PAN Localization Project
Science Technology and Environment Agency of Lao PDR
National University Computer and Emergency Science of Pakistan
International Development Research Center

Syllabification of Lao Script for Line Breaking

Phonpasit PHISSAMAY, National Project Director
Valaxay DALALOY, National Project Coordinator
Nadir DURRANI, Project consultant
Dr. Sarmad Hussain, Regional Project Director
(Ms. Oulaiphone SILIMASAK);
(Ms. Chitthaphone CHANHSILILATH);
(Mr. Khamphay INTHARA);
(Mr. Phonexay VILAKONE);

I. Definition:

Lines are broken as result of one of two conditions. The first condition is the presence of an explicit line breaking character. The second condition results from a formatting algorithm having selected among available line break opportunities; ideally the chosen line break results in the optimal layout of the text.

Different formatting algorithms may use different methods to determine an optimal line break. For example, simple implementations consider a single line at a time, trying to find a *locally optimal* line break. A basic, yet widely used approach is to allow no compression or expansion of the inter-character and inter-word spaces and consider the longest line that fits. When compression or expansion is allowed, a locally optimal line break seeks to balance the relative merits of the resulting amounts of compression and expansion for different line break candidates.

When expanding or compressing inter-word space according to common typographical practice, only the spaces marked by U+0020 SPACE, U+00A0 NO-BREAK SPACE, and U+3000 IDEOGRAPHIC SPACE are subject to compression, and only spaces marked by U+0020 SPACE, U+00A0 NO-BREAK SPACE, and occasionally spaces marked by U+2009 THIN SPACE are subject to expansion. All other space characters normally have fixed width. When expanding or compressing inter-character space the presence of U+200B ZERO WIDTH SPACE or U+2060 WORD JOINER is always ignored.

Three principal styles of context analysis determine line break opportunities.

1. *Western* — spaces and hyphens are used to determine breaks: It is commonly used for scripts employing the space character. Hyphenation is often used with space-based line breaking to provide additional line break opportunities. However, it requires knowledge of the language and in addition; it may need user interaction or overrides.
2. *East Asian* — lines can break anywhere, unless prohibited: In these scripts, lines can break anywhere, except before or after certain characters. The precise set of prohibited line breaks may depend on user preference or local custom and is commonly tailorable.
3. *South East Asian* — line breaks require morphological analysis: The third style is used for scripts such as Lao, which do not use spaces, but which restrict word-breaks to syllable boundaries, the determination of which requires knowledge of the language comparable to that required by a hyphenation algorithm. Such an algorithm is beyond the scope of the Unicode Standard.

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

II. Syllabification of Lao Script

Introduction

Syllable is a unit of spoken language, which may have a common meaning or not. A unit of spoken language may have one single syllable or many syllables. Additional complexity is introduced by lack of a space character. Though humans can process multiword string while reading and extract words from it, this process is very difficult for computers. However, unless words can be separated, it is impossible to perform even the simple task of determining how to break at the end of a typed line when characters exceed the line length, without the possibility of breaking between a word and worse between syllables.

In Lao language, the breaking of the text flow after each word (like in English) is not common. One possible way line breaking can be achieved in Lao is to use a Lao lexicon which is currently incompletion. Lao collation is complex because it does not sort on key-press order but on basis of its intricate syllable structure. Thus, before any processing is done, a Lao character string has to be syllabified. This paper develops a Lao string syllabification algorithm. This algorithm is essential for processing basic input character string to enable more advanced language processing including searching, sorting, line breaking and lexical development.

Lao Character Set

Lao syllable structure contains characters and marks for consonants, vowels and tones. Table 1 below lists these possible characters and marks. The marks are combining characters and are shown adjacent to 'x', latter representing a consonant. Character names and their Unicode is also given (ref. to Unicode)

Table 1: Lao Characters

(a) Vowels ('x' is a placeholder for a consonant character)

Vowel	Name	Vowel	Name
Xะ	LAO VOWEL SIGN A (0EB0)	Xู	LAO VOWEL SIGN UU (0EB9)
Xั	LAO VOWEL SIGN MAI KAN (0EB1)	Xົ	LAO VOWEL SIGN MAI KON (0EBB)
Xา	LAO VOWEL SIGN AA (0EB2)	Xຽ	LAO SEMIVOWEL SIGN NYO(0EBD)
Xำ	LAO VOWEL SIGN AM (0EB3)	เX	LAO VOWEL SIGN E (0EC0)
Xิ	LAO VOWEL SIGN I (0EB4)	แX	LAO VOWEL SIGN EI (0EC1)
Xี	LAO VOWEL SIGN II (0EB5)	โX	LAO VOWEL SIGN O (0EC2)
Xึ	LAO VOWEL SIGN Y (0EB6)	ใX	LAO VOWEL SIGN AY (0EC3)
Xื	LAO VOWEL SIGN YY (0EB7)	ไX	LAO VOWEL SIGN AI (0EC4)
Xุ	LAO VOWEL SIGN U (0EB8)	Xໍ	LAO NIGGAHITA (0ECD) (It is a Vowel Sign "OR"

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

(b) Consonants ('x' marks a placeholder for a consonant character)

Cons.	Name	Cons.	Name
ກ	LAO LETTER KO (0E81)	ฝ	LAO LETTER FO TAM (0E9D)
ຂ	LAO LETTER KHO SONG (0E82)	พ	LAO LETTER PHO TAM (0E9E)
ຄ	LAO LETTER KHO TAM (0E84)	ฟ	LAO LETTER FO SONG (0E9F)
ງ	LAO LETTER NGO (0E87)	ม	LAO LETTER MO (0EA1)
ຈ	LAO LETTER CO (0E88)	ย	LAO LETTER YO (0EA2)
ຊ	LAO LETTER SO TAM (0E8A)	ร	LAO LETTER LO LING (0EA3)
ຍ	LAO LETTER NYO (0E8D)	ล	LAO LETTER LO LOOT (0EA5)
ດ	LAO LETTER DO (0E94)	ว	LAO LETTER WO (0EA7)
ຕ	LAO LETTER TO (0E95)	ส	LAO LETTER SO SONG (0EAA)
ຖ	LAO LETTER THO SONG (0E96)	ห	LAO LETTER HO SONG (0EAB)
ທ	LAO LETTER THO TAM (0E97)	อ	LAO LETTER O (0EAD)
ນ	LAO LETTER NO (0E99)	ฮ	LAO LETTER HO TAM (0EAE)
ບ	LAO LETTER BO (0E9A)	◌ີ	LAO VOWEL SIGN LO (0EBC) (It functioning as consonant)
ປ	LAO LETTER PO (0E9B)	ໜ	LAO HO NO (OEDC)
ຜ	LAO LETTER PHO SONG (0E9C)	ໝ	LAO HO MO (OEDD)

(c) Tones ('x' is a placeholder for a consonant character)

Tone	Name	Tone	Name
x◌́	LAO TONE MAI EK (0EC8)	x◌̋	LAO TONE MAI TI (0ECA)
x◌̂	LAO TONE MAI THO (0EC9)	x◌̌	LAO TONE MAI CATAWA (0ECB)

(d) Sign ('x' is a placeholder for a alternate consonant character)

Sign	Name	Sign	Name
ฯ	LAO ELLIPSIS (0EAF)	ๅ	LAO KOLA (0EC6)
x◌̽	LAO CANCELLATION MARK (0ECC)		

Table 1 shows there are 18 vowel marks and characters, 30 character marks and characters, 4 tonal marks and 3 special symbols. Vowels can occur before, above, below a consonantal character or on the baseline. Characters occur on the baseline. Tonal marks always occur on top of consonantal characters.

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

Basic Syllable Structure

Lao language writing system is based on a central or nuclear consonantal character. This consonant may have optionally vowel character or marks around it (before, after, above or below). In addition, this nuclear consonantal character may also have optional a tonal mark above it and optionally more consonantal characters following it. This is illustrated in Figure 1 below. Capital C indicates the nuclear consonantal character. The subscripts indicate that all are optional except the nucleus ***C***.

		$t_{0,1}$		
		$v_{0,1}$		
$V_{0,1}$	$c_{0,1}$	***C***	$v_{0,1}$	$c_{0..4}$
		$c_{0,1}$		
		$v_{0,1}$		

Constraints on Syllable Structure

Not all characters and marks are allowed at all the placeholders shown in Figure 1. Figure 2 and Table 2 below collectively further classify these consonants, vowels and tones to indicate these limitations. The number *n* in X*n* represents the key-press order or typing sequence of these characters and marks relative to each other, except X (which is typed between X1 and X2). Thus tone is typed after the initial vowel X0, consonant X1, nucleus X, and vowel mark X3 and X4 (if X0, X1 and X2 are not null; X can never be null).

		X5						
		X4						
X0	X1	X	X6	X7	X8	X9	X10
		X2						
		X3						

The following table is shown the characters exist in each structure position

X0	X1	X				X2	X3	X4	X5	X6	X7	X8	X9	X10
ເX$_1$	ຫ	ກ$_{01}$	ຂ$_{02}$	ຄ$_{03}$	ງ$_{04}$	Xຼ$_1$	Xຸ$_1$	Xິ$_1$	X່$_1$	ວ$_1$	Xະ$_1$	ກ$_1$	ຈ$_1$	ຯ$_1$
ແX$_2$		ຈ$_{05}$	ສ$_{06}$	ຊ$_{07}$	ຍ$_{08}$	ຣ$_2$	Xູ$_2$	Xີ$_2$	X້$_2$	ອ$_2$	Xາ$_2$	ງ$_2$	ສ$_2$	ໆ$_2$
ໂX$_3$		ດ$_{09}$	ຕ$_{10}$	ຖ$_{11}$	ທ$_{12}$	ວ$_3$		Xຶ$_3$	X໊$_3$	ຽ$_3$	Xຳ$_3$	ຍ$_3$	ຊ$_3$	X໌$_3$
ໄX$_4$		ນ$_{13}$	ບ$_{14}$	ປ$_{15}$	ຜ$_{16}$	ລ$_4$		Xື$_4$	X໋$_4$			ດ$_4$	ພ$_4$	
ໃX$_5$		ຝ$_{17}$	ພ$_{18}$	ຟ$_{19}$	ມ$_{20}$			Xໍ$_5$				ນ$_5$	ຟ$_5$	
		ຢ$_{21}$	ຣ$_{22}$	ລ$_{23}$	ວ$_{24}$			Xົ$_6$				ມ$_6$	ລ$_6$	
		ຫ$_{25}$	ອ$_{26}$	ຮ$_{27}$	ໜ$_{28}$			Xັ$_7$				ບ$_7$		
		ໝ$_{29}$										ວ$_8$		

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

- X0 represents a vowel which occurs before the nuclear consonant. It is can always defined the beginning of syllable.
- X1 is a combination consonant which comes before the nuclear consonant, only if nuclear consonant is one of **{ໆ, ຍ, ລ, ວ, ◌ຼ, ມ, ນ, ຣ}**.
- X is represents the nuclear consonants.
- X2 is a combination consonant which comes after the nuclear consonant, which placing under or next to the nuclear consonant.
- X3 is represents a vowel which occurs under the nuclear consonant.
- X4 is represents a vowel which occurs upper the nuclear consonant.
- X5 is represents a tone marks which occurs upper the nuclear consonant or upper vowels.
- X6 is represents consonant vowel, which occurs after nuclear consonant. It functions when the syllable doesn't have any vowels. And it always exists with X8.
- X7 is represents an after vowels. However $X7_1$ is always represents the end of syllable and it is never exist with tone mark.
- X8 is represents alternate consonants.
- X9 is represents alternate consonant to pronounce foreign language, it always exist with $X10_3$.
- X10 represents a sign mark. X10 is always occurs at the end of syllable, but mostly people keep it separate from syllable.

The nuclear consonant always exists in syllable with some vowel or alternate consonants. The position at which vowels appear can guide to define the beginning and end of each syllable. Using following rules we can find potential syllable boundary.

1. **For $x0_1$ = ເx**
   - 1.1. ເx = $x0_1$ (x1)X(x2)(x5)(x8)(x9:x10)
   - 1.2. ເ◌ິ, ເ◌ີ = $x0_1$ (x1)X(x2)$x4_{1\text{-}2}$(x5)(x8) (x9:x10)
   - 1.3. ເ◌ຶອ, ເ◌ືອ = $x0_1$ (x1)X(x2)$x4_{3\text{-}4}$(x5) $x6_2$ (x8) (x9:x10)
   - 1.4. ເxະ, ເxໆະ = $x0_1$ (x1)X(x2)($x7_2$)$x7_1$
   - 1.5. ເ◌ິໆ = $x0_1$ (x1)X(x2) $x4_6$ (x5) $x7_2$
   - 1.6. ເ◌ັ(x8) = $x0_1$ (x1)X(x2) $x4_7$ (x5) x8 (x9:x10)
   - 1.7. ເ◌ັ(x8) = $x0_1$ (x1)X(x2) $x4_7$ (x5) x8 (x9:x10)
   - 1.8. ເxᩮ, ເ◌ັᩮ = $x0_1$ (x1)X(x2) ($x4_7$)(x5)$x6_3$
2. **For $x0_2$ = ແx**
   - 2.1. ແx = $x0_2$ (x1)X(x2)(x5)(x6)(x8) (x9:x10)
   - 2.2. ແxະ = $x0_2$ (x1)X(x2)$x7_1$
   - 2.3. ແ◌ັ(x8) = $x0_2$ (x1)X(x2) $x4_7$ (x5) x8 (x9:x10)
3. **For $x0_3$ = ໂx**
   - 3.1. ໂx, ໂxວ = $x0_3$ (x1)X(x2)(x5)(x8) (x9:x10)
   - 3.2. ໂxະ = $x0_3$ (x1)X(x2)$x7_1$
   - 3.3. ໂ◌ັວ, ໂ◌ັຍ = $x0_3$ (x1)X(x2)$x4_7$(x5) $x8_{3:8}$

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

4. **For x0$_4$ = ໄx** = x0$_4$ (x1)X(x2)(x5)*(x6$_1$)* (x9:x10)
5. **For x0$_5$ = ໃx** = x0$_5$ (x1)X(x2)(x5) *(x6$_1$)*
6. **For x3 = ◌ຸ & ◌ູ** = (x1)X(x2)x3(x5)(x8) (x9:x10)
7. **For x4$_{1-4}$ = ◌ິ & ◌ີ & ◌ຶ & ◌ື** = (x1)X(x2)x4$_{1-4}$(x5)(x8) (x9:x10)
8. **For x4$_5$ = ◌ໍ =** (x1)X(x2)x4$_5$(x5)(x7$_2$)(x9:x10)
9. **For x4$_6$ = ◌ິ**
   9.1. ◌ິ(x8) = (x1)X(x2)x4$_6$(x5)x8(x9:x10)
   9.2. ◌ິ◌ະ = (x1)X(x2)x4$_6$(x5) x6$_1$ x7$_1$
10. **For x4$_7$ = ◌ັ =** ◌ັ◌, ◌ັ◌◌ = (x1)X(x2)x4$_7$(x5)(x6$_1$)x8(x9:x10)
11. **For x6 = x◌ & x◌ & x◌ =** (x1)X(x2)(x5)x6$_{1-3}$x8(x9:x10)
12. **For x7$_1$ = x◌ะ** = (x1)X(x2)(x5)x7$_1$
13. **For x7$_2$ = x◌ า** = (x1)X(x2)(x5)x7$_2$ (x8)(x9:x10)
14. **For x7$_3$ = x◌ำ** = (x1)X(x2)(x5)x7$_3$ (x9:x10)

Lao Syllabification Algorithm

Lao Syllabification algorithm can be defined as follows:

1. Traverse through the input array and mark syllable boundary as soon as you encounter a punctuation mark, space or a character that does not belong to Lao character set. Example:

   ຄົນ | 10 | ລາວ

   Syllable (boundaries due to non-Lao character)

2. Filter out Lao characters out of input array leaving behind punctuation marks, spaces, and non-Lao characters. Example:

   ຄົນ 10 ລາວ ⟶ ຄົນລາວ

3. Reorder character in case of typing variations (Sometime people maybe typed X5 before X2, X3, X4 and X4 maybe typed before X2). Example:

   ກຸ່ມນີ້

ກ	◌່	◌ຸ	ມ	ນ	◌ັ	◌ີ
X	X5$_1$	X3$_1$	X8	X	X5$_2$	X4$_4$

   Syllable In this case: X3$_1$ should typed before X5$_1$ and X4$_4$ should typed before X5$_2$

ກ	◌ຸ	◌່	ມ	ນ	◌ີ	◌ັ
X	X3$_1$	X5$_1$	X8	X	X4$_4$	X5$_2$

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

4. Mark each character in run with all possible X_n values it can take. Example:

ຄົນລາວ	ຄ	◌ົ	ນ	ລ	າ	ວ
	X	$X4_6$	X	X	$X7_2$	X
			X8	X2		X2
				X9		X6
						X8

5. Use rules discussed in previous section to find out syllable boundaries.
   a. In case if more then one condition suggests a syllable boundary chose the one with longest run.
   b. In case none of the conditions suggest syllable boundary try including last character from previous syllable to current syllable.
   c. Test all the conditions for previous syllable because removing last character might make it invalid. If it is still a valid syllable then try conditions starting from newly added character, other wise restore the previous syllable and skip first character from current syllable and try testing from next character. Keep skipping characters till find a valid syllable boundary.
   d. If program finds boundary for the new syllable it should continue naturally other wise restore previously disturbed syllable by putting back the removed character. Skip the current characters and re-test the conditions, keep skipping unless you find valid syllable.

For example: In case of ເຮືອນີ້ program should identify syllable boundary like

ເຮືອນ | ◌ີ◌້ then it would be unable to detect any condition working for◌ີ◌້ so it should shift syllable boundary one step back and try to include 'ນ' with ◌ີ້ . But before testing this new syllable it should test the previous syllable ເຮືອ and find if it is still a valid syllable which in the current scenario is a valid syllable. Now try testing for ນີ້, this is also a valid syllable so program would continue naturally and go for testing next syllable.

Now consider this example ເຮີະນີ້ program suggests syllable boundary after ເຮີ now as we try test ະນີ້ we would find none of the conditions working so we try to remove ◌ີ from previous syllable and test if ເຮ is still a valid. In case it is we test ◌ີ ະນີ້ we still find none of the conditions working so we restore previous syllable ເຮີ skip

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

current character ะ and carry on testing conditions from ปิ้. Keep skipping characters unless you find valid syllable. In the current example you only have to skip it once.

6. Traverse till the end of array.

7. Put the Lao characters back into original array that have punctuation marks and other non-Lao characters.

Example 1: ໂຄງການL10Nພັດທະນາພາສາລາວ

– Step 1 & 2: Traverse through the input array mark syllable boundary where counter non-Lao character. Extract only Lao text and put it in new array.
ໂຄງການ | ພັດທະນາພາສາລາວ

– Step 3: Look for possible re-ordering of text. For example user might type ‘ ’ X5 before ‘ ’ X47 so we must do the re-ordering.
– Step 4: Mark each character with possible X_N.

ໂ	ຄ	ງ	ກ	າ	ນ	B1	ພ	◌ັ	◌่	ດ	ທ	ะ	B1	ນ	າ	ພ	າ	ສ	າ	ລ	າ	ວ
A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	A20	A21	A22
$X0_3$	X_{03}	X_{04}	X_{01}	$X7_2$	X_{13}		X_{18}	$X4_7$	X5	X_{09}	X_{12}	$X7_1$		X_{13}	$X7_2$	X_{18}	$X7_2$	X_{06}	$X7_2$	X_{23}	$X7_2$	X_{24}
		$X8_2$	$X8_1$		$X8_5$		$X9_4$			$X8_4$				$X8_5$		$X9_4$		$X9_2$		$X2_4$		$X2_3$
																				$X9_8$		$X6_1$
																						$X8_8$

– Step 5: Use rules and conditions discussed above to define syllable boundaries.

Rule 3.1		**Rule 13**			**Rule 10**					**Rule 12**			**Rule 13**			**Rule 13**			**Rule 13**			**Rule 13**			
ໂ	ຄ	ງ		ກ	າ	ນ		ພ	◌ັ	◌่	ດ		ທ	ะ		ນ	າ		ພ	າ		ສ	າ		ລ	າ	ວ
A0	A1	A2		A3	A4	A5		A7	A8	A9	A10		A11	A12		A14	A15		A16	A17		A18	A19		A20	A21	A22

– Step 6: Put the Lao characters back into original array.

Example 2: ແຂວງຫລວງພະບາງ14ກຸມພາ2005

– Step 1 & 2**:** Traverse through the input array mark syllable boundary where you counter non-Lao character. Extract only Lao text and put it in new array.

ແຂວງຫລວງພະບາງ | ກຸມພາ |

- Step 3: Look for possible re-ordering of text. For this example there is none.
- Step 4: Mark each character with possible Xn**.**

ແ	ຂ	ວ	ງ	ຫ	ລ	ວ	ງ	ພ	ະ	B	ບ	າ	ງ	B	ກ	ຸ	ມ	ພ	າ
A0	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**A11**	**A12**	**A13**	**A14**	**A15**	**A16**	**A17**	**A18**	**A19**
$X0_2$	**X**	**X**	**X**	**X1**	**X**	**X**	**X**	**X**	**$X7_2$**		**X**	**$X7_1$**	**X**		**X**	**X3**	**X**	**X**	**$X7_1$**
		X2	**X8**	**X**	**X2**	**X2**	**X8**	**X9**			**X8**		**X8**		**X8**		**X8**	**X9**	
		$X6_1$			**X9**	**$X6_1$**													
		X8				**X8**													

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la

- Step 5: Use rules and conditions discussed above to define syllable boundaries.

Rule 2.1					**Rule 11**					**Rule 12**			**Rule 13**				**Rule 6**				**Rule 13**	
ແ	ຂ	ວ	ງ	B	ຕ	ລ	ວ	ງ	B	ພ	ະ	B	ບ	າ	ງ	B	ກ	ຸ	ມ	B	ພ	າ
A0	A1	A2	A3	B	A4	A5	A6	A7	B	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	B	A18	A19

- Step 6: Put the Lao characters back into original array.

III. Lao Line Breaking Utility

Screen Dumps

Given below are some screen shots that might be helpful in facilitating the understanding.

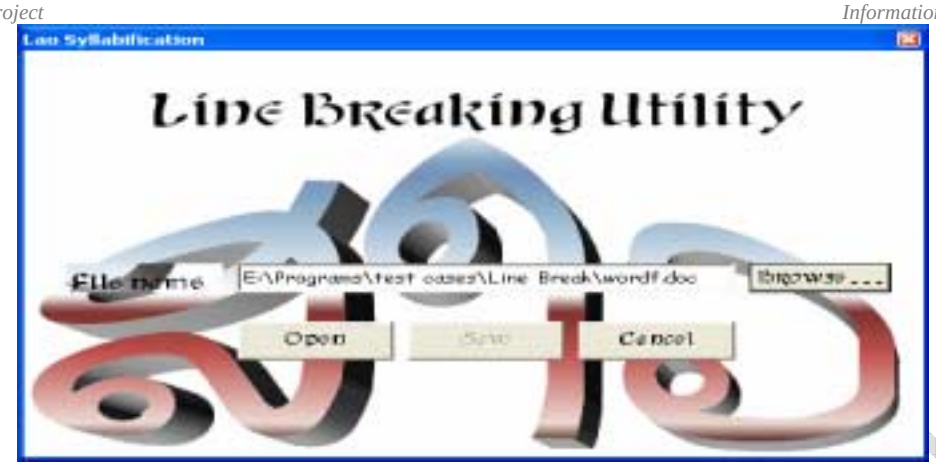
ສູນເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ (IT Center) ຮັບຜິດຊອບ ການຄົ້ນຄ້ວາ, ກ
ານພັດທະນາ, ການບໍລິການ ການເຜີ່ຍ ເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ, ເຊິ່ງນັບແ
ຕ່ສ້າງຕັ້ງມາ ທາງສູນແມ່ນໄດ້ປະກອບ ສ່ວນໃນການພັດທະນາ ເຕັກໂນໂລຊີ
ສື່ສານຂໍ້ມູນຂ່າວສານ ຢູ່ ສປປ ລາວ ຢ່າງແຂງແຮງ

Highlighted text represents the problem. Word is unable to break the line properly due to the continuum of words without any space or any other clue which can be used to decide where to break the line. .We have to insert these clues inform of zero width space.

For instance in first-second lines consider ນັບແຕ່ since there is no space for remaining ຕ່ in the currently line word shifts it to next line blindly without realizing that it was part of ແ. However, if there were a zero width space in between ນັບ and ແຕ່ then it would shift whole syllable to next line. Similar is the case with ກາຫ in line number first where ຫ being part of the syllable is dragged to next line.

User can browser for the file and let the program read the input and do the rest.

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la



Program displays a success message and prompts the user to save the modified file with new name or replace the existing after which the output file is displayed automatically.

Output file:

ສູນເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ (IT Center) ຮັບຜິດຊອບ ການຄົ້ນຄ້ວາ, ການພັດທະນາ, ການບໍລິການ ການເຜີ່ຍ ເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ, ເຊິ່ງນັບແຕ່ ສ້າງຕັ້ງມາ ທາງສູນແມ່ນໄດ້ປະກອບ ສ່ວນໃນການພັດທະນາ ເຕັກໂນໂລຊີ ສື່ສານຂໍ້ມູນຂ່າວສານ ຢູ່ ສປປ ລາວ ຢ່າງແຂງແຮງ

As you can see that syllable boundaries are now properly defined as word finds the zero width space after every syllable which provides it help to properly segment the words and break the line properly. If we put a space between ນັບ and ແຕ່ it would not shift only ຕ່ but complete syllable ແຕ່ to the next line as shown below.

ສູນເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ (IT Center) ຮັບຜິດຊອບ ການຄົ້ນຄ້ວາ, ການພັດທະນາ, ການບໍລິການ ການເຜີ່ຍ ເຕັກໂນໂລຊີຂໍ້ມູນຂ່າວສານ, ເຊິ່ງນັບ ແຕ່ສ້າງຕັ້ງມາ ທາງສູນແມ່ນໄດ້ປະກອບ ສ່ວນໃນການພັດທະນາ ເຕັກໂນໂລຊີ ສື່ສານຂໍ້ມູນຂ່າວສານ ຢູ່ ສປປ ລາວ ຢ່າງແຂງແຮງ

And as you can see there is a corresponding adjustment in line three where entire syllable ການ moves onto next line.

P.O Box 2279, Vientiane, Lao PDR, Tel: (856-21) 217706 ext 108; Fax: (856-21) 213472
Email: itc@stea.gov.la; Web: www.stea.gov.la