

LAO NEW CODING (LNC)

H. Thongvilu (c) 2003-01-01

ABSTRACT

LNC is a complete new digital encoding of Lao language. A lack of native Lao script handling technology had bred a number of solutions for Lao language computerization, however all are implementation of similar techniques. LNC is designed to address to some of the issues of current technology deficiencies, from a new perspective. Currently, keystrokes sequence or bytes stream of Lao characters determine the correct positioning of these elements. Failing to follow the right sequence, even though a correct display may be achieved, it could complicate lexical search and data matching. LNC offers a new way of encoding Lao words with precise symbol elements' position information, in less number of bytes. By efficiently using less byte for data representation and storage, LNC is also a compression scheme and technique.

FOREWORD

As Lao language computerization development is out of mainstream research due to market size commercial interest, we make special dedication to the many interest groups and individuals who had given their time and efforts in bringing Lao language to the digital world.

1. INTRODUCTION

Lao New Coding (LNC) is a complete new digital encoding of Lao language. It is designed to address to some of the issues of current technology deficiencies. The research stemmed from my work on computerization of Lao language over the past ten years. The idea was conceived at least five years ago but I was never inspired to publish it. It was stirred by frustration of a lack of an effective way of digital encoding of Lao writing. The situation is not help by the small market for Lao language support on computers. However despite of the language low profile, it is a challenge to special small interest groups that had soldier on to produce some workable solutions.

For the least of its worth, it is another mental exercise.

2. TECHNOLOGY

2.1 Lao writing

Lao writing is derived from Indian Sanskrit script and many words are borrowed to build Lao dictionary, as well as some imported foreign words. Lao word could be of a single or multiple syllables. The syllable is formed by: [consonant/cluster consonants][vowel/derived vowel] [consonant modifier (optional)][tone (optional)]; this is being referred to as a *word-clusters* in this paper.

It is interesting to note that Lao grammar in classical term as a rule of speech and writing; that Lao grammar only defines word constructions and little of inter-linkages of words.

Multi-words words are built by joining a number of words; Lao words do not have extensive word identifier¹ to explicitly define words relationship in the multi-words construction. *Lao grammar* is referred to in this text, as word-clusters construction rule.

Lao vowels compose of multiple sub-elements; each element occupies a predefined position within the word-cluster, some elements proceed and displace consonant when invoke. To date such symbols compounding is not supported within any operating system, and the lack of standards had bred many incoherent implementations that had confused the computerization of Lao language.

Traditionally Lao text is not punctuated, other than being organized in paragraphs of *un-segmented text*².

2.2 Characters encoding

With the advent of PC and MS Windows with its code-page³solution, it has boosted foreign languages support in applications. However with only 128 characters available at any one time and requiring prior knowledge of the language and font used, it only provides a limited solution.

Unicode is emerging as the standard capable of supporting all known languages all at once, making a

¹ Noun, verb, adverb, pronoun, adjective

² Long continuous stream of Lao characters

³ Code-page reserves the top 128 characters of a block of 256 for non-English characters and symbols.

full multi-lingual document possible. Unicode is yet to prove itself in applications, it is a heavier burden on computer resource by using 16 bits word to represent a character, a trade off, compare with 8 bits (ISO8) as used in a code-page system.

Most Unix systems today are also based on similar encoding method.

2.3 Current technology of Lao language computerization.

Lao language support, due to a lack of standards and lack of native Lao script handling within existing operating systems, the industry had developed many incoherent solutions, mainly on the Microsoft Windows operating system platform.

A full foreign language system at operating system command level is still far off. It is even questioned of its necessity if foreign languages are fully supported at application level. Such full foreign language support is possible as developed for Unix operating system under the Internationalisation (I18N) localization scheme where system commands are encoded and mapped to foreign words and messages, thus all applications that are compiled with I18N awareness would be full internationalised.

There are two broad main issues in computerizing foreign languages:

- 1) Display, fonts.
- 2) Input methodology.

They both deal with data representation; how the information is seen, retrieved and reproduced on screen, printer or transmitted. How input data is interpreted and stored. Some uniformity and consistency is required so that foreign languages would work across applications and platforms.

In current technology, Lao words require some five bytes on average. Keystrokes or bytes stream of Lao words require a correct order of sequence to correctly position the elements within the word-cluster. Possible solution is to assign extra bytes that would preserve elements' attributes and position information. The recent font standard, OpenType⁴ developed by Adobe and Microsoft, combines TrueType and Postscript technologies to provide new typographic features such as the desirable capability of elements compounding, promises a possible solution to Lao script problems.

Reliance on keystrokes sequence alone would complicate lexical search and matching of words.

⁴ : <http://www.adobe.com/type/opentype/main.html>

3. KNOWN ISSUES

3.1. Display, fonts

Due to a lack of standards in Lao language computerization and limitation of current technology, namely in how to handle compounding symbols, it had created many diverging implementations; such as multiple pre compounded symbols and non-consistence byte codes assignment to character symbols that had affected the digital information processing and exchange. All these confusions are of course resolvable with an industry accepted workable set of conventions and standards.

The quality of fonts and typefaces had reached a satisfactory level, and font technology has also come of age with Unicode that would standardise the byte(s) coding of all known languages. Combined with a new OpenType that would allow symbols compounding, the emerging technology promises a more coherent future as far as foreign languages are concerned.

3.2. Input methodology

Inevitably any data coding comes down to some interaction with human input such as from keystrokes, scanned text, voice or whatever. The problem of input is inherited with the deficiency of operating systems to handle foreign native scripting. Smart keyboard drivers, macros, rule based or database lookup of words construction, could create a pre-emptive input process. It is possible with current technology but the solution needs to be interfaced and integrated into the operating system and applications, as addressed to by I18N project on Unix platform.

4. COMPATABILITY

A question that may be ask is will LNC compromise other character encoding scheme such as ISO8859 and Unicode and what are the Font issues.

4.1. Other character encoding scheme such as ISO8859 and Unicode.

LNC does not define any character coding see section 8. "INTERPRETATION OF LNC DATA". The LNC engine maps and remaps LNC internal hashed data to existing character codes.

4.2. Font issues.

LNC only relates to fonts particulars only by maps of internal LNC bit patterns to character codes and associate fonts. The relevancy issue of font comes in with hidden, zero width and control character codes that are used to format the visual output. It is merely for the purpose of cross referencing (table lookup) when packing and unpacking LNC data to match the original data stream.

5. LAO GRAMMAR

5.1. Word construction

The basic of Lao grammar is very simple each single syllable word is constructed with some well define rules.

- 1) Consonants
Any single consonant character from the alphabets.
- 2) Cluster consonant
Dual consonants⁵, although in Lao only a combination of:
 - a) [ກຂຄງຈສຊຕຖ] + ວ
 - b) ຫ + [ຍລມນ] are used, these change form to: ຫຼ ຫູ ຫ໊ ຫ໋
- 3) Consonant qualifier/modifier [optional]
ກ ງ ຍ ດ ນ ບ ມ ວ ອ
- 4) Vowels
Vowel elements surround the consonant cell, occupying cells as shown in Table 5.1.
 - i. Pure vowels
 - ii. Derived sounds
Pure sounds could be qualified by a subset of consonants: ກ ງ ຍ ດ ນ ບ ມ ວ ອ, some change forms in the process, e.g. ຂ + ກ = ຂັກ
- 5) Intonation [optional]

Lao syllable falls within word-cluster boundary; all symbols occupying a well define cells.

	2/4		
2	1	2/3	3
	2		

Table. 5.1

Word-Cluster cells

Baseline

⁵ A combination of two consonants from a set of 27 would cater for 2²⁷ possible foreign sounds, although some may be hard to articulate.

5.1. Multi-words word formation

As Lao single words fall within a syllable boundary, multi-words are multi-syllables words.

Lao text is written from left to right in a continuous stream of characters, traditionally with no punctuation in paragraphs. Joining single or other multiple-words together to form new words.

As mentioned in section '2.1 Lao writing' Lao words do not have extensive word identifier⁶ to explicitly define words relationship in the multi-words construction.

Such words relationship and dependency defines sentence formation and extraction from the continuous stream of characters in a Lao text. Lao reading or utterance also follows these word relationships.

6. LNC OBJECTIVES

Address to and develop rule base symbols compounding mechanism for Lao language, thus establish a concise and uniform method of compiling Lao digital information.

6.1. SCOPE OF LNC

LNC is herein described how single word syllable words are encoded, to be successively replicate in multiple-words words.

6.2. OUT OF SCOPE OF LNC

LNC does not deal with how a byte stream is tagged as Lao text, nor does it resolve LNC text in a multi-lingual document.

⁶ Noun, verb, adverb, pronoun, adjective

7. LNC DESIGN

Analysing the use of Lao character symbols the design of LNC was developed, as an expansion of Table 5.1.

Note, that the cells represent only symbols position, the visual rendering and output is a function of the device driver to place the symbol correctly. LNC is only the data encoding part of a Lao word, data representation would still relies on font capability to do symbols compounding or substitution with pre-compounded symbols.

7.1. Cells description

- Most symbols are non-composite (unique symbol), with a single exception in cell 5 (ຊຍ).
- Cells 1,2,3,5,6,7 of Table 7.1 are bits assigned to represent their respective (Table 7.2.1) symbol occupying the cell
- Cell 4c represents descendant or under script of a consonant e.g. ທຸ ທຸ similar to cell 5 ຍ/ຽ consonant modifiers change.
- Cell 4a & 4b requires special dealing, see section 7.2 "Consonants and Cluster-consonants".

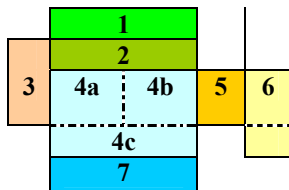


Table 7.1

7.2. Consonants and Cluster-consonants

Cells 4a and 4b are to accommodate consonants.

- In a *single consonant* word only 3 bytes are required for simple word construction. The bits are layout as in table 7.2.1
- When a dual or cluster-consonants are used, it is flagged by bit 23rd of the 3rd byte. Bit 24-28 of an additional 4th byte is required to represent the 2nd consonant of the pair.
- In cases where a cell is empty, cell 1 shift down to cell 2, 6 to cell 5 positions.

15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
6								7		5		4a			

31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	
4b								23	2				1			

Table 7.2.1

Elements in cell cluster (cell/elements/bits req.)					
1/4/3	Tones (4) ໊ ໋ ໌ ໋	2/7/3	໊ ໋ ໌ ໋ ໊ ໋ ໌ ໋	3/5/3	໊ ໋ ໌ ໋
4/27/5 section 7.2	27 consonants & cluster consonants	5/3/2	ຊ ຍ ຊຍ	6/9/4	ກ ງ ຍ/ຽ ດ ນ ບ ມ ວ ອ
					7/2/2

Table 7.2

8. INTERPRETATION OF LNC DATA

It has to be stressed that LNC is only *data representation*. It is how a number of bytes are used to represent some information.

Where as font encoding of a character set, is how data byte(s) represent a particular symbol. In non-compressed or hashed up data, the bytes stream are represented by the character encoding bytes and thereby readily examinable.

LNC data stream is considered as hashed up data, it therefore is not examinable with the data being interpreted by some utility ware of the its hashing.

It is anticipate the LNC can be built as an engine within application that supports it. Typically the de/compression process will happen every time data is fetch or store, transparently.

8.2. Storing/writing LNC data

Once data is fetch from an input device, probably in character set code bytes, it is analysed and using a lookup translation table, convert the code bytes to a bit pattern of each cell (ref. Table 7.1 and 7.2). The resulting hashed data is stored.

It is worth noting that using LNC, it helps resolve the necessity of keystrokes order (ref section 2.3 Current technology of Lao language computerization) that defines the lexical of the word.

8.2.1. Phonetic input

By pre defining a sound, such as a vowel, in LNC bytes composing all required elements, it is possible to write a Lao word with 2 keystrokes: [Consonant][LNC Vowel], the LNC engine would translate the keystrokes into a stream of LNC bytes; paving a way for a phonetic input, by keyboard or even by voice by matching the two classes of sound.

8.3. Fetching, unpacking LNC data
Step 1. Make sure the data is a hashed up LNC data block. This can be done by some signature stored within the data block, as supported by the application.

Step 2. Three bytes are retrieve from data storage or direct from input device:

8.3.1. Check the top of the third byte bit 23. If it is not set 3rd step follows.

8.3.2. If bit 23 is set, an extra byte is fetch that represents the second consonant for the cluster-consonant pair.

These are the maximum bytes required to represent a Lao word syllable.

Steps 1 & 2, would already achieve *significant data stream search*, compare with traditional character code base data block of un-define numbers of bytes, requiring data search engine to do string searches and testings for syllable word boundary.

Thirdly, the bytes are broken up at bits boundary according to table 7.2.1. Each cell bit value is translated to some character code matching the original data input, and stream to the output device.

As mentioned above in section 7.2 "Storing/writing LNC data", since LNC guarantees the correct lexical of a word, LNC remapping would consistently repack the data into the right character code order to produce the correct compounding of symbols.

This is a significant problem for lexical searches using the character code system, a different order of keystrokes could produce the right visual effect, but the word is different lexically.

9. CONCLUSION

Lao New Coding (LNC) is a new concept of encoding of data in the computerization of Lao Language. It addresses primarily in preserving character symbols positions and lexical encoding of syllable words. It is an exercise of the mind, where it will end depends on how adaptable it is within existing applications and drivers.