Microsoft Office Plug-ins for Lao Encoding Standardization and for Line Breaking

Nation Authority of Science and Technology

1. Introduction

Most of the Lao language and script standardization tasks have been completed in the Phase-1 of the Pan Localization project [1]. It includes Lao encoding standardization, fonts, keyboard standard, Syllabification for line breaking, Collation and spelling checker. Basic utilities are developed to perform these algorithms on simple text files. But the use of the these utilities is very limited or cumbersome as most of the text processing tasks have been performed using word processors like MS Office or OpenOffice. The simple text can be processed using the developed utilities but the text formatting is lost in this processes. There is strong demand to develop MS Office plug-ins for these utilities.

2. Methodology

There are mainly two parts of each plug-in; the language processing module and the MS Office automation. The function of language processing module is to perform language dependent algorithms for encoding standardization and syllabification. MS Office automation functionality is to get the data/text from Office applications and put the data back (keeping the formatting intact) after performing language dependent algorithm. Different automation modules are required for each of Office applications MS Word, MS Excel and MS PowerPoint as shown in Figure 1. For the language processing modules the API's are developed using the existing code (developed in Phase-1). For the MS Office automation the code is taken from Cambodian component of PAN Localization component (who has already developed the similar utilities for Khmer language) and customized for Lao language.

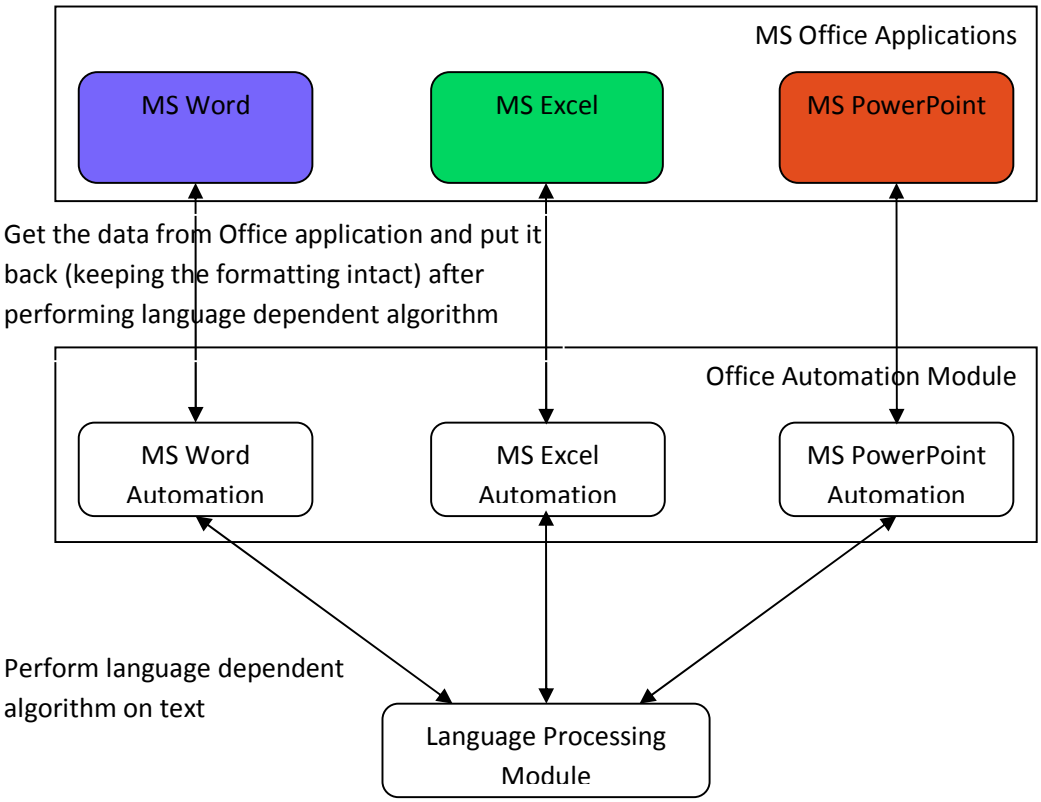

Figure 1: High level architecture of MS Office plug-ins

3. Lao Encoding Standardization

Lao was traditionally written using as ASCII encoding by replacing Latin characters with Lao characters. With the inclusion of Lao language in Unicode now it is possible to define Unicode fonts, display and store Lao data in Unicode format. But the operating systems (until MS Windows XP) do not provide any mechanism for inputting or editing the data directly. Users need to install proprietary software (Lao Script for Windows) to achieve this functionality. Lack of Unicode awareness and need of proprietary software for inputting the data are the factors that still force users to use modified ASCII encoding for Lao. There is a need to develop automatic encoding converts from ASCII to Unicode for text data and as well as for MS Office applications.

Due to Lack of encoding standards (before Unicode), vendors have defined their custom encoding based on ASCII. There are about 140 different non-Unicode fonts and their versions exist for Lao language. In first step the fonts are categories into 46 different groups according

to the encoding they used. And 46 mapping tables are defined for converting non-Unicode encoding to Unicode. These 46 groups are further reduced to 24 groups after taking the union of mapping tables.

Mapping tables are simple text files with the following format

NON-UNICODE ENCODING: UNICODE

Example:

```
005C:0EDD

59:0EB4+0EC9
```

To associate the fonts with their coressponding mapping tables, Font Groups table is defined. The format of this table is:

SOURCE FONT = TARGET FONT = SIZE FECTOR = MAP TABLE

Example:

```
#Group=13
     LaokeyB1 = Saysettha OT = 1 = 13
     LaofontNM = Saysettha OT = .75 = 13
#Group=14
     Kmodern3 Sample = Saysettha OT = 1.2 = 14
```

Where the *source font* is the non-Unicode font and the *target font* is the Unicode font. *Size fector* is a muliplying fector to change the type size of Unicode fonts to match it with the type size of non-Unicode fonts. *Map table* defines the coressponding mapping table for conversion.

3.1 MS Office Plug-ins for Lao Encoding Standardization

Three different plug-ins are developed of each of MS Office application: MS Word, MS Excel and MS PowerPoint. A setup program is also developed to easily install these plug-ins.

Figure 2: Installer program for plug-ins

After Installing the plug-ins the option to convert the doument is available under the add-ins ribon of Office 2007 as shown in Figure 3.

Figure 3: Conversion plug-in in MS Word 2007

The plug-in automatically detects the active document and creates a new document with Unicode encoding.

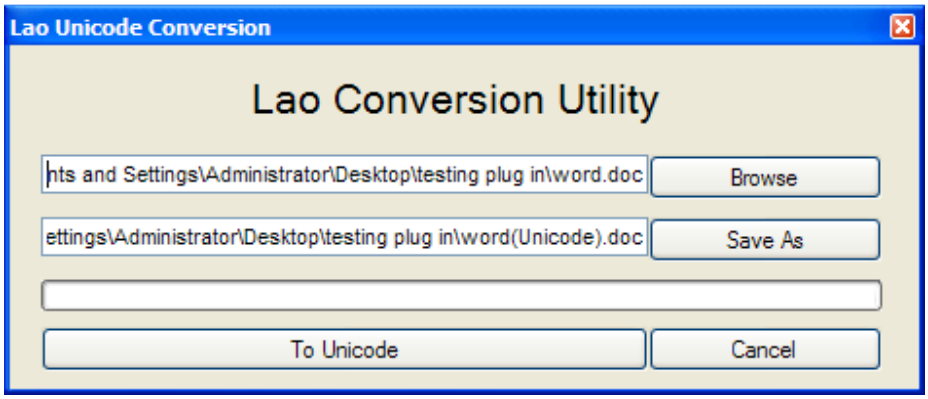

Figure 4: Plug-in for conversion

4. MS Office plug-in for line breaking

Lao language does not use spacing between words. Apart from other word segmentation issues (e.g. spelling checker, find next word etc.) word wrapping is also a major problem. It is always preferable to break lines at word boundaries but in worst conditions when word segmentation algorithms are not available yet, Lao script can also be break at syllable boundaries for word wrapping. Fortunately, Lao syllables can easily be detected by simple rule based methods and P. Phissamay et al. discussed these rules in detail in Syllabification of Lao Script for Line Breaking [2].

4.1 Syllable structure of Lao script

Lao character set is composed of 30 consonants, 18 vowels, 4 tone marks and three signs. 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. The Basic syllable structure of Lao script is shown in Figure 5.

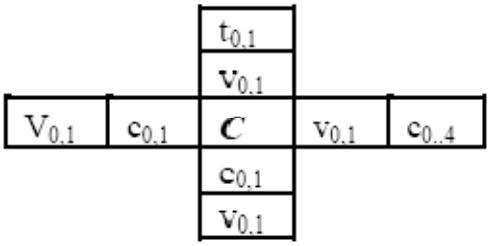

Figure 5: Basic syllable structureof Lao script

The capital 'C' indicates the nuclear consonant. Whereas, 'v' and 'c' and the subscripts indicate that all are optional except the nucleus C.

4.2 MS Office Plug-in

Two plug-ins are developed for MS Word and MS PowerPoint. The line break plug-in has a simple interface as shown in Figure 6.

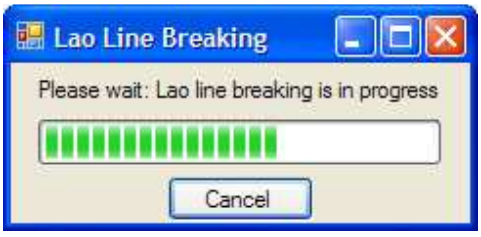

Figure 6: MS Office Plug-in for line Breaking

4.3 Results of Line Breaking Plug-in

Figure 7 and 8 show the results before and after the line breaking respectively.

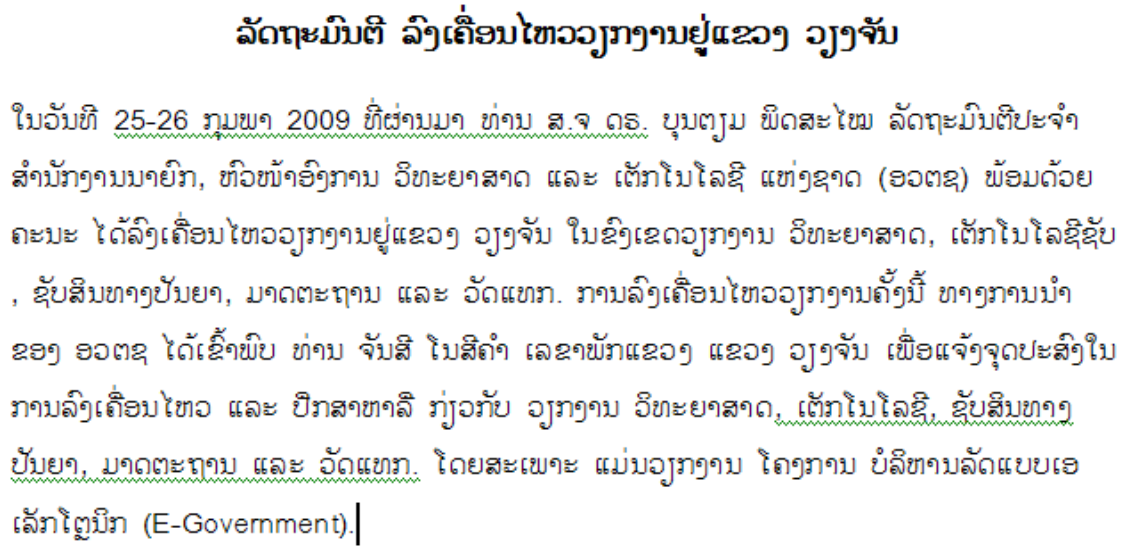

Figure 7: Text in MS Word without line breaking

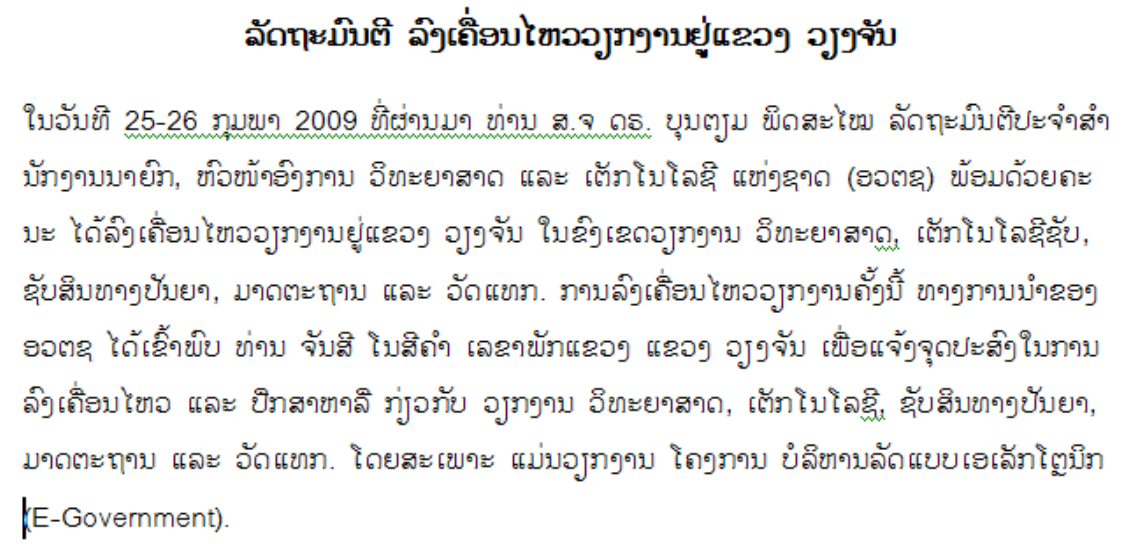

Figure 8: Text in MS Word after line breaking

Refrences:

[1]: Phase-1 outputs of Lao country component of Pan Localization Project. http://panl10n.net/english/OutputsLaos1.htm

[2]: Syllabification of Lao Script for Line Breaking by P. Phissamay,V. Dalolay,C, Chanhsililath, O. Silimasak, S. Hussain, N. Durrani, Science Technology and Environment Agency and CRULP