Lao Character Set

Phonpasit Phissamay and Nadir Durrani
Science Technology and Environment Agency, Center for Research in Urdu Language Processing
phonpasit@stea.gov.la, nadirdurrani@yahoo.com

Abstract

The objective of this paper is to discuss the characteristics of Lao script and Lao character set. Some popularly used previous encodings are discussed and are compared with existing Unicode standard.

1. Definition

Character is a smallest component of written language that has semantic value; refers to the abstract meaning and/or shape, rather than a specific shape, though in code tables some form of visual representation is essential for the reader's understanding.[1]

Like all other forms of information stored on computer, text is stored as a sequence of binary numbers. The computing "map" for particular numbers corresponding to particular characters in a common accepted set is usually referred to as the Character Set or the Code Page. The Character set is an ordered enumeration of characters (alpha/numeric and symbolic) used for keyboard mapping, textual information display, and the exchange of textual information in digital format. Without code pages, the input/output/storage and transport of textual information using computer or other digital devices would not be possible.[2]

ASCII (American Standard Code for Information Interchange) defined in 1968 by American National Standards Institute (ANSI) only contained English characters. This table was extended to 8 bits to include other Latin based languages spoken in Europe and South America but it did not include Lao and other Asian languages. Consequently to help themselves they exploit the ASCII encoding just replacing roman characters by Lao characters and using ASCII code table.

Unicode is the universal character encoding, maintained by the Unicode Consortium. This encoding standard provides the basis for processing, storage and interchange of text data in any language in all modern software and information technology protocols. The Unicode Standard is a single, universal character encoding. All characters are equally accessible, and the blocks have no implementation expression in most Unicode software. Unicode characters may be encoded at any code point from U+0000 to U+10FFFF. Unicode characters are expressed with 16-bit code units.[3]

There are four levels of the Unicode Character Encoding Model:

- Abstract Character Repertoire. This means the set of characters to be exactly encoded, such as some alphabet or symbol set
- Coded Character Set. This is the mapping from an abstract character repertoire to become a set of nonnegative integers
- Character Encoding Form. This is a mapping from a set of nonnegative integers that are the particular elements of a CCS to become the particular code units' sequence. Furthermore, some characters themselves have a specified width, such as 32-bit integers
- Character Encoding Scheme. This is a reversible alteration from a set of code units' sequence (altered from one or more CEFS to a serialized sequence of bytes).

[1] www.cicc.or.jp/english/hyoujyunka/mlit4/7-5LAOPDR

[2] www.undplao.org/unv%20site/Reports

[3] www.unicode.org/unicode/faq/basic_q.html

2. Lao script

Lao characters originate from two sources Pali and Sanskrit and written Buddhist prayer from Tong Pali. Lao alphabets have their roots with Sanskrit currently used as communicative language and/or Buddhist language from the Pali alphabet used in written Buddhist prayers. The Lao language is composed of 78 characters and signs: 33 Consonants, 28 Vowels, 04 Tone marks, 03 Special symbols and 10 Lao digits.

2.1. Consonants

The Lao alphabet has 33 consonant letters including 27 single consonants and 6 mixed consonants.

- Single consonants: 27 single consonants are respectively

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

- Six mixed consonants are as follows:

ຫ+ງ = ຫງ
ຫ+ຍ = ຫຍ
ຫ+ລ = ຫລ or ຫຼ
ຫ+ວ = ຫວ
ຫ+ນ= ໜ
ຫ+ມ= ໝ

2.2. Vowels

The Lao alphabet has 28 vowels, they are:

ະ າ ◌ິ ◌ີ ◌ຶ ◌ື ◌ຸ ◌ູ ເະ ແະ ແ ໂະ ໂ ເາະ ◌ໍ ເ◌ິ ເ◌ີ ເ◌ັຽ ເຽ ◌ົວະ ◌ົວ ເ◌ືອ ໄ ໃ ຳ ເ◌ົາ

2.3. Tone Mark

Tonal marks are another group of alphabet letters which are symbols used to characterize the change of the sounds of a syllable to follow the rule of consonants sounds with short, medium, low, high tone. 4 tonal marks are characterized with the followings symbol:

◌່◌້◌໊◌໋

2.4. Symbols or signs

Some symbols used in the Lao language often have a certain meaning and characteristic similarly to those which are used in the English language. But some may have different meaning such as: Indicating repetition of syllable ໆ Indicating and others (etc.) ຯ Indicating long sound of syllable

2.5. Lao digits

໐ ໑ ໒ ໓ ໔ ໕ ໖ ໗ ໘ ໙

3. Lao character Set (Lao Code Table)

The character set are consonant characters, vowel characters, tone marks, Lao digits and special symbols. It contains many combined glyph characters. The mixing with alphanumeric characters can be handled only by switching the font list. Mixing is not possible on the internal code level. Therefore the number of Lao characters existed into any code table are less then defined in the Lao language grammar, because there is not allowed to create a single glyphs (shape) for mixed consonant and combination vowels. Therefore most of Lao code table consists of 65 character sets which are: 27 Single consonants; 3 mixed consonants; 18 Single vowels; 4 Tone marks; 3 Special symbols and 10 Lao digits.
There are several coding conventions which have allowed Lao language to be used both on DOS and windows- based computers

Such as Lao Script for Windows, IBM-8-1033, Alice0, Hongkard, etc. The most intractable problem with these code tables is that they might encode different character sets, leading to characters being in a different position on the keyboard. A document written in the Lao language on one computer, saved on floppy disc and then opened on another computer might appear to be nonsense because the internal coding of the characters of the second computer might be incompatible with that of the first computer.

Most of Lao character sets (Code table) can be divided into 3 categories:

3.1. Upper ASCII structure

It was simple technology that employed replacement of English glyphs by Lao character glyph and using codes for English characters to write Lao text. The code started from 65 to onwards. The problem with this approach was if the user wanted to mix Lao-English text he had to face problems

switching font. Moreover there was not enough room for possible permutations and combinations of Lao consonants as they combined with tone marks and vowels.

3.2. Lower ASCII structure

To solve the problem of mixed Lao-English text they used another mapping scheme not to use the numerals and English ASCII codes but to make use of the higher ASCII characters the 8th bit that was added to facilitate European and South American characters. In this fashion the user would be able to write English-Lao text with same font without having trouble to repeatedly change the font.

The Lao characters were added into the ASCII code followed by the roman character. Usually character is started from the range of 128. However it is require add-on software or some programming application in order to executed all those characters. Many developers created the new shape (Character code) for solving the issue of variety of tone mark position. Therefore number of Lao characters existed in this type of code table are bigger then usual.

But there are still problems with this approach. With each updated version of an operating system there has to be an update and maintenance of fonts because position of new shape may be interrupted with some function in new operating system or application operating in the prohibited zone. And then Lao fonts are unworkable on the systems which can effectively use the Thai language and also unworkable on Japanese, Chinese and Korean versions of windows.

3.3. Lao Unicode structure

In 1993 the International Standard Organization created the International standard of Universal Multiple Octet Coded Character Set for information technology known as ISO-10646 and it has transformed into the Unicode. It's fortunately or unfortunately the standard of Lao character sets has been existed with this code table by unknown proposal. This code table is composed of 65 symbols with almost existed in Lao language: 27 Single consonants, 3 mixed consonants, 18 Single vowels, 4 Tone marks, 3 Special symbols and 10 Lao digits.

The specification of the Unicode is that the Lao alphabet has been particularity assigned to the code range 0x0E80 to 0x0EFF (hexadecimal). No other scripts are able to utilize the code values in this range. Punctuation and Arabic numerals (typically used with Lao) have their normal (ANSI) code values in the range 32 to 255. With the Lao Character Unicode range, the characters themselves are mostly based on the Thai TIS (620-2533:1990) standard. Although this standard enables a full, correct representation of the language, it is not ideal in terms of Lao character ordering.

Fonts that use "pure" Unicode must compromise typographic quality unless additional positioning information is provided in the font and used at the operating system level. An alternative is to include typographic enhancements of Lao (contextual variants and composite glyphs) in the Private Use Area (PUA) range of Unicode (0xE000 to 0xF8FF), which will not conflict with other languages, but may limit compatibility between applications and across platforms.[4]

3.3.1. Problem with the current Lao encoding in Unicode

a. The whole structure based on Thai is not so suitable to Lao. The Code table is very wasteful because of so many sporadic vacant slots. The order of characters is not following the Lao ways. A clear example is the position of character ฑ, it should be placed between character and ล =m but it is actually placed before ร. This makes the result of simple binary sorting incorrect.

b. The wrong naming of "ฝ" and "ฟ"The current block defined as follows:

U+0E9C is ຜ Lao single character PHO SUNG
U+0E9D is ຝ Lao single character FO TAM
U+0E9E is ພ Lao single character PHO TAM
U+0E9D is ຟ Lao single character FO SUNG

SUNG means "high" and TAM means "low". This classification, which is common in linguistics, intends grouping of characters by their tonal characteristics and does not match the actual tone. Based on this, ຝ should have been named Lao LETTER FO SUNG, because its tonal property is the same as that of ຜ and other "high" consonants, while ຟ should have been named LAO LETTER FO TAM.

[4] www.laoscript.net/support (Font: Lao Language and Unicode)

c. Possible wrong naming of ຣ and ລ: The current Lao block defined as follows:

   U+0EA3 is ລ Lao letter LO LING, but originally is called Lao Letter RO

   U+0EA5 is ຣ Lao letter LO LOOT, but originally is called Lao Letter LO

d. Possible improper naming of: ◌ຼ The current Lao block defined as follows:

   U+0EBC is ◌ຼ Lao SEMIVOWEL SIGN LO, but originally is Lao Combination Consonant LO

e. Possible improper naming and annotation of: ◌ໍ The current Lao block defined as follows:

   U+0ECD is ◌ໍ Lao NIGGAHITA (final nasal), but originally is Lao vowel sign OR

However, it seem that the said above problems will not causing any serious technical problem of using Lao language based on Unicode, since all of Lao single character are existed in Unicode. And the programmer and software are referring to the code table, not to the name of characters.

4. The proposal for the standard of Lao character set

Based From several official meetings in which the issue of Lao character sets was seriously discussed, it was decided that using Unicode is the only solution to this difficult issue for the following reasons:

- Unicode is considered a universal coding standard and being increasingly employed by the respective software developers; Apple Macintosh, Windows and UNIX (Linux).

- All Lao single characters are supported by Unicode, enabling the writing of any word in the Lao language. A sorting mechanism is required because Lao language is syllabic, not alphabetic. However, there will be sporadic unused slots which may affect the sorting order by the code table but generally this is not considered to be an issue causing much concern.

- Local development: All new growing developers in this area are intending to utilize Unicode. The new version of LSWIN supports Unicode.

- The Unicode and ISO 10646 has clearly stated that no native character sets, which already exists in this code table can be re-moved, changed or rename. This means that after a character set has been defined, there cannot be any changes done. Therefore, if a different code table is proposed, then ISO and Unicode will never accepted the new proposal of standard code table of Lao Character sets.[5]

- Market trends: most new operating systems and software are based on Unicode. The system will generally facilitate all kinds of information processing of the native language with Unicode

- The users: The user can write in Lao font without any installation or downloading each time they use their computer. This convenience and universal adaptation encourages standardization.

The following functions should be implemented in order to completely support Lao Unicode at the operating system level:

- Lao Unicode keyboard layout.

- Implementation of typographically-enhanced Lao fonts

- Convert data from ASCII to Unicode characters
- Be able to automatically break the lines according to the fixed boundary of each line (word wrap)

- Sorting words correctly by the Lao grammatical order

- Utilize Lao “locale” information (Identifiers of language, numeric symbols, etc.)

[5] Submitted to International University of Japan on 2006-05-19

Reference and appreciations:

- The Asian Character sets, by Mr. TAKAYUKI SATO, Center of the International Cooperation for Computerization, JAPAN
- Issues related with the existing international standard encoding of Lao script, by Prof HARADA SHIRO, University of Tokyo

Lao script and Unicode, by Dr. Jonh M. Durdin, Lao Script for Window software.