would mean the full established encoding schema would need to be completely re-architected, impacting the current active user base of an extremely digital language/writing system.
Character repertoire

Every character that is used in Thai Noi / Lao Buhan should be evidenced with attestations and descriptions explaining function, behaviour and any variant forms found. It is essential to have an exhaustive list. See Appendix B for a proposed format.

Because Thai Noi / Lao Buhan is a historical script, using modern pronunciation values is fraught with inconsistencies between the varieties of spoken languages. As such, the inventory should name consonants according to their original Brahmic varga position using proto-Taiic sound values: a suggested scheme is presented in the appendices.

For vowels it should specify where the signs sit in relation to base consonants (above, below, left, right), and which signs can be used together. The numerals in the character names (e.g. LA2, KHA1) should be explained; or a different naming system should be adopted (see appendix A). Care should be taken to talk about glyphs and characters distinctly: some are base consonants, some are so-called sub-consonants, and some are semivowels, ligatures or conjuncts. The following sections aim to list the proposed characters by categorising them this way, but much more study is needed to be certain.

Theppitak Karoonbanyanan has done a lot of work to document the characters: see https://linux.thai.net/~thep/esaan-scripts/tn-issues/tn-encoding.html for a more comprehensive list with attestations. A full proposal will need to include numerals, punctuation and other symbols along with consonants and vowels.

Consonants

It is unclear why \texttt{OE63} THA, \texttt{OE67} LA3, and \texttt{OE68} SA6 would be required, if this is intended as an extension to the Thai block, since these appear to be mere typographic/glyph variants of \texttt{OE16} THO THUNG, \texttt{OE25} LO LING and \texttt{OE2A} SO SUA. In Thai Noi / Lao Buhan, these letters, and some others, have different forms than in standard Thai or Lao (often \texttt{ລ} and \texttt{ສ} instead of \texttt{ລ} and \texttt{ສ}); this is another factor that suggests separate encoding.

Proposed \texttt{OE6C} O+YA (.ORDER) is a character that can be found historically in some writing systems of northern Thailand/. While in modern Thai it would be considered a ligature or digraph, this character is not a ligature in Tham or Lao, where it is encoded at 1A40 and 0EA2 respectively, and should not be named as a ligature in the Thai Noi / Lao Buhan encoding.

\footnote{The reader will note the lack of attestations in this document; this is due to a lack of time to locate and scour the required resources.}
\footnote{Reflexes in Thai, Lao and other languages can then be annotated in the character descriptions.}
\footnote{Other differences generally include variants of \texttt{ປຝຟ} in the forms \texttt{ປຝຟ}; and the numerals \texttt{໒໔໕໖} which can be \texttt{໒໔໕} and \texttt{໖}. Appendix A contrasts each consonant with its counterpart in sibling scripts.}
\footnote{In modern Thai, \texttt{อย} as a consonant cluster represents a mid-class YA, where the \texttt{ย} effectively converts the \texttt{ย} from a low tone to a mid tone consonant. Historically it represented a glottalised palatal continuant, and in a standard Thai font could be potentially accessed with a discretionary <hlig> digraph glyph.}
Conjunct mechanism

Thai Noi / Lao Buhan, like many other Southeast Asian and Indic scripts, uses various consonant forms (above-base, below-base and post-base glyph alternates) of consonants to indicate conjuncts. Whether these forms are also used to indicate medial consonants\(^8\) (as in Burmese) and/or final consonants\(^9\) (as in Tham and some orthographies of Khmer) should be documented.

In Thai Noi / Lao Buhan, many consonants, for example ອ, have both below-base (◌ໝ) and post-base (◌ໜ) forms\(^10\) for use in different contexts\(^11\). It will be necessary to distinguish between these, and establish a mechanism to select the appropriate glyph. Unusually, in Thai Noi / Lao Buhan, consonant forms may also employed in polysyllabic words to indicate word-internal syllable onsets (such as ສາ/phāː/ = language).

Conjunct forms are not found in Thai, and in other Indic scripts are generally not encoded, but accessed by contextual substitution. For example, Khmer and Myanmar encodings operate on a virama model, which could be suitable, as this allows any arbitrary sequence of consonants to be clustered using a control character. However, Tibetan provides an alternative model, in which conjunct forms are explicitly encoded. This would more easily allow consonants to have more than one alternate glyph, as well as determining their above-base, below-base or post-base status\(^12\). One of these two possible approaches should be chosen\(^13\).

Conjunct forms

Proposed 0E3C LA2 would be cognate with 0EBC LAO SEMIVOWEL SIGN LO, and is a combining mark that attaches below the baseline (◌ໝ). As mentioned above, it should be explained whether this represents the second part of a consonant cluster, as in Lao, or a medial or final consonant. Note also the post-base form of LO (◌ໜ) is not mentioned.

Proposed 0E3D YA2 is a spacing, post-base form of the consonant Ya. As such, it would be cognate with 0EBD LAO SEMIVOWEL SIGN NYO (◌ຽ), or with the 103B MYANMAR CONSONANT SIGN MEDIAL YA (◌ျ). Whether it is used as a vowel/semivowel, a medial or a

---

8 Medial consonants in Burmese inflect a syllable-initial consonant or subsequent vowel, without initiating a new syllable.
9 Syllable-final (coda) consonants in Tham or Khmer may follow an inherent or explicit vowel; this necessitates an implementation in which vowel glyphs can carry consonant forms. This has been a difficulty with the Universal Shaping Engine (USE), which does not currently allow consonants to operate on vowels.
10 Suggested terminology: ตัวเฟื้องหอย take different forms below the baseline while ตัวซ้อน take the same form as the base consonant, but sit below the baseline. ตัวเฟื้องข้าง are the post-base forms. Proper terms should be discussed and agreed.
11 For example ໝ vs ຘໝ.
12 Transferring user choice of positioning into the input method is likely to be easier than getting locale-specific or time-period-specific font positioning rules to operate without issue. It also makes text exchange fully transparent: if positioning is left to the font, users with different fonts or different versions of fonts may not see the same displayed text.
13 Tham script mixes both approaches, which is also possible, though not so streamlined.
Towards a Thai Noi / Lao Buhan proposal 28 February 2018

14. An underlying difficulty with the Taiic languages is their propensity to reuse glyphs in several categories (as consonants, vowels, medial or final consonants).
post-form, depending on the font style.

A post-base form of MA (лежа) is found in proposed 0E61 KHA1+MA (กก) and 0E6B SA3+MA (กก); these should follow the same encoding principle. Below-base forms of NA (ำ/ฦ) and MA (ฦ) borrowed from Tham script also exist in Thai Noi / Lao Buhan, and should be discussed.

Other ligatures are possible: one source about the Fakkham script attests a combining version of  souha nya with the form (אם), to make ligatures such as (າซอ) and (aho). Although perhaps rather rare, an encoding should aim to provide maximum flexibility for transcribing different kinds of texts.

Vowel signs

0E3B MAI KONG is cognate with 0EBB LAO MAI KON(G), an above-base combining mark of the form (◌๑). Its sound value, and valid combinations should be described.

0E6E SARA A2 appears to be a glyph variant16 of 0EBO LAO VOWEL SIGN A (◌๑) or 0E30 THAI CHARACTER SARA A (◌่), rather than a separate character. (If it is a separate character from the normal sara A, it would need to be shown contrastively).

Proposed 0E6D O+YA (◌๑) is not a glyph variant (allograph) of 0E6C O+YA (ๅ) mentioned above, and should be contrasted with it. In fact, although Thai represents both characters homographically with the letter pair (อย), here it represents the vowel -oy- and is the same character as 1A6D VOWEL SIGN OY in Tham script. In Thai Noi / Lao Buhan the sign is called YO YAT NAM (หยอยยตั้น).

Other conjunct forms

As shown in Appendix B, we can also find:

- subjoined CA (ฃ)
- subjoined DDA (ฅ)
- subjoined DXA (ฅ or ฅ)
- subjoined TA (ฅ)
- subjoined DHA (ฅ or ฅ)
- subjoined NA (ฅ or ฅ)
- subjoined BXA (ฅ)
- subjoined BA (ฅ)

15 Because of the way computers process text, using actual characters encoded in the Tham block as the subjoined components in Thai Noi / Lao Buhan is not possible: this would break script runs meaning Tham marks could not attach to Thai Noi / Lao Buhan bases. It would also mean that fonts supporting Thai Noi / Lao Buhan would also have to support (a subset of) Tham characters.

16 Various forms of Sara A are found in Thai Noi / Lao Buhan, such as ◌ฮ and ◌ฮ. These, however, are not contrastive — they are merely stylistic and do not serve different functions from ◌ฮ.

17 The letter pair อย in Thai can act as a consonant cluster or a diphthong.
The presence of all these conjunct forms, and potentially other undocumented ones, suggests a virama model might be most appropriate\textsuperscript{18}. The Myanmar model would be a useful point of reference here.

\textsuperscript{18} Following the Tibetan model would require each subjoined form to be attested and given a separate codepoint; because of potential undocumented subjoined forms, following the Myanmar virama model may be better.
Further reading

English language


Grabowsky, V (2011) “Manuscript culture of the Tai” in Manuscript Cultures 4 pp145–156, Centre for the study of manuscript cultures, Hamburg University [pdf retrieved 21 February 2018]


Ronnakiat, N (1992) “Evidence of the Thai Noi alphabet found in inscriptions” from the third international symposium on language and linguistics, Bangkok, Thailand [pdf retrieved 21 February 2018]


Wikipedia Isan language

Lao language


พระภิรมย์เทียบ (1951) ดรุณ์อักษรปั้นสิทธิ์ มรรจ.บุชชวาระ [Learn to read Tham Lao, in Lao] vol. 2, Institut Bouddhique, Vientiane [pdf retrieved 21 February 2018]

Thai language


บุญเจิม, ส (2009) ตําราเรียนอักษรธรรมอีสำน [Learn Tham Isaan script], มรดกอีสำน [Isaan heritage publisher] สุบาระราชณี [Ubon Ratchathani], Thailand.


ผดุงสุนทรารักษ์, วิโรจน์ (2004) อักษรไทยและอักษรขอมไทย [Thai and Khom Thai alphabets], Ramkhamhaeng University, Bangkok.

ไพบูลย์หวังเจริญ, พิมพ์พรรณ (2013) วิเคราะห์สารนิเทศจากใบจุ้ม [Analysing the Bai Jum manuscripts], สํานักหอสมุดแห่งชาติ กรมศิลปกร National Library Fine Arts Department.


วรรณอุดร, สมัย (2004?) อักษรไทยน้อย [pdf retrieved 21 February 2018]

—— (2005?) ความเป็นมาของอักษรไทยน้อย [pdf retrieved 21 February 2018]

ศิริวัฒน์, วัฒนา (undated) ซุดความรู้ ๔ อักษรไทยน้อย [pdf retrieved 21 February 2018]

—— (undated) ซุดความรู้ ๔ อักษรไทยน้อยกับภาษาลาว [pdf retrieved 21 February 2018]

## Consonants by varga, compared to sibling scripts

<table>
<thead>
<tr>
<th></th>
<th>Glottalised</th>
<th>Unvoiced unaspirated</th>
<th>Unvoiced aspirated</th>
<th>Unvoiced other</th>
<th>Voiced unaspirated</th>
<th>Voiced other</th>
<th>Voiced aspirated</th>
<th>Nasal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Velar</strong></td>
<td>KA</td>
<td>KHA</td>
<td>KXA</td>
<td>GA</td>
<td>GXA</td>
<td>GHA</td>
<td>NGA</td>
<td></td>
</tr>
<tr>
<td><strong>Palatal</strong></td>
<td>CA</td>
<td>CHA*</td>
<td>JA</td>
<td>ZA</td>
<td>JHA*</td>
<td>NYA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Retroflex</strong></td>
<td>QDDA*</td>
<td>TTA</td>
<td>TTHA</td>
<td>DDA*</td>
<td>DDHA*</td>
<td>NNA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Dental</strong></td>
<td>QDA</td>
<td>TA</td>
<td>THA</td>
<td>DA</td>
<td>DHA</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Labial</strong></td>
<td>QBA</td>
<td>PA</td>
<td>PHA</td>
<td>FA</td>
<td>BA</td>
<td>VA</td>
<td>BHA</td>
<td>MA</td>
</tr>
<tr>
<td><strong>Non-varga</strong></td>
<td>YA</td>
<td>QYA</td>
<td>RA</td>
<td>VOCALIC R</td>
<td>VOCALIC RR</td>
<td>LA</td>
<td>VOCALIC L</td>
<td>VOCALIC LL</td>
</tr>
<tr>
<td></td>
<td>WA</td>
<td>SHA</td>
<td>SSA</td>
<td>SA</td>
<td>HA</td>
<td>LLA</td>
<td>QA</td>
<td>HHA</td>
</tr>
</tbody>
</table>

- Top left of each cell shows Thai, top right is Lao (including Pali letters), bottom left is Myanmar and bottom right is Tham Isaan/Lao.
- In the letter names, X represents friction (reconstructions vary by scholar, e.g. KXA = *xa/*kʰa etc and GXA = *ɣ/*gʰ etc) and Q represents glottality in the proto-Thai sound values.
- Thai Noi / Lao Buahan glyphs are indicative only — several forms are attested for some characters. See Appendix B for allographs.
- * Absence of Thai Noi glyphs should not be construed as evidence these letters do not exist. It is suggested codepoints are reserved for empty slots in case / until attestations are found.
- † Many different glyphs are found to represent vocalic consonants in Thai Noi, Tham Lao and Lao Pali; there are also inconsistencies in which glyphs are used for which letters.
- ‡ In different sources,  dna is transcribed as /pʰ/ and sometimes as /l/1a/. More research required.
### Appendix B

**Consonant list with allographs and suggested names**

<table>
<thead>
<tr>
<th>Base form*</th>
<th>Allographs</th>
<th>Suggested name</th>
<th>Conjunct forms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ວ</td>
<td>ວ</td>
<td>THAI NOI / LAO BUHAN LETTER KA</td>
<td></td>
</tr>
<tr>
<td>ໄ</td>
<td>ໄ</td>
<td>THAI NOI / LAO BUHAN LETTER KHA</td>
<td></td>
</tr>
<tr>
<td>໅</td>
<td>໅</td>
<td>THAI NOI / LAO BUHAN LETTER KXA</td>
<td></td>
</tr>
<tr>
<td>ໆ</td>
<td>ໆ</td>
<td>THAI NOI / LAO BUHAN LETTER GA</td>
<td></td>
</tr>
<tr>
<td>໇</td>
<td>໇</td>
<td>THAI NOI / LAO BUHAN LETTER GXA</td>
<td></td>
</tr>
<tr>
<td>່</td>
<td>່</td>
<td>THAI NOI / LAO BUHAN LETTER GHA</td>
<td></td>
</tr>
<tr>
<td>້</td>
<td>້</td>
<td>THAI NOI / LAO BUHAN LETTER NGA</td>
<td></td>
</tr>
<tr>
<td>໊</td>
<td>໊</td>
<td>THAI NOI / LAO BUHAN LETTER CA</td>
<td></td>
</tr>
<tr>
<td>໋</td>
<td>໋</td>
<td>THAI NOI / LAO BUHAN LETTER CHA</td>
<td></td>
</tr>
<tr>
<td>໌</td>
<td>໌</td>
<td>THAI NOI / LAO BUHAN LETTER JA</td>
<td></td>
</tr>
<tr>
<td>ໍ</td>
<td>ໍ</td>
<td>THAI NOI / LAO BUHAN LETTER ZA</td>
<td></td>
</tr>
<tr>
<td>໎</td>
<td>໎</td>
<td>THAI NOI / LAO BUHAN LETTER NYA</td>
<td></td>
</tr>
<tr>
<td>໏</td>
<td>໏</td>
<td>THAI NOI / LAO BUHAN LETTER QDDA</td>
<td>ທ</td>
</tr>
<tr>
<td>໐</td>
<td>໐</td>
<td>THAI NOI / LAO BUHAN LETTER TTA</td>
<td></td>
</tr>
<tr>
<td>໑</td>
<td>໑</td>
<td>THAI NOI / LAO BUHAN LETTER TTHA</td>
<td>ຟ</td>
</tr>
<tr>
<td>໒</td>
<td>໒</td>
<td>THAI NOI / LAO BUHAN LETTER JHA</td>
<td></td>
</tr>
<tr>
<td>໓</td>
<td>໓</td>
<td>THAI NOI / LAO BUHAN LETTER NNNA</td>
<td></td>
</tr>
<tr>
<td>໔</td>
<td>໔</td>
<td>THAI NOI / LAO BUHAN LETTER QDA</td>
<td>ຟ</td>
</tr>
<tr>
<td>໕</td>
<td>໕</td>
<td>THAI NOI / LAO BUHAN LETTER TA</td>
<td></td>
</tr>
<tr>
<td>໖</td>
<td>໖</td>
<td>THAI NOI / LAO BUHAN LETTER THA</td>
<td></td>
</tr>
<tr>
<td>໗</td>
<td>໗</td>
<td>THAI NOI / LAO BUHAN LETTER DA</td>
<td></td>
</tr>
<tr>
<td>໘</td>
<td>໘</td>
<td>THAI NOI / LAO BUHAN LETTER DHA</td>
<td></td>
</tr>
<tr>
<td>໙</td>
<td>໙</td>
<td>THAI NOI / LAO BUHAN LETTER NA</td>
<td></td>
</tr>
<tr>
<td>໚</td>
<td>໚</td>
<td>THAI NOI / LAO BUHAN LETTER QBA</td>
<td></td>
</tr>
<tr>
<td>໛</td>
<td>໛</td>
<td>THAI NOI / LAO BUHAN LETTER PA</td>
<td></td>
</tr>
<tr>
<td>ໜ</td>
<td>ໜ</td>
<td>THAI NOI / LAO BUHAN LETTER PHA</td>
<td></td>
</tr>
<tr>
<td>ໝ</td>
<td>ໝ</td>
<td>THAI NOI / LAO BUHAN LETTER FA</td>
<td></td>
</tr>
<tr>
<td>ໞ</td>
<td>ໞ</td>
<td>THAI NOI / LAO BUHAN LETTER BA</td>
<td></td>
</tr>
<tr>
<td>ໟ</td>
<td>ໟ</td>
<td>THAI NOI / LAO BUHAN LETTER YA</td>
<td></td>
</tr>
<tr>
<td>໠</td>
<td>໠</td>
<td>THAI NOI / LAO BUHAN LETTER BHA</td>
<td></td>
</tr>
<tr>
<td>໡</td>
<td>໡</td>
<td>THAI NOI / LAO BUHAN LETTER MA</td>
<td></td>
</tr>
<tr>
<td>໢</td>
<td>໢</td>
<td>THAI NOI / LAO BUHAN LETTER YA</td>
<td></td>
</tr>
<tr>
<td>໣</td>
<td>໣</td>
<td>THAI NOI / LAO BUHAN LETTER QYA</td>
<td></td>
</tr>
<tr>
<td>໤</td>
<td>໤</td>
<td>THAI NOI / LAO BUHAN LETTER RA</td>
<td></td>
</tr>
<tr>
<td>໥</td>
<td>໥</td>
<td>THAI NOI / LAO BUHAN LETTER RY / VOCALIC R</td>
<td></td>
</tr>
<tr>
<td>໦</td>
<td>໦</td>
<td>THAI NOI / LAO BUHAN LETTER RRY / VOCALIC RR</td>
<td></td>
</tr>
<tr>
<td>໧</td>
<td>໧</td>
<td>THAI NOI / LAO BUHAN LETTER LA</td>
<td></td>
</tr>
</tbody>
</table>

* Further sources and experts should be consulted to ensure the presented forms are as representative as they can be. The vocalic consonants especially should be disentangled.

† As Theppitak Karoonbanyanan notes, these forms of subjoined Da derive from the original subjoined Sa.