

5. Lao

Lao language is derived from Kam-Tai branch of Tai-Kadai language spoken by approximately 3 million people in Laos and Thailand [17]. Traditional Lao literature has been written in Lao and Tham scripts. The Lao script emerged in 13th [18] or 14th century [19], deriving mutually with old Thai script from Brahmi writing system. Lao script was simplified in 1960, making it more regular [18].

5.1. Writing System

5.1.1. Character Set

Like Indic scripts, Lao script consonants also carry an inherent vowel, and in addition an inherent tone, both of which can be over-ridden by explicitly specifying them. Lao script has 27 consonants which are divided into three classes, high, middle and low. This grouping helps in determining the tone of the syllable, along with the tone marks and vowels. These consonants are given in Figure 5.1. Vowels are always written around a central consonant. Vowels occur in full form or as marks which can attach before, after, above or below the consonant. Lao vowels are shown in Figure 5.2 [21]. Slightly variant vowel list is reported in [4].

ກ ຂ ຄ ງ ຈ ສ ຊ ຍ ດ ຕ ຖ ທ ນ ບ ປ
 ຜ ຝ ພ ຟ ມ ຢ ຣ ລ ວ ຫ ອ ຮ

Figure 5.1. Lao Consonants

ຂ ື ື ູ ເXຂ ແXຂ ໂXຂ ເXາຂ ເ ື ເັຍ ເືອ ືວຂ
 Short Vowels
 Xາ ື ື ູ ເX ແX ໂX ະ ເ ື ເXຍ ເ ືອ ືວ
 Long Vowels
 ໂX ໃX ເ ືາ ະາ
 Diphthongs

Figure 5.2. Lao Vowels [21]
 (X used as a placeholder for a consonant)

Lao script also has four tone marks, shown in Figure 5.3.

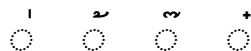


Figure 5.3. Lao Tone Marks

Lao also possesses special characters shown in Table 5.1.

Table 5.1. Lao Special Characters

Name	Glyph
Mai Sum (Sentence Repetition)	໑
Mai Sum (Word Repetition)	໒
Mai Kalan	໐

Mai Sum (໑, ໒) are used for sentence and word repetition. These are used instead of writing the whole sentence or whole word again. Mai Kalan is used with foreign words and is optional.

Lao has its own set of numerals given in Figure 5.4.



Figure 5.4. Lao Digits

5.1.2. Script Details

5.1.2.1. No Word Spacing

Like other South-East Asian scripts such as Thai and Burmese, Lao does not have spaces between words. Native readers identify word boundaries using their tacit knowledge of the language. Text is written in continuum and space is only used at the end of sentence or clause.

5.1.2.2. Vowel and Tone Marks

Vowels are used in conjunction with consonants to modify the way they are pronounced. They attach at the front, back, top or bottom of the consonant. Unlike Indic languages multiple vowels can attach to a consonant. These variations are shown in Table 5.2.

Table 5.2. Lao Vowels with Consonant KO

ເກ	ເກ	Connects to Left
ກະ	ກະ	Connects to Right
ກຸ	ກຸ	Connects at Bottom
ກິ	ກິ	Connects at Top
ເກິ	ເກິ	Connects to Left and Top

The tone marks are always placed above the consonants. If there is already a vowel above consonant, the tone mark will stack above the vowel, as shown in Table 5.3.

Table 5.3. Placement of Lao Tone Marks

ເກິ	ເກິ	Above the Consonant
ກິ	ກິ	Above the Vowel

Further details are given in the discussion on syllable structure later.

5.1.2.3. Syllable and Syllabification

Lao is a syllable based language. The syllables are structured around a central consonant (also known as main or nuclear consonant). A syllable might optionally have combinational consonants, at least one vowel which may be placed before, after, above or below the main consonant, and up to one tone mark. This is illustrated in Figure 5.5 below. Capital C indicates the nuclear consonant. The subscripts “0..n” mean zero to *n*, indicating that all are optional (in case of zero) except the nuclear **C**.

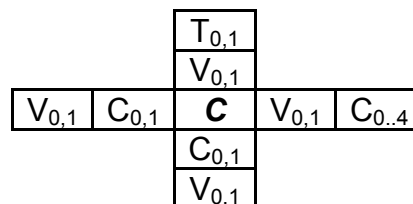


Figure 5.5. Generic Syllable Structure for Lao (C = Consonant; V = Vowel; T = Tone Mark)

A detailed syllable template for Lao is shown Figure 5.6. X₀ through X₁₀ are explained below.

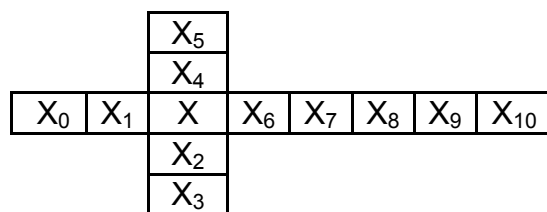


Figure 5.6. Detailed Syllable Structure for Lao

- X_0 represents a vowel which always occurs before the nuclear consonant X.
- X_1 is a combination consonant ຫ which comes before the nuclear consonant, only if the nuclear consonant is one of {ງ, ຍ, ນ, ມ, ລ, ວ}. It can also occur before ູ.
- X represents the nuclear consonants.
- X_2 is ູ and comes only when ຫ occurs as X_1 (in this case, there will be no nuclear consonant) and the combination forms the nuclear consonant.
- X_3 represents vowels which occur under the nuclear consonant.
- X_4 represents vowel which occur above the nuclear consonant.
- X_5 represents tone marks which appear above nuclear consonant or above vowels.
- X_6 represents consonant vowel, which occurs after nuclear consonant. This functions as vowel when the syllable does not have any vowels, and always appear with X_8 .
- X_7 represents an after-vowel. However X_{71} always indicates the end of syllable and it never exists with a tone mark.
- X_8 represents alternate consonants.
- X_9 represents alternate consonant to pronounce foreign language words. It always exists with X_{10} .
- X_{10} represents different marks as discussed in Table 5.1. Mai Sum may be considered outside the syllable.

The following Table 5.4 further classifies where each Lao character can occur. A character can fall under multiple categories depending upon its position in syllable.

Table 5.4. Positional Restrictions on Lao Characters in a Syllable

X_0	X_1	X	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
ໄ= X_{01}	ຫ	ກ ຂ ຄ	ູ	ຸ	ົ= X_{41}	ັ	ວ= X_{61}	ຮ= X_{71}	ກ	ຈ	ຢ
ໄ໊= X_{02}		ງ ຈ ຊ		ູ	ົ໊= X_{42}	ັ໊	ອ= X_{62}	າ= X_{72}	ງ	ສ	ງ
ໄ໋= X_{03}		ຍ ດ ຕ			ົ໋= X_{43}	ັ໋	ຯ= X_{63}	າ໋= X_{73}	ຍ	ຊ	ັ

ໄ=X ₀₄ ໃ=X ₀₅	ຖ ບ ປ	ື=X ₄₄ ື̇=X ₄₅ ື̈=X ₄₆ ື̉=X ₄₇	ື̇						ດ ພ
	ຜ ຝ ພ								ນ ພ
	ຟ ມ ຢ								ມ ລ
	ຮ ລ ຫ								ບ
	ອ ຮ ຫ								ວ
	ໝ ວ ສ								
ທ ນ									

Syllable boundaries are detected based on a set of conditions. For example the syllable ເກີດ satisfies condition: $X_{01}(X_1) X(X_2) X_{4_1} | X_{4_2} (X_5) (X_8) (X_9: X_{103}) (X_{10_1} | X_{10_2})$. It states that a syllable that fulfills this condition must have vowel X_{01} ຕ. Combinational consonants X_1 and X_2 are optional. It should have a main consonant X which is ກ in this example string. It must have one of the two vowels X_{41} or X_{42} (ື or ື̇). Tone mark X_5 and consonants X_8 and X_9 are also optional. Moreover if X_9 occurs it must be followed by X_{103} . One of the X_{101} or X_{102} can occur optionally. The syllable template is filled for this string in Figure 5.7.

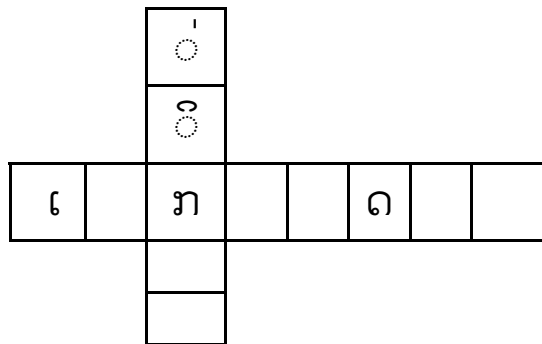


Figure 5.7. Syllable Template Filled for Lao String ເກີດ

Further algorithmic details and a complete set of syllabification rules are given in [20].

5.2. Collation

Two different strategies are commonly used in Lao for collation. One of these uses base characters and collapses them into bigger linguistic units and assigns a single collation element

per unit. The second strategy does not collapse the input characters and assigns a single collation element to each character in the script. The two mechanisms are known as language based versus script based collation.

Lao language has syllable based collation. The word is subdivided into a sequence of syllables for sorting. Then, given two words, their initial syllables are compared. The second syllables of these words are only compared if the first syllables are identical, and so on. This strategy is significantly different from Unicode Collation Algorithm [2] discussed in Chapter 2. In the earlier algorithm, after collation elements are assigned, a single sort key is generated for each word for a single comparison with other sort keys from other words. However, in the case of Lao, there will be a sort key generated for each syllable (not word!). The comparison of words will be an iterative process which compares sort keys of each syllable in one word in sequence, with corresponding sort keys of syllables in other word. These comparisons will be done until a difference is found.

Within the syllable, Lao sorts at four levels, with nuclear consonants getting the primary weight, vowels getting the secondary weight, alternate consonants getting the tertiary weight and tone marks getting the quaternary level weight. Punctuation marks and some other Lao characters are ignorable at all levels. Most popular Lao dictionaries generally agree on the order of consonants, though differences lie in the ordering of vowels and combining consonants.

5.2.1. Text Processing

5.2.1.1. Syllabification

Lao strings are collated based on syllable sequence. Thus, it is critical to syllabify the strings to be compared. This can be done through advanced language processing techniques, both by extensive rule-based systems [4], or using statistical methods. Some initial details for rule based syllabification are provided in Section 5.1.2.3. There has been very limited work on statistical solutions for Lao syllabification.

5.2.1.2. Syllable Parsing

Lao characters behave differently in collation depending on where they occur in a syllable. For example, consonants get primary weight in collation if they occur as main consonant X, combinational consonant X₁, X₂ but tertiary weight if they occur in a secondary role as X₈ and X₉, as given in Table 5.4. Thus, the syllabification process should not only return syllable boundaries

but also label the role of each character within the syllable string. This implies that complete internal parsing of syllable is also desired. Lao Letter WO ວ can play as X, X₂, X₆ and X₈ in a syllable. Therefore it can acquire primary (when X or X₂), secondary (when X₆) or tertiary (when X₈) weight.

5.2.1.3. Reordering

The syllable structure illustrated in Figures 5.5 and 5.6 shows that a main consonant can be preceded optionally by another consonant and a dependent vowel. Like other Indic scripts, these characters are logically treated to occur after the central consonant. However, unlike encoding of South Asian scripts, Unicode encodes Lao characters in visual order (for backward compatibility with earlier systems for Thai and Lao). Thus, the characters have to be reordered into the logical order for collation.

In addition to typing the initial vowel and combinational consonant before the main consonant, there can be many ways of typing the characters following this main consonant in a syllable. For example, the string ຫຼີ can be generated by the sequence ຫ + ື + ື or alternatively by the sequence ຫ + ື + ື. These differences can also cause inconsistent collation results. This inconsistency in collation is shown in Table 5.5. below. Different character sequences result in different sort keys using the same collation elements causing different sorting order. Thus, reordering of all characters in a syllable needs to be conducted in a consistent order before sorting can proceed.

Table 5.5. Differences in Sort Keys Caused by Variation in Character Sequence

Syllable	Collation Elements	Sort Key
ກ		
ກ+ເ	[0820 0200 0020 0002] [0000 021A 0020 0002]	[0820 0000 <u>0200 021A</u> 0000 0020 0020 0000 0002 0002]
ເ+ກ	[0000 021A 0020 0002] [0820 0200 0020 0002]	[0820 0000 <u>021A 0200</u> 0000 0020 0020 0000 0002 0002]

The desired order of characters in a syllable in Figure 5.6. is X (main consonant) X₁ X₂ (combinational consonants), X₀ X₃ X₄ X₆ X₇ (vowels), X₈ X₉ (alternate consonants), X₅ (tone

However, each combined form maps onto a single vowel and thus a single collation element. Therefore, the contractions in Table 5.7 are needed to achieve Lao collation.

Table 5.7. Contraction to Single Collation Element from Multiple Encoded Characters

Glyph	Unicode for Contraction
Ꞩ + ິ = ິ	0EC0 + 0EB4
Ꞩ + ື = ື	0EC0 + 0EB5
ົ + ວ + ຮ = ົຽ	0EBB + 0EA7 + 0EB0
ົ + ວ = ົ	0EBB + 0EA7
Ꞩ + ື + ອ = ືອ	0EC0 + 0EB6 + 0EAD
Ꞩ + ື + ອ = ືອ	0EC0 + 0EB7 + 0EAD
Ꞩ + ົ + າ = ົາ	0EC0 + 0EBB + 0EB2
Ꞩ + ຮ = ຮ	0EC0 + 0EB0
Ꞩ + າ + ຮ = ຮາ	0EC0 + 0EB2 + 0EB0
ແ + ຮ = ແຮ	0EC1 + 0EB0
ໂ + ຮ = ໂຮ	0EC2 + 0EB0
Ꞩ + ຍ = ຮຍ	0EC0 + 0E8D
Ꞩ + ັ + ຍ = ຮັຍ	0EC0 + 0EB1 + 0E8D
ໍ + າ = ົາ	0ECD + 0EB2

5.2.2. Unicode Collation Elements

Lao language dictionaries follow two different collation sequences, which may be termed as Lao language-based (e.g. [22]) and script-based collation. Language-based collation uses the encoded vocalic symbols (given in Figure 5.8) to do the context based contractions (given in

Table 5.7) to form singular vowels (given in Figure 5.2). A Collation element is then assigned to each vowel in Figure 5.2.

Script-based collation does not perform the contractions discussed but assigns collation element to each script symbol given in Figure 5.8. Thus, the collation is not done on basis of vowels but individual script symbols used for forming these vowels.

Syllabification, syllable based parsing, re-ordering and normalization is done in the same manner as discussed for both strategies. The difference in the strategies is just in contraction and eventual collation assignment process. Collation elements for the two strategies are also different and are given in Tables 5.8 and 5.9.

5.2.2.1. Language Based Sorting

The collation elements for language based sorting are given in Table 5.8.

Table 5.8. Lao Collation Elements for Language Based Sorting

Glyph	Unicode	Collation Elements	Unicode Name
← Consonants →			
ກ	0E81	0820 0200 0020 0002	LAO LETTER KO
ຂ	0E82	0822 0200 0020 0002	LAO LETTER KHO SUNG
ຄ	0E84	0824 0200 0020 0002	LAO LETTER KHO TAM
ງ	0E87	0826 0200 0020 0002	LAO LETTER NGO
ຈ	0E88	0828 0200 0020 0002	LAO LETTER CO
ສ	0EAA	082A 0200 0020 0002	LAO LETTER SO SUNG
ຊ	0E8A	082C 0200 0020 0002	LAO LETTER SO TAM
ຢ	0E8D	082E 0200 0020 0002	LAO LETTER NYO
ດ	0E94	0830 0200 0020 0002	LAO LETTER DO
ຕ	0E95	0832 0200 0020 0002	LAO LETTER TO
ຖ	0E96	0834 0200 0020 0002	LAO LETTER THO SUNG

ທ	0E97	0836 0200 0020 0002	LAO LETTER THO TAM
ນ	0E99	0838 0200 0020 0002	LAO LETTER NO
ບ	0E9A	083A 0200 0020 0002	LAO LETTER BO
ປ	0E9B	083C 0200 0020 0002	LAO LETTER PO
ຜ	0E9C	083E 0200 0020 0002	LAO LETTER PHO SUNG
ຝ	0E9D	0840 0200 0020 0002	LAO LETTER FO TAM
ພ	0E9E	0842 0200 0020 0002	LAO LETTER PHO TAM
ຟ	0E9F	0844 0200 0020 0002	LAO LETTER FO SUNG
ມ	0EA1	0846 0200 0020 0002	LAO LETTER MO
ຢ	0EA2	0848 0200 0020 0002	LAO LETTER YO
ຣ	0EA3	084A 0200 0020 0002	LAO LETTER LO LING
ຣ̣	0EBC	084C 0200 0020 0002	LAO SEMI VOWEL SIGN LO
ລ	0EA5	084E 0200 0020 0002	LAO LETTER LO LOOT
ວ	0EA7	0850 0200 0020 0002	LAO LETTER WO
ຫ	0EAB	0852 0200 0020 0002	LAO LETTER HO SUNG
ຫງ	0EAB+0E87	0854 0200 0020 0002	LAO LETTER HO SUNG+ LAO LETTER NGO
ຫຍ	0EAB+0E8D	0856 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER NYO
ຫນ	0EAB+0E99	0858 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER NO
ໜ	0EDC	0858 0200 0020 0002	LAO LETTER HO NO
ໜມ	0EAB+0EA1	0860 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER MO
ໝ	0EDD	0860 0200 0020 0002	LAO LETTER HO MO

ຫລ	0EAB+0EA5	0864 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER LO LOOT
ຫຼ	0EAB+0EBC	0864 0200 0020 0002	LAO LETTER HO SUNG + LAO SEMIVOWEL SIGN LO
ຫວ	0EAB+0EA7	0868 0200 0020 0002	LAO LETTER HO SUNG + LAO LETTER WO
ອ	0EAD	086A 0200 0020 0002	LAO LETTER O
ຮ	0EAE	086C 0200 0020 0002	LAO LETTER HO TAM
← Vowels →			
ຮ	0EB0	0000 0202 0020 0002	LAO VOWEL SIGN A
້+X8/X9	0EB1+X8/X9	0000 0204 0020 0002	LAO VOWEL SIGN MAI KAN + CONSONANTAL
າ	0EB2	0000 0206 0020 0002	LAO VOWEL SIGN AA
ິ	0EB4	0000 0208 0020 0002	LAO VOWEL SIGN I
ີ	0EB5	0000 020A 0020 0002	LAO VOWEL SIGN II
ົງ	0EB6	0000 020C 0020 0002	LAO VOWEL SIGN Y
ົງ	0EB7	0000 0210 0020 0002	LAO VOWEL SIGN YY
ູ	0EB8	0000 0212 0020 0002	LAO VOWEL SIGN U
ູ	0EB9	0000 0214 0020 0002	LAO VOWEL SIGN UU
ເXຮ	0EC0+X+0EB0	0000 0216 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT + LAO VOWEL SIGN A
້+X8/X9	0EC0+0EB1+X8/X9	0000 0218 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN MAIN KAN + CONSONANTAL
ເX	0EC0+X	0000 021A 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT
ເXຮ	0EC1+X+0EB0	0000 021C 0020 0002	LAO VOWEL SIGN EI + MAIN CONSONANT + LAO VOWEL SIGN A
້+X8/X9	0EC1+0EB1+X8/X9	0000 0220 0020 0002	LAO VOWEL SIGN EI + LAO VOWEL SIGN MAI KAN + CONSONANTAL
ເX	0EC1+X	0000 0222 0020 0002	LAO VOWEL SIGN EI + MAIN CONSONANT

ໂຂະ	0EC2+X+0EB0	0000 0224 0020 0002	LAO VOWEL SIGN O + MAIN CONSONANT + LAO VOWEL SIGN A
ົ	0EBB	0000 0226 0020 0002	LAO VOWEL SIGN MAI KON
ໂຂ	0EC2+X	0000 0228 0020 0002	LAO VOWEL SIGN O + MAIN CONSONANT
ເຂາະ	0EC0+X+0EB2+0EB0	0000 022A 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT + LAO VOWEL AA + LAO VOWEL SIGN A
ໍ	0ECD	0000 022C 0020 0002	LAO NIGGAHITA
Xອ+X8/X9	X+0EAD+X8/X9	0000 022E 0020 0002	MAIN CONSONANT + LAO LETTER O + CONSONANTAL
ເື	0EC0+0EB4	0000 0230 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN I
ເືີ	0EC0+0EB5	0000 0232 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN II
ເື້	0EC0+0EB1+0EBD	0000 0234 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN MAI KAN + LAO SEMIVOWEL SIGN NYO
ເຂງ	0EC0+X+0EBD	0000 0236 0020 0002	LAO VOWEL SIGN E + MAIN CONSONANT + LAO SEMI VOWEL SIGN NYO
ງ+X8/X9	0EBD+X8/X9	0000 0238 0020 0002	LAO SEMI VOWEL SIGN NYO + CONSONANTAL
ົວະ	0EBB+0EA7+0EB0	0000 023A 0020 0002	LAO VOWEL SIGN MAI KON + LAO LETTER WO + LAO + VOWEL SIGN A
້ວ+X8/X9	0EB1+0EA7+X8/X9	0000 023C 0020 0002	LAO VOWEL SIGN MAI KON + LAO LETTER WO + CONSONANTAL
ົວ	0EBB+0EA7	0000 023E 0020 0002	LAO VOWEL SIGN MAI KON + LAO LETTER WO
ື້ອ	0EC0+0EB6+0EAD	0000 0240 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN Y + LAO LETTER O
ື້ອ	0EC0+0EB7+0EAD	0000 0242 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN YY + LAO LETTER O
Xວ+X8/X9	X+0EA7+X8/X9	0000 0244 0020 0002	MAIN CONSONANT + LAO LETTER WO + CONSONANTAL
ໂຂ	0EC4+X	0000 0246 0020 0002	LAO VOWEL SIGN AI + MAIN CONSONANT
ໃຂ	0EC3+X	0000 0248 0020 0002	LAO VOWEL SIGN AY + MAIN CONSONANT

ົ້າ	0EC0+0EBB+0EB2	0000 024A 0020 0002	LAO VOWEL SIGN E + LAO VOWEL SIGN MAI KON + LAO VOWEL SIGN AA
ໍ່າ	0EB3	0000 024C 0020 0002	LAO VOWEL SIGN AM
ໍ່າ	0ECD+0EB2	0000 024C 0020 0002	LAO NIGGAHITA + LAO VOWEL SIGN AA
← Alternate Consonants →			
ກ	0E81	0000 0000 0022 0002	LAO LETTER KO
ງ	0E87	0000 0000 0024 0002	LAO LETTER NGO
ຍ	0E8D	0000 0000 002C 0002	LAO LETTER NYO
ດ	0E94	0000 0000 002E 0002	LAO LETTER DO
ນ	0E99	0000 0000 0030 0002	LAO LETTER NO
ບ	0E9A	0000 0000 0032 0002	LAO LETTER BO
ມ	0EA1	0000 0000 0038 0002	LAO LETTER MO
ວ	0EA7	0000 0000 003C 0002	LAO LETTER WO
← Tone Marks →			
◌̄	0EC8	0000 0000 0000 0004	LAO TONE MAI EK
◌̅	0EC9	0000 0000 0000 0006	LAO TONE MAI THO
◌̆	0ECA	0000 0000 0000 0008	LAO TONE TI
◌̇	0ECB	0000 0000 0000 0008	LAO TONE MAI CATAWA
← Numerals →			
໐	0ED0	0700 0200 0020 0002	LAO DIGIT ZERO
໑	0ED1	0702 0200 0020 0002	LAO DIGIT ONE
໒	0ED2	0704 0200 0020 0002	LAO DIGIT TWO
໓	0ED3	0706 0200 0020 0002	LAO DIGIT THREE

໔	0ED4	0708 0200 0020 0002	LAO DIGIT FOUR
໕	0ED5	070A 0200 0020 0002	LAO DIGIT FIVE
໖	0ED6	070C 0200 0020 0002	LAO DIGIT SIX
໗	0ED7	070E 0200 0020 0002	LAO DIGIT SEVEN
໘	0ED8	0710 0200 0020 0002	LAO DIGIT EIGHT
໙	0ED9	0712 0200 0020 0002	LAO DIGIT NINE
← Various Symbols →			
່	0ECC	0000 0000 0000 0000	MAI KALAN
໊	0EC6	0000 0000 0000 0000	MAI SUM
໋	0EAF	0000 0000 0000 0000	MAI SUM

5.2.2.2. Script Based Sorting

The collation elements for language based sorting are given in Table 5.9.

Table 5.9. Lao Collation Elements for Script Based Sorting

Glyph	Unicode	Collation Elements	Unicode Name
← Consonants →			
ກ	0E81	0820 0200 0020 0002	LAO LETTER KO
ຂ	0E82	0822 0200 0020 0002	LAO LETTER KHO SUNG
ຄ	0E84	0824 0200 0020 0002	LAO LETTER KHO TAM
ງ	0E87	0826 0200 0020 0002	LAO LETTER NGO
ຈ	0E88	0828 0200 0020 0002	LAO LETTER CO
ສ	0EAA	082A 0200 0020 0002	LAO LETTER SO SUNG
ຮ	0E8A	082C 0200 0020 0002	LAO LETTER SO TAM

ຢ	0E8D	082E 0200 0020 0002	LAO LETTER NYO
ດ	0E94	0830 0200 0020 0002	LAO LETTER DO
ຕ	0E95	0832 0200 0020 0002	LAO LETTER TO
ຖ	0E96	0834 0200 0020 0002	LAO LETTER THO SUNG
ທ	0E97	0836 0200 0020 0002	LAO LETTER THO TAM
ນ	0E99	0838 0200 0020 0002	LAO LETTER NO
ບ	0E9A	083A 0200 0020 0002	LAO LETTER BO
ປ	0E9B	083C 0200 0020 0002	LAO LETTER PO
ຜ	0E9C	083E 0200 0020 0002	LAO LETTER PHO SUNG
ຝ	0E9D	0840 0200 0020 0002	LAO LETTER FO TAM
ພ	0E9E	0842 0200 0020 0002	LAO LETTER PHO TAM
ຟ	0E9F	0844 0200 0020 0002	LAO LETTER FO SUNG
ມ	0EA1	0846 0200 0020 0002	LAO LETTER MO
ຢ	0EA2	0848 0200 0020 0002	LAO LETTER YO
ຣ	0EA3	084A 0200 0020 0002	LAO LETTER LO LING
ລ	0EA5	084E 0200 0020 0002	LAO LETTER LO LOOT
ວ	0EA7	0850 0200 0020 0002	LAO LETTER WO
ຫ	0EAB	0852 0200 0020 0002	LAO LETTER HO SUNG
ຫຼ	0EAB+0EBC	0866 0200 0020 0002	LAO LETTER HO SUNG + LAO SEMIVOWEL SIGN LO
ອ	0EAD	086A 0200 0020 0002	LAO LETTER O
ຮ	0EAE	086C 0200 0020 0002	LAO LETTER HO TAM
ໜ	0EDC	0870 0200 0020 0002	LAO LETTER HO NO

ໝ	0EDD	0872 0200 0020 0002	LAO LETTER HO MO
← Vowels →			
ຮ	0EB0	0000 0202 0020 0002	LAO VOWEL SIGN A
ຯ	0EB2	0000 0206 0020 0002	LAO VOWEL SIGN AA
໐	0EB4	0000 0208 0020 0002	LAO VOWEL SIGN I
໑	0EB5	0000 020A 0020 0002	LAO VOWEL SIGN II
໒	0EB6	0000 020C 0020 0002	LAO VOWEL SIGN Y
໓	0EB7	0000 0210 0020 0002	LAO VOWEL SIGN YY
໔	0EB8	0000 0212 0020 0002	LAO VOWEL SIGN U
໕	0EB9	0000 0214 0020 0002	LAO VOWEL SIGN UU
໖	0EC0	0000 0216 0020 0002	LAO VOWEL SIGN
໗	0EC1	0000 0222 0020 0002	LAO VOWEL SIGN EI
໘	0EC2	0000 0224 0020 0002	LAO VOWEL SIGN O
໙	0ECD	0000 022C 0020 0002	LAO NIGGAHITA
໑໐	0EC4	0000 0246 0020 0002	LAO VOWEL SIGN AI
໑໑	0EC3	0000 0248 0020 0002	LAO VOWEL SIGN AY
໑໒	0EB1	0000 024A 0020 0002	LAO VOWEL SIGN MAI KAN
໑໓	0EBB	0000 024C 0020 0002	LAO VOWEL SIGN MAI KON
໑໔	0EBD	0000 0250 0020 0002	LAO SEMI VOWEL SIGN NYO
໑໕	0EA7	0000 0252 0020 0002	LAO LETTER WO
໑໖	0EAD	0000 0254 0020 0002	LAO LETTER O
← Consonantal →			
ກ	0E81	0000 0000 0022 0002	LAO LETTER KO

ງ	0E87	0000 0000 0024 0002	LAO LETTER NGO
ຈ	0E88	0000 0000 0026 0002	LAO LETTER CO
ສ	0EAA	0000 0000 0028 0002	LAO LETTER SO SUNG
ຊ	0E8A	0000 0000 002A 0002	LAO LETTER SO TAM
ຢ	0E8D	0000 0000 002C 0002	LAO LETTER NYO
ດ	0E94	0000 0000 002E 0002	LAO LETTER DO
ນ	0E99	0000 0000 0030 0002	LAO LETTER NO
ບ	0E9A	0000 0000 0032 0002	LAO LETTER BO
ຟ	0E9E	0000 0000 0034 0002	LAO LETTER PHO TAM
ຟ	0E9F	0000 0000 0036 0002	LAO LETTER FO SUNG
ມ	0EA1	0000 0000 0038 0002	LAO LETTER MO
ລ	0EA5	0000 0000 003A 0002	LAO LETTER LO LOOT
ວ	0EA7	0000 0000 003C 0002	LAO LETTER WO
← Tone Marks →			
◌̄	0EC8	0000 0000 0000 0004	LAO TONE MAI EK
◌̅	0EC9	0000 0000 0000 0006	LAO TONE MAI THO
◌̆	0ECA	0000 0000 0000 0008	LAO TONE TI
◌̇	0ECB	0000 0000 0000 0008	LAO TONE MAI CATAWA
← Numerals →			
໐	0ED0	0700 0200 0020 0002	LAO DIGIT ZERO
໑	0ED1	0702 0200 0020 0002	LAO DIGIT ONE
໒	0ED2	0704 0200 0020 0002	LAO DIGIT TWO
໓	0ED3	0706 0200 0020 0002	LAO DIGIT THREE

໔	0ED4	0708 0200 0020 0002	LAO DIGIT FOUR
໕	0ED5	070A 0200 0020 0002	LAO DIGIT FIVE
໖	0ED6	070C 0200 0020 0002	LAO DIGIT SIX
໗	0ED7	070E 0200 0020 0002	LAO DIGIT SEVEN
໘	0ED8	0710 0200 0020 0002	LAO DIGIT EIGHT
໙	0ED9	0712 0200 0020 0002	LAO DIGIT NINE
← Various Symbols →			
ໍ	0ECC	0000 0000 0000 0000	MAI KALAN
ງ	0EC6	0000 0000 0000 0000	MAI SUM
ຯ	0EAF	0000 0000 0000 0000	MAI SUM

Results

Data sorted by different strategies gives different output sequences. Sample output sequences for each technique are given in Tables 5.10 and 5.11.

Table 5.10. Input and Corresponding Sorted Output for Lao Using Language Based Strategy

Sample Input		Sample Output	
ເງິນແຮງຖົງ	ກົກເສົາ	ກະໂຄງການ	ກິ້ນໜັກ
ກາ	ກິ້ນຂີ້	ກະຕື້ລີ້ລິ້ນ	ກໍຈິງຢູ່ແລ້ວ
ຈົດທະບຽນ	ກິ້ນຂີ້ທັງ	ກະຕື້ລີ້ລິ້ນ	ກອກນ້ໍາ
ສຽງມ້າແຫມ	ຈົດທະບຽນການຄ້າ	ກະແຕະ	ກອງກິ້ນ
ສຽງທອງ	ງານຂຶ້ນເຮືອນໃໝ່	ກະແຕະ	ກອງໂຈນ
ເກືອບໝົດ	ກາກະບາດ	ກະແຕ	ກອງສອດແນມ
ສີກຸຫລາບ	ສຽງຟ້າຮ້ອງ	ກະເຕາະກະແຕະ	ເກີດຈາກ
ເງິນຮາງ	ແກ້ວໂກເມນ	ກັບຄືນມາ	ເກີດໄພ
ກ້ວຍ	ກິ້ນຂວດ	ກາ	ເກີດມາ
ໂກເທົ້າໃດ	ກິ້ນໜັກ	ກາກະບາດ	ກົວ
ກະແຕ	ກໍຈິງຢູ່ແລ້ວ	ກາຄໍາຊອບ	ກົ່ວ

ເກີດມາ	ເຄື່ອງວັດຄວາມໄວ	ກາໂຕລິກ	ເກືອບຫມົດ
ກາຄຳຊອບ	ກອກນ້ຳ	ກ້າ	ກວຍ
ເກົ່າ	ກົກຂາ	ກ້າ	ກ້ວຍ
ກະໂຄງການ	ກ້າແກ່ນ	ກ້າ	ກ້ວຍມືນາງ
ກ້າ	ກັບຄືນມາ	ກ້າກັນ	ໂກເທົາໃດ
ກ້າ	ເກີດຈາກ	ກ້າແກ່ນ	ໂກປານໃດ
ກິນິນ	ເກີດໄພ	ກິນິນ	ໃກ້ຊິດກັນ
ກາໂຕລິກ	ກອງສອດແນມ	ກິຣິຍາ	ເກົາ
ກິຣິຍາ	ຈັກກະພັດນິຍົມ	ກີ້ເຕາ	ເກົາ
ກີ້ເຕາ	ກະເຕາະກະແຕະ	ກຶ້ງຕາໃສ່	ເຄື່ອງວັດຄວາມໄວ
ກຶ້ງຕາໃສ່	ກອງກິນ	ກຸຫລາບ	ເຄື່ອງວັດຄວາມຮ້ອນແ
ກ້າກັນ	ກົວ	ກູລີ	ຍັນ
ກ້າ	ກວຍ	ເກັດປາ	ງານຂຶ້ນເຮືອນໃຫມ່
ກູລີ	ກ້ວຍມືນາງ	ເກັບ	ເງິນຮາງ
ກຸຫລາບ	ໃກ້ຊິດກັນ	ເກັບກຸ່ວ	ເງິນແຮຖົງ
ເກັດປາ	ກະແຕະ	ເກັບກຸ່ວ	ຈັກກະພັດນິຍົມ
ສຽງແທບ	ໂກປານໃດ	ເກັບພາສີ	ຈົດທະບຽນ
ເກັບ	ເກົາ	ແກ້ວໂກເມນ	ຈົດທະບຽນການຄ້າ
ກະຕື້ລີ້ລົ້ນ	ກອງໂຈນ	ແກ້ວນາເຝານເນື້ອແ	ສີກຸຫລາບ
ເກັບກຸ່ວ	ຊັກອອກພູດນຶ່ງ	ຂງ	ສີແກ່
ແກ້ວນາເຝານເນື້ອແ	ກົວ	ກົກ	ສຽງທອງ
ຂງ	ສີແກ່	ກົກຂາ	ສຽງປົກກະຕິ
ກົກ	ເຄື່ອງວັດຄວາມຮ້ອນແ	ກົກແຂນ	ສຽງພ້າຮ້ອງ
ເກັບກຸ່ວ	ຍັນ	ກົກເສົາ	ສຽງມ້າແທມ
ເກັບພາສີ	ກະແຕະ	ກັນຂີ້	ສຽງແທບ
ກະຕື້ລີ້ລົ້ນ	ສຽງປົກກະຕິ	ກັນຂີ້ທັງ	ຊັກອອກພູດນຶ່ງ
ກົກແຂນ		ກັນຂວດ	

Table 5.11. Input and Corresponding Sorted Output for Lao Using Script Based Strategy

Sample Input		Sample Output	
ເງິນແຮຖົງ	ກົກເສົາ	ກະໂຄງການ	ໂກປານໃດ
ກາ	ກັນຂີ້	ກະຕື້ລີ້ລົ້ນ	ໃກ້ຊິດກັນ
ຈົດທະບຽນ	ກັນຂີ້ທັງ	ກະຕື້ລີ້ລົ້ນ	ກັບຄືນມາ
ສຽງມ້າແທມ	ຈົດທະບຽນການຄ້າ	ກະເຕາະກະແຕະ	ກົກ
ສຽງທອງ	ງານຂຶ້ນເຮືອນໃຫມ່	ກະແຕ	ກົກຂາ
ເກືອບຫມົດ	ກາກະບາດ	ກະແຕະ	ກົກແຂນ

ສີກຸຫລາບ	ສຽງຟ້າຮ້ອງ	ກະແຕະ	ກົກເສົາ
ເງິນຮາງ	ແກ້ວໂກເມນ	ກາ	ກິ້ນຂີ້
ກ້ວຍ	ກິ້ນຂວດ	ກາກະບາດ	ກິ້ນຂີ້ທັງ
ໄກເທົ່າໃດ	ກິ້ນຫນັກ	ກາຄຳຊອບ	ກິ້ນຂວດ
ກະແຕ	ກໍຈິງຢູ່ແລ້ວ	ກາໂຕລິກ	ກິ້ນຫນັກ
ເກີດມາ	ເຄື່ອງວັດຄວາມໄວ	ກ້າ	ກົວ
ກາຄຳຊອບ	ກອກນ້ຳ	ກ້າ	ກົວ
ເກົາ	ກົກຂາ	ກ້າ	ກວຍ
ກະໂຄງການ	ກ້າແກ່ນ	ກ້າແກ່ນ	ກ້ວຍ
ກ້າ	ກັບຄືນມາ	ກ້າກິ້ນ	ກ້ວຍມືນາງ
ກ້າ	ເກີດຈາກ	ກິນິນ	ກອກນ້ຳ
ກິນິນ	ເກີດໄພ	ກິຣິຍາ	ກອງກິ້ນ
ກາໂຕລິກ	ກອງສອດແນມ	ກີ້ເຕາ	ກອງໂຈນ
ກິຣິຍາ	ຈັກກະພັດນິຍົມ	ກີ້ງຕາໃສ່	ກອງສອດແນມ
ກີ້ເຕາ	ກະເຕາະກະແຕະ	ກຸຫລາບ	ເຄື່ອງວັດຄວາມໄວ
ກີ້ງຕາໃສ່	ກອງກິ້ນ	ກູລີ	ເຄື່ອງວັດຄວາມຮ້ອນເ
ກ້າກິ້ນ	ກົວ	ເກີດຈາກ	ຍັນ
ກ້າ	ກວຍ	ເກີດໄພ	ງານຂຶ້ນເຮືອນໃຫມ່
ກູລີ	ກ້ວຍມືນາງ	ເກີດມາ	ເງິນຮາງ
ກຸຫລາບ	ໃກ້ຊິດກັນ	ເກືອບຫມົດ	ເງິນແຮ່ຖົງ
ເກັດປາ	ກະແຕະ	ເກັດປາ	ຈັກກະພັດນິຍົມ
ສຽງແທບ	ໂກປານໃດ	ເກັບ	ຈົດທະບຽນ
ເກັບ	ເກົາ	ເກັບກຸ່ວ	ຈົດທະບຽນການຄ້າ
ກະຕື່ລີລົ້ນ	ກອງໂຈນ	ເກັບກຸ່ວ	ສີກຸຫລາບ
ເກັບກຸ່ວ	ຊັກອອກພູດນຶ່ງ	ເກັບພາສີ	ສີແກ່
ແກ້ວນາເຝານເນື້ອແ	ກົວ	ເກົາ	ສຽງທອງ
ຂງ	ສີແກ່	ເກົາ	ສຽງບົກກະຕິ
ກົກ	ເຄື່ອງວັດຄວາມຮ້ອນເ	ແກ້ວໂກເມນ	ສຽງຟ້າຮ້ອງ
ເກັບກຸ່ວ	ຍັນ	ແກ້ວນາເຝານເນື້ອແ	ສຽງມ້າແທມ
ເກັບພາສີ	ກະແຕະ	ຂງ	ສຽງແທບ
ກະຕື່ລີລົ້ນ	ສຽງບົກກະຕິ	ກໍຈິງຢູ່ແລ້ວ	ຊັກອອກພູດນຶ່ງ
ກົກແຂນ		ໄກເທົ່າໃດ	

Conclusion

Lao presents one of the most challenging scenarios for collation. First, Lao text does not have spaces so processing is required to segment text into words (not discussed in detail in this chapter; much work has been done on this for Thai, e.g. see [23, 24, 25]). Once the word sequence is available, words are required to be syllabified and individual characters need to be tagged for different roles depending on the context (details of this process are discussed in [20]). Then re-ordering and normalization need to be done. Finally, depending on collation strategy, which could be based on language or script, collation elements need to be assigned. Within the syllable, Lao sorts at four levels, with nuclear consonants getting the primary weight, vowels getting the secondary weight, non-nuclear consonants getting the tertiary weight and tone marks getting the quaternary level weight. The sort keys generated are also at syllable level (and not at word level). Thus, the Unicode collation algorithm [2] discussed in the second chapter needs to be modified to do a sequence of comparisons of sort keys generated by syllables from words. Though the current work has been tested, much more work needs to be done in this area. Standards also need to be defined by relevant organizations.