

# Lao Collations

By:

Phonpasit PHISSAMAY

Nadir DURRANI

Lao Localization Project Team



# For Nuclear consonants

## A. All ligature characters followed after all single consonant:

ກ(1); ຂ(2); ຄ(3); ງ(4); ຈ(5); ສ(6); ຊ(7); ຍ(8); ດ(9); ຕ(10); ຖ(11); ທ(12);  
ນ(13); ບ(14); ປ(15); ຜ(16); ຝ(17); ພ(18); ຟ(19); ມ(20); ຢ(21); ຣ(22); ລ(23);  
ວ(24); ຫ(25); ອ(26); ຮ(27); ຫງ(28); ຫຍ(29); ຫນ(30); ຫມ(31); ຫມ(32);  
ໝ(33); ຫລ(34); ຫຼ(35); ຫວ(36).

## B. All ligature characters allocated in the group of its first combination consonant “ຫ(25)” :

ກ(1); ຂ(2); ຄ(3); ງ(4); ຈ(5); ສ(6); ຊ(7); ຍ(8); ດ(9); ຕ(10); ຖ(11); ທ(12);  
ນ(13); ບ(14); ປ(15); ຜ(16); ຝ(17); ພ(18); ຟ(19); ມ(20); ຢ(21); ຣ(22); ລ(23);  
ວ(24); ຫ(25); ຫງ(26); ຫຍ(27); ຫນ(28); ຫມ(29); ຫລ(30); ຫຼ(31); ຫວ(32);  
ອ(33); ຮ(34); ຫມ(35); ຫມ(36);

## C. All ligature characters allocated in the group of its nuclear consonant:

ກ(1); ຂ(2); ຄ(3); ງ(4); ຫງ(5); ຈ(6); ສ(7); ຊ(8); ຍ(9); ຫຍ(10); ດ(11); ຕ(12); ຖ(13);  
ທ(14); ນ(15); ຫນ(16); ຫມ(17); ບ(18); ປ(19); ຜ(20); ຝ(21); ພ(22); ຟ(23); ມ(24);  
ຫມ(25); ຫມ(26); ຢ(27); ຣ(28); ລ(29); ຫລ(30); ຫຼ(31); ວ(32); ຫວ(33); ຫ(34); ອ(35);  
ຮ(36).

# For vowels

**A. All combination vowels followed the all single vowels:**

ຂະ(1); ຂາ(2); ິ(3); ື(4); ື(5); ື(6); ູ(7); ູ(8); ເຂະ(9); ເຂ(10); ແຂະ(11);  
ແຂ(12); ໂຂະ(13); ໂຂ(14); ເຂາະ(15); ະ(16); ື(17); ື(18); ື້(19); ເຂງ(20);  
ືວະ(21); ືວ(22); ືອ(23); ືອ(24); ໂຂ(25); ໃຂ(26); ະ(27); ືາ(28)

**B. The combination allocated into the group of its first component vowels:**

ຂະ(1); ຂາ(2); ິ(3); ື(4); ູ(5); ູ(6); ເຂ(7); ເຂະ(8); ື(9); ື(10); ືອ(11);  
ືອ(12); ເຂາະ(13); ືາ(14); ື້(15); ື້(16); ເຂງ(17); ແຂະ(18); ແຂ(19);  
ແ້(20); ໂຂ(21); ໂຂວ(22); ໂຂະ(23); ໂ້ກ(24); ະ(25); ືວ(26); ືວະ(27); ື(28);  
ືວ(29); ະ(30); ຂວ(31); ຂອ(32); ຂງ(33)

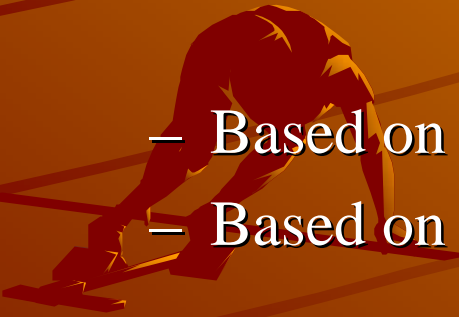
# Lao Collation

- ◆ Lack of National Standards

- ◆ Looking at text books and prevailing lexicons we find two most commonly used approaches.

- Based on Combinational Consonants and Vowels

- Based on Alphabetical Order



# What is Collation ?

- ✦ A term generally taken as a process of sorting out the set of strings based on lexicographic order.
- ✦ For example sorting order of English characters based on their ASCII code is “ABCDE...Z abcde...z” which is different then its lexicographic order which is “AaBbCcDdEd...Zz”.
- ✦ Collation is something, driven on base of cultural and national standards.

# Possible Solutions

## ✦ Based on Lexicon

- Giving weight to each word. So collation involves primarily searching the word in lexicon and comparing the weights to sort them.

## ✦ Discrepancies of this approach

- Brute Force Computing
- Consumes Space
- Consumes Time




# Contd..

## ✦ Based on Syllabification

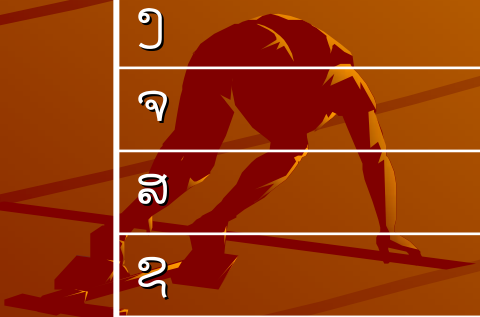
- Before we can proceed we need to segment the word into syllables.
- Many Lao characters exhibit hybrid behavior. With each identity they act different when doing collation. We need to know what role each character is playing within the syllable.
- To further aggravate the situation a Lao character needs to be reordered once it is syllabified because the collation is based on the syllable structure and not the key press order.

# Four Levels of Lao Collation

- ◆ Nuclear and Combining Consonants + Digits
  - ◆ Vowels
  - ◆ Consonantal
  - ◆ Tone Marks
  - ◆ Punctuation Marks and several Lao characters are left ignorable at all levels.
- 

# Combinational Consonants and Vowels

## ◆ Consonants (Approach A)



ກ	ຖ	ຢ	ໝ
ຂ	ທ	ຮ	ຫມ
ຄ	ນ	ຼ	ຫຼ
ງ	ບ	ລ	ຫລ
ຈ	ປ	ວ	ຫວ
ສ	ຜ	ຫ	ອ
ຊ	ຝ	ຫງ	ຮ
ປ	ພ	ຫຍ	
ດ	ຟ	ໜ	
ຕ	ມ	ຫນ	

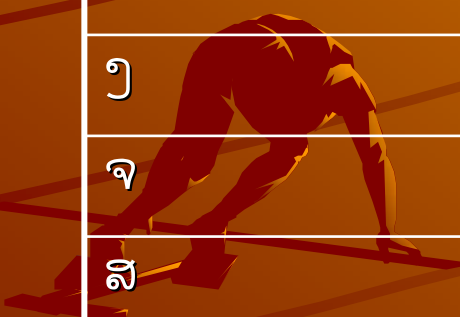
# Contd..

## ◆ Vowels (Approach A)

ເຂ	ເັ້+X8/X9	Xອ+X8/X9	ເື້ອ
ເັ້+X8/X9	ເX	ເື້	Xວ+X8/X9
າ	ແXະ	ເື້	ໄX
ົວ	ເັ້+X8/X9	ເັ້ງ	ໃX
ົວ	ແX	ເXງ	ເື້າ
ົວ	ໄXະ	ງ+X8/X9	ໍາ
ົວ	ົ	ົວະ	ໍ່+າ
ຸ	ໄX	ັວ+X8/X9	
ູ	ເXາະ	ົວ	
ເXະ	ໍ	ເື້ອ	

# By Alphabetical Order

## ◆ Consonants (Approach B)



ກ	ຄ	ຝ	ຫ
ຂ	ຕ	ພ	ຫຼ
ຄ	ຖ	ຟ	ອ
ງ	ທ	ມ	ຮ
ຈ	ນ	ຢ	ໝ
ຊ	ບ	ຮ	ໝ
ຮ	ປ	ລ	
ເ	ຜ	ວ	

# Contd..

## ◆ Vowels (Approach B)

ॐ	ॐ	ॐ	ॐ
ॐ	ॐ	ॐ	ॐ
ॐ	ॐ	ॐ	ॐ
ॐ	ॐ	ॐ	ॐ
ॐ	ॐ	ॐ	ॐ

# Contd..

## ◆ Consonantal

ກ	ຍ	ຟ
ງ	ດ	ມ
ຈ	ນ	ລ
ສ	ບ	ວ
ຊ	ພ	

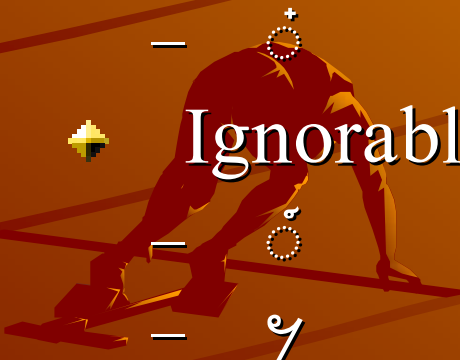
# Contd..

## ◆ Tone Marks

- ̀
- ́
- ̂
- ̃
- ̄

## ◆ Ignorable Level

- ̅
- ̆
- ̇



# Contd..

## ◆ Digits

၀	၂	၆	၈
၁	၃	၇	
၂	၄	၉	



# Lao Collation Algorithm

- ◆ Based on Unicode Collation Algorithm

- ◆ Steps Involved

- Segmentation: Finding out syllable boundaries and getting information what role each character plays in the syllable. Some characters can act hybrid. For example ‘໊’ can act as nuclear consonant, combining consonant, vowel and consonantal.

- ◆ Example: For ກາງຫລາວ we find out that it has two syllables ‘ກາງ’ and ຫລາວ. We also find what role is each character playing within each syllable. For example ‘ກ’ is nuclear consonant, ‘າ’ and ‘ງ’ is consonantal in ‘ກາງ’.

# Contd..

- Reordering: This is different than reordering done during syllabification. This involves reordering into collation levels i.e. the syllable is reordered as ‘X’ ‘X1’ ‘X2’ (Consonants), ‘X0’ ‘X1’ ‘X3’ ‘X4’ ‘X6’ ‘X7’ (Vowels), ‘X8’ ‘X9’ (Consonantal), ‘X5’ (Tone Marks) and finally ‘X10’ (Special characters).

For example ‘ຫລາວ’ is reordered as ‘ລຫາວ’ and ເນັ້ນ is reordered as ນ ເ ັ ນ ັ .

# Contd..

– Form Collation Element Array: This is where two approaches are different.

✦ With Combinational Approach we have to look at the neighbors to decide what weight we should give to a consonant or a vowel.

✦ For example ‘ㄹ’ can attain different weights as it plays combinational politics differently in cases  
‘ㄹXㄷ’, ‘ㄹX8/9’, ‘ㄹX’,  
‘ㄹXㄷ’, ‘ㄹX’, ‘ㄹX’, ‘ㄹX’, ‘ㄹX’, ‘ㄹX’, ‘ㄹX’ .

✦ With alphabetical order there is only one possible collation element for each character.

# Contd..

- ◆ Example (forming collation element array) ກາງຫລາວ

Combinational Approach		Alphabetical Approach	
ກ	[0820.0200.0020.0002]	ກ	[0820.0200.0020.0002]
າ	[0000.020F.0020.0002]	າ	[0000.020A.0020.0002]
ງ	[0000.0000.002A.0002]	ງ	[0000.0000.002A.0002]
ຫລ	[08B1.0200.0020.0002]	ລ	[088E.0200.0020.0002]
າ	[0000.020F.0020.0002]	ຫ	[0898.0200.0020.0002]
ວ	[0000.0000.0066.0002]	າ	[0000.020A.0020.0002]
		ວ	[0000.0000.0066.0002]

# Contd..

- Form Sort Key: Sort key is formed by successively appending weights from the collation element array picking primary, secondary, tertiary and then final level one by one. Zero is placed only as level separator.

✦ For instance sort key for: ‘ນອນ’ is

[085c 0000 0200 0269 0000 0020 0020 0048 0000  
0002 0002 0002]

✦ ‘ເນ’ is [085c 0000 0200 023C 0000 0020 0020  
0000 0002 0002]



# Contd..

- Compare the Sort Key: Last step involves comparison of the sort keys. We start comparing from the first element and continue only till difference is found or till the end of sort key.



◆ Level 4 differences are ignored if there are any Level 1,2 or 3 differences

◆ Level 3 differences are ignored if there are any Level 1 or 2 differences

◆ Level 2 differences are ignored if there are any Level 1 differences

◆ Level 1 differences are never ignored.

# Contd..

✦ For instance when comparing

ນອນ' [085c 0000 0200 0269 0000 0020 0020 0048 0000  
0002 0002 0002]

‘ເນ’ [085c 0000 0200 023C 0000 0020 0020 0000 0002  
0002]

we find out that difference is at secondary i.e. at vowel level.

Thank you

