Lao Syllabification for line breaking

By
Valaxay DALALOY
Science Technology and Environment Agency
8-2006

Lao Character Set

- Lao character set consists of 65 such as:

- 27 consonants and 3 special consonants

ກ ຂ ຄ ງ ຈ ສ ຊ ຍ ດ ຕ ຖ ທ ນ ບ ປ

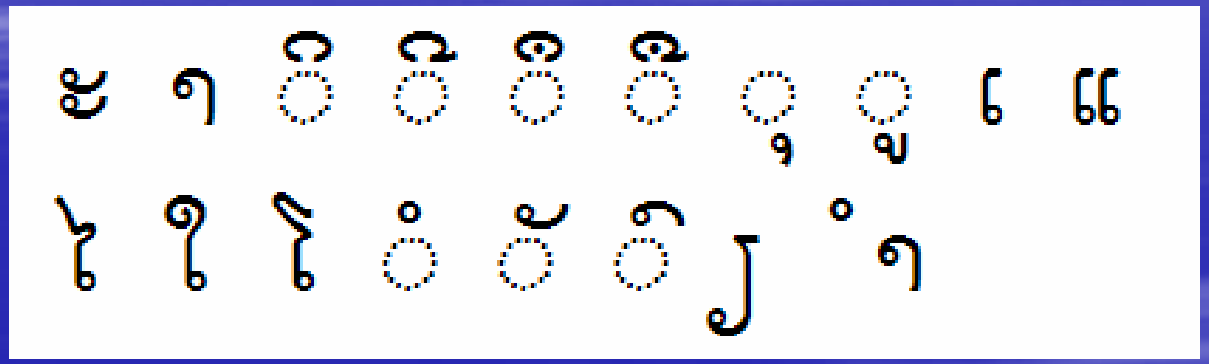
ຜ ຝ ພ ຟ ມ ຢ ຣ ລ ວ ອ ຮ ◌ ຫ ໝ

- 6 mix consonants

ห + ง = หง	ห + ย = หย
ห + น = หน (ห)	ห + ม = หม (ห)
ห + ล = หล (ฬ)	ห + อ = หอ

Lao Character Set (Cont.)

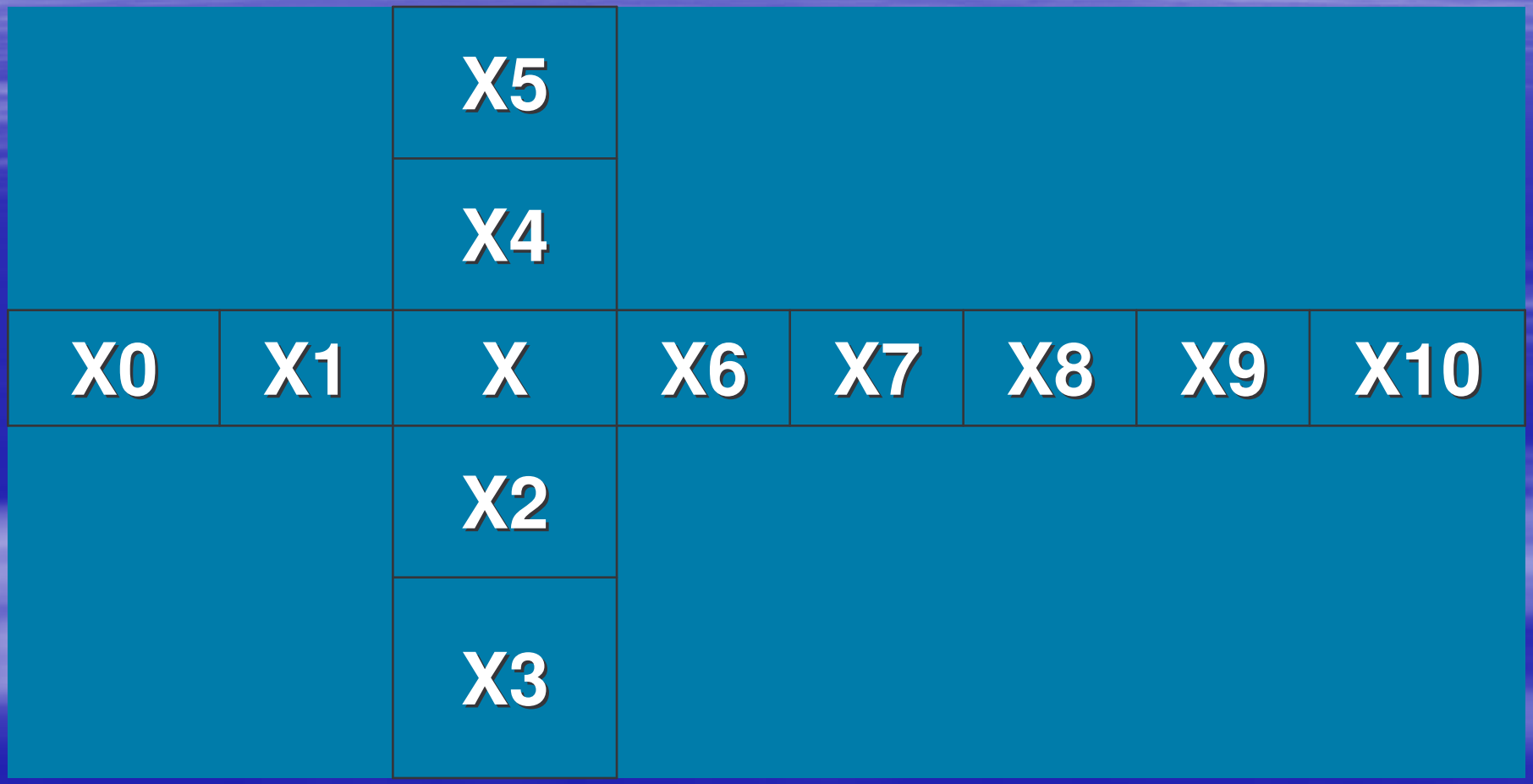
18 Vowels



4 tone marks and 3 signs

10 Lao digits

Lao Syllable Structure



Lao Syllable Structure (Cont.)

X0	X1	X			X2	X3	X4	X5	X6	X7	X8	X9	X10
เX1	ຫ	ກ01	ຂ02	ຄ03	X1	X1	X1	X1	ອ1	Xະ1	ກ1	ຈ1	◌ 1
แX2		ງ04	ຈ05	ສ06	◌ 2	X2	X2	X2	ຮ2	Xา2	ງ2	ສ2	◌ 2
ไX3		ຊ07	ຍ08	ດ09	◌ 3		X3	X3	◌ 3	Xำ3	ຍ3	ຊ3	X3
ໄX4		ຕ10	ຖ11	ท12	◌ 4		X4	X4			ด4	ฟ4	
ใX5		น13	บ14	ป15			X5				น5	ฟ5	
		ผ16	ฝ17	พ18			X6				ม6	ล6	
		ฟ19	ม20	ย21			X7				บ7		
		ร22	ล23	ว24							อ8		
		ห25	อ26	ฮ27									
		ฬ28	◌ 29										

Rules for Lao Syllabification Segmentation

1. For $x0_1$ = ◌x

1.1. ◌x = $x0_1$ (x1)X(x2)(x5)(x8)(x9:x10)

1.2. ◌ิx, ◌ีx = $x0_1$ (x1)X(x2)$x4_{1\text{-}2}$(x5)(x8) (x9:x10)

1.3. ◌ึxอ, ◌ืxอ = $x0_1$ (x1)X(x2)$x4_{3\text{-}4}$(x5) $x6_2$ (x8) (x9:x10)

1.4. ◌xะ, ◌xาะ = $x0_1$ (x1)X(x2)($x7_2$)$x7_1$

1.5. ◌ำxา = $x0_1$ (x1)X(x2) $x4_6$ (x5) $x7_2$

1.6. ◌ัx(x8) = $x0_1$ (x1)X(x2) $x4_7$ (x5) x8 (x9:x10)

1.7. ◌ัx(x8) = $x0_1$ (x1)X(x2) $x4_7$ (x5) x8 (x9:x10)

1.8. ◌xๅ, ◌ัxๅ = $x0_1$ (x1)X(x2) ($x4_7$)(x5)$x6_3$

Rules for Lao Syllabification Segmentation (Cont.)

2. For $x0_2$ = ◌ັx

2.1. ◌ັx = $x0_2$ (x1)X(x2)(x5)(x6)(x8) (x9:x10)

2.2. ◌ັxะ = $x0_2$ (x1)X(x2)$x7_1$

2.3. ◌ັx̌(x8) = $x0_2$ (x1)X(x2) $x4_7$ (x5) x8 (x9:x10)

3. For $x0_3$ = ໃx

3.1. ໃx, ໃxอ = $x0_3$ (x1)X(x2)(x5)(x8) (x9:x10)

3.2. ໃxะ = $x0_3$ (x1)X(x2)$x7_1$

3.3. ໃx̌อ, ໃx̌ย = $x0_3$ (x1)X(x2)$x4_7$(x5) $x8_{3:8}$

Rules for Lao Syllabification Segmentation (Cont.)

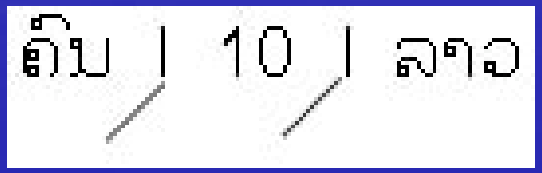
4. **For $x0_4$ = ไx** = $x0_4$ (x1)X(x2)(x5)*($x6_1$)* (x9:x10)
5. **For $x0_5$ = ใx** = $x0_5$ (x1)X(x2)(x5) *($x6_1$)*
6. **For x3 = ◌ุx & ◌ูx** = (x1)X(x2)x3(x5)(x8) (x9:x10)
7. **For $x4_{1\text{-}4}$ = ◌ิx & ◌ีx & ◌ึx & ◌ืx** = (x1)X(x2)$x4_{1\text{-}4}$(x5)(x8) (x9:x10)
8. **For $x4_5$ = ◌ัx =** (x1)X(x2)$x4_5$(x5)($x7_2$)(x9:x10)
9. **For $x4_6$ = ◌ิ้x**
   - 9.1. ◌ิ้x(x8) = (x1)X(x2)$x4_6$(x5)x8(x9:x10)
   - 9.2. ◌ิ้xอะ = (x1)X(x2)$x4_6$(x5) $x6_1$ $x7_1$

Rules for Lao Syllabification Segmentation (Cont.)

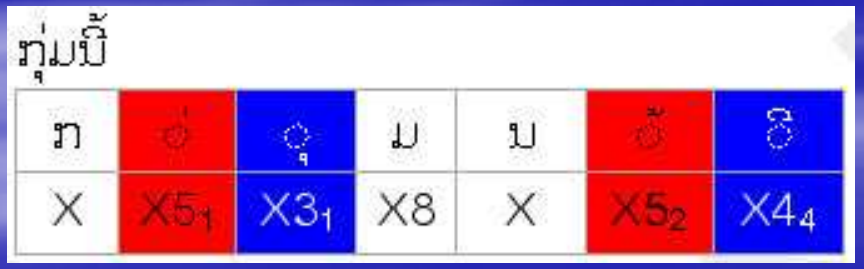
10. **For x4$_7$ = ◌ັ = ◌ັດ, ◌ັອດ** = (x1)X(x2)x4$_7$(x5)(x6$_1$)x8(x9:x10)
11. **For x6 = ◌ວ & ◌ອ & ◌ຽ =** (x1)X(x2)(x5)x6$_{1\text{-}3}$x8(x9:x10)
12. **For x7$_1$ = ◌ະ** = (x1)X(x2)(x5)x7$_1$
13. **For x7$_2$ = ◌ໆ** = (x1)X(x2)(x5)x7$_2$ (x8)(x9:x10)
14. **For x7$_3$ = ◌ໍ** = (x1)X(x2)(x5)x7$_3$ (x9:x10)

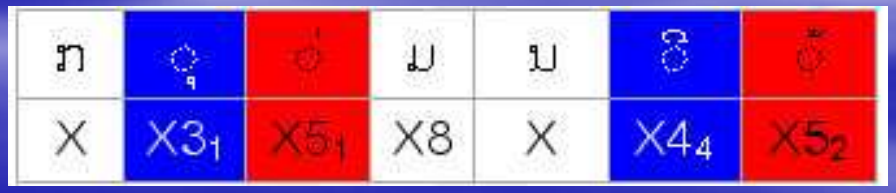
Lao Syllabification Algorithm

- mark syllable boundary by a punctuation mark, space or a character that does not belong to Lao character set



- Reorder character in case of typing variations





Lao Syllabification Algorithm (Cont.)

Mark each character in run with all possible Xn values it can take

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

THANK YOU