Open 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 all most of these utilities are not supported to Open Office, so we have to start reprogramming them again using open source and plug-in into it in Phrase-2.

2. Methodology

There are mainly two parts of each plug-in; the language processing module and the Open Office automation. The function of language processing module is to perform language dependent algorithms for encoding standardization and syllabification. Open Office automation functionality is to get the data/text from Office applications and put the data back after performing language dependent algorithm. For the language processing modules, the API's are developed in Phase-2. For the Open 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. The tool we used for developing these utilities is Open office – Net Beans Integration.

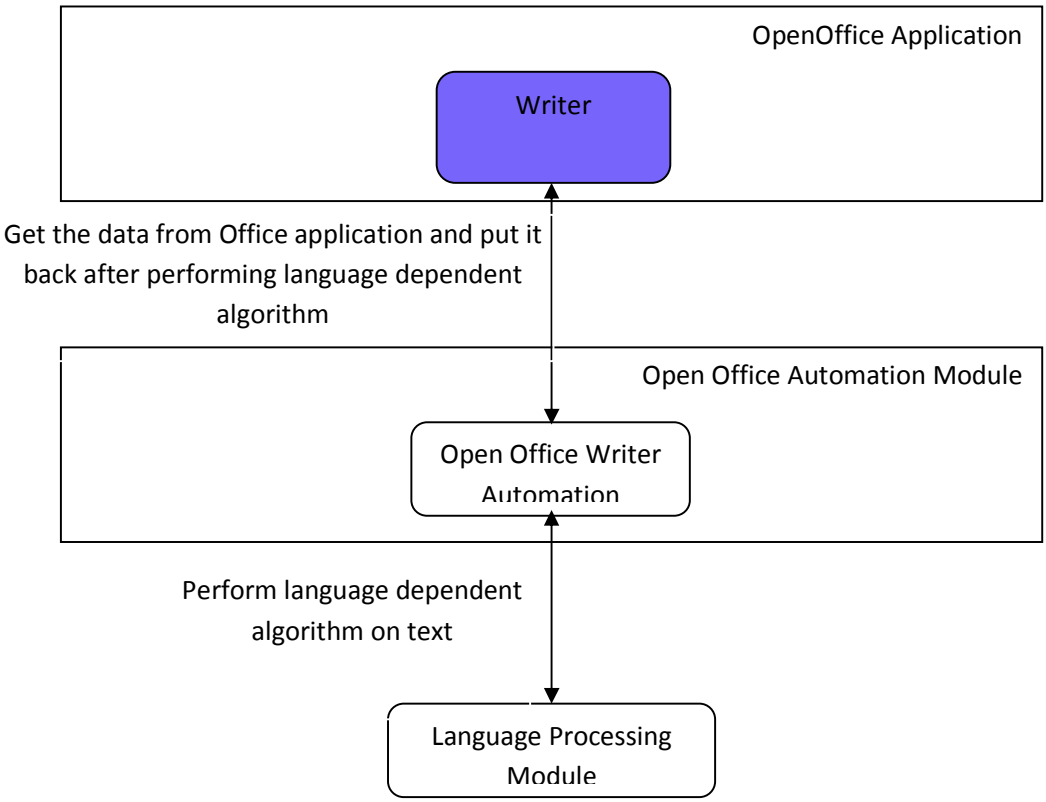


Figure 1: High level architecture of MS Office plug-ins

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 Open 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 25 mapping tables are defined for converting non-Unicode encoding to Unicode.

Mapping tables are simple text files with the following format

NON-UNICODE ENCODING: UNICODE

Example:

```
00FE:0ECC

2122:0EB4+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=1
     Alice0 95 = Alice5 OT = 1 = 1
     Saysettha Lao = Saysettha OT = 1 = 1
#Group=2
     Alice0 2000 = Alice5 OT = 1 = 2
     Saysettha 2000 = Saysettha OT = 1 = 2
```

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 OpenOffice Plug-ins for Lao Conversion

The plug-in is developed for Open Office Writer application. The setup steps of plug-in for Lao Conversion are shown as below.

Step 1 Setup the necessary components

Double click the “Lao **Open Office Plugin.msi** “and follow the instruction.

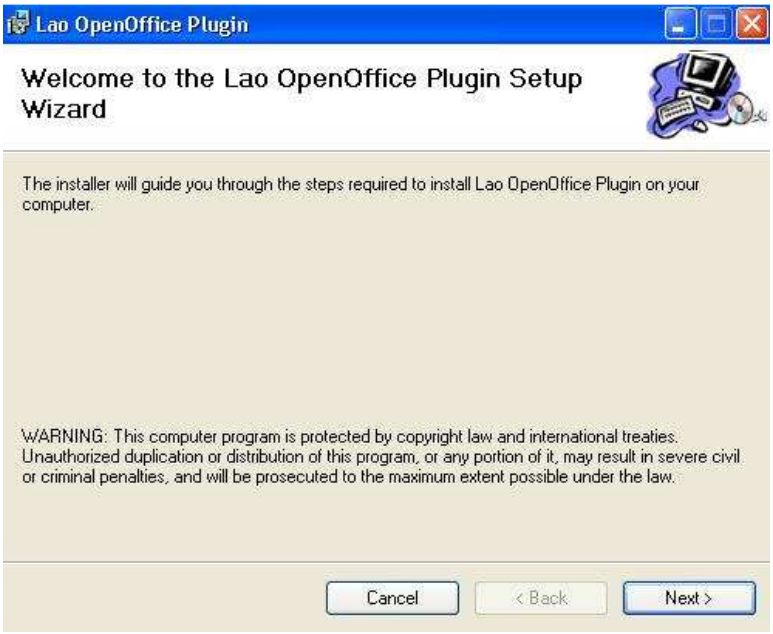


Figure 2

Step 2 Deploy the Lao Conversion Extension (Figure3 and Figure4)

Double click the “LaoConversion.oxt”

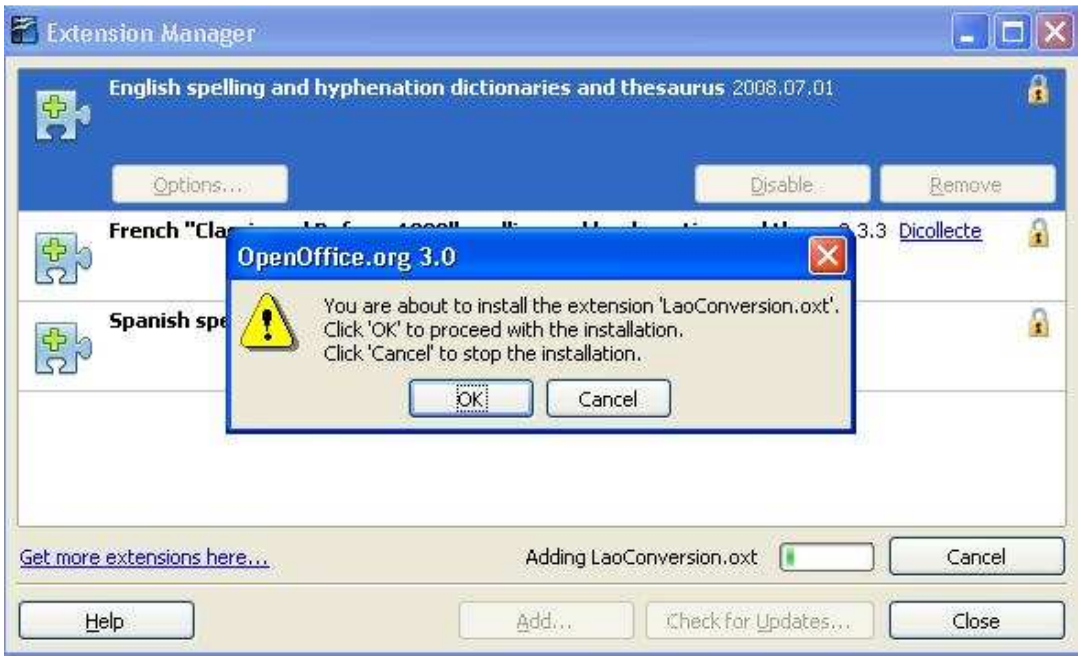


Figure 3

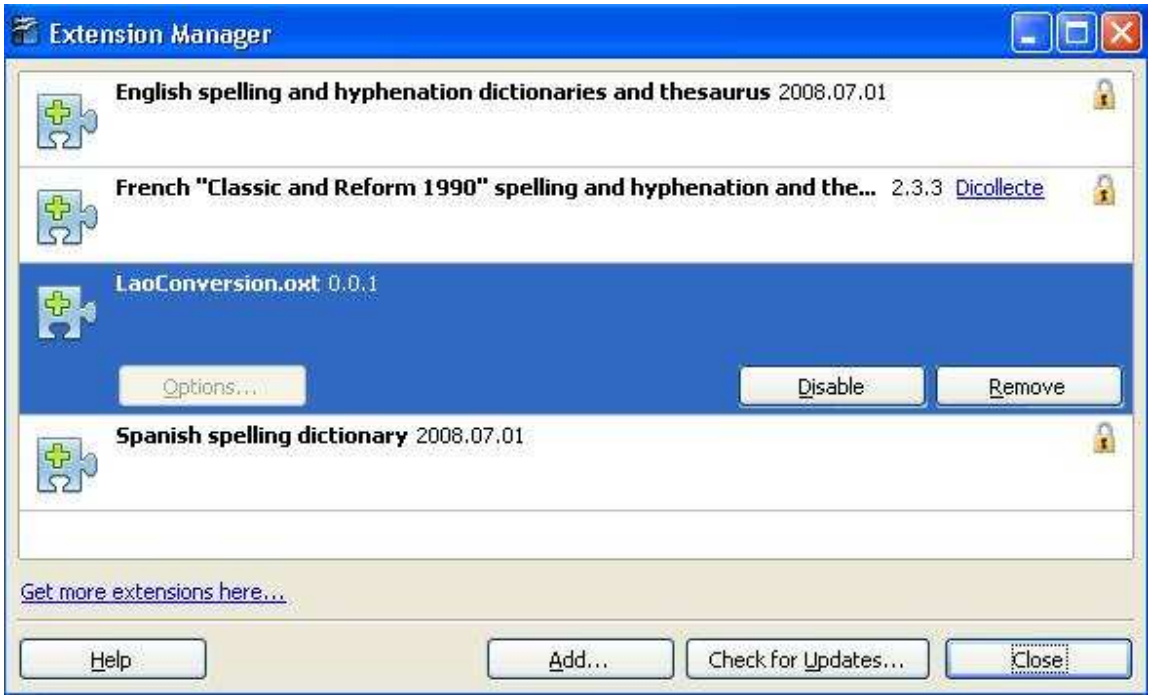


Figure 4

After Installing the plug-ins the option to convert the doument is available under the AddOn Menu of OpenOffice.org Writer as shown in Figure 5.

Figure 5: Conversion plug-in in Open Office Writer

The plug-in automatically detects the active document and the document will be overridden with Unicode encoding.

4. Open 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 (developed in Phase 1).

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 6.

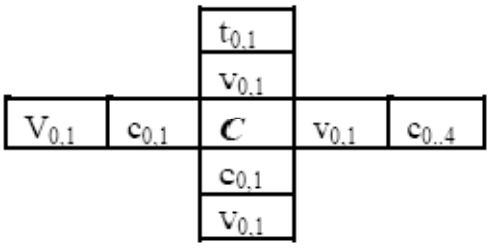


Figure 6: Basic syllable structure of Lao script

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

4.1 OpenOffice Plug-in

The plug-in is developed for Open Office Writer application. The setup steps of plug-in for Lao LineBreaking are same as Lao Conversion. If the step 1 (mentioned in the Lao Conversion plug-in) is already installed, you can skip it and continue the following. Otherwise you have to start from step1 and then continue the following.

Deploy the Lao Line Breaking Extension by

Double click at " **Lao LineBreaking.oxt** "

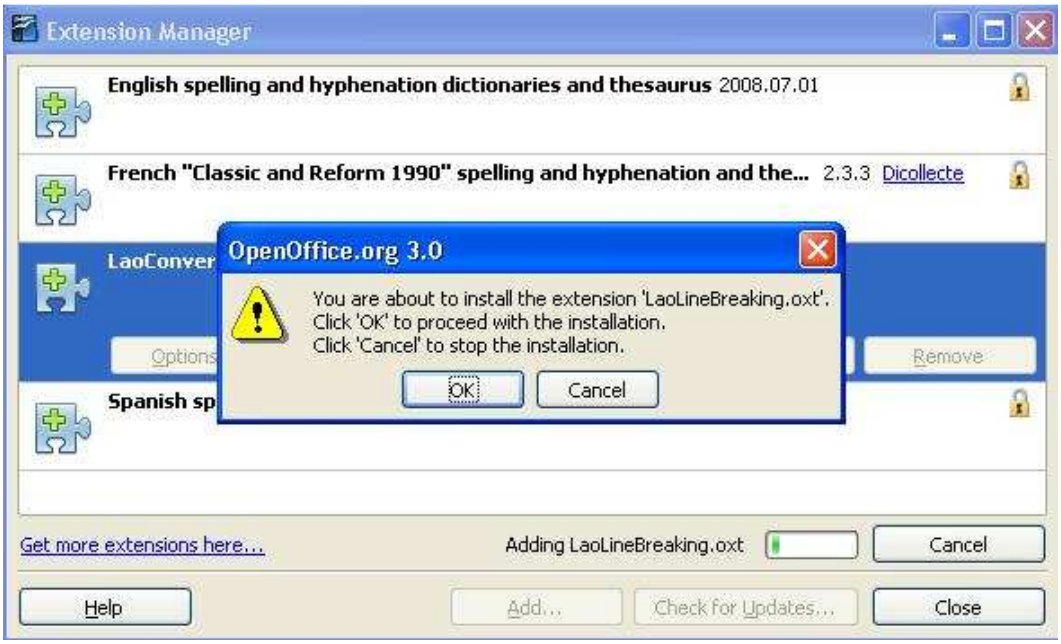


Figure 7

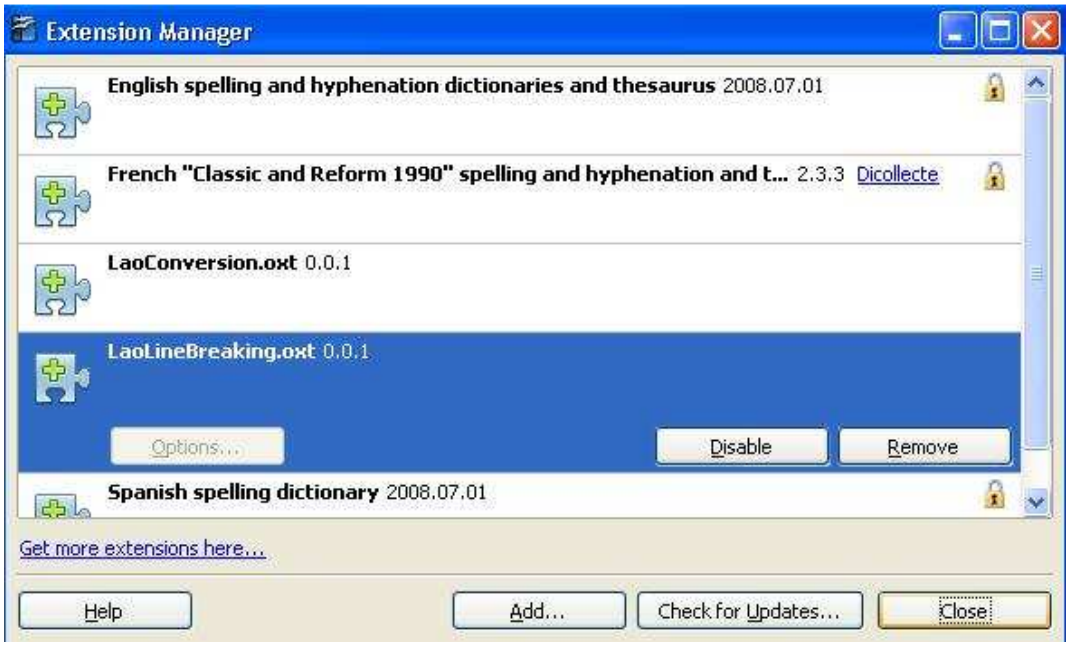


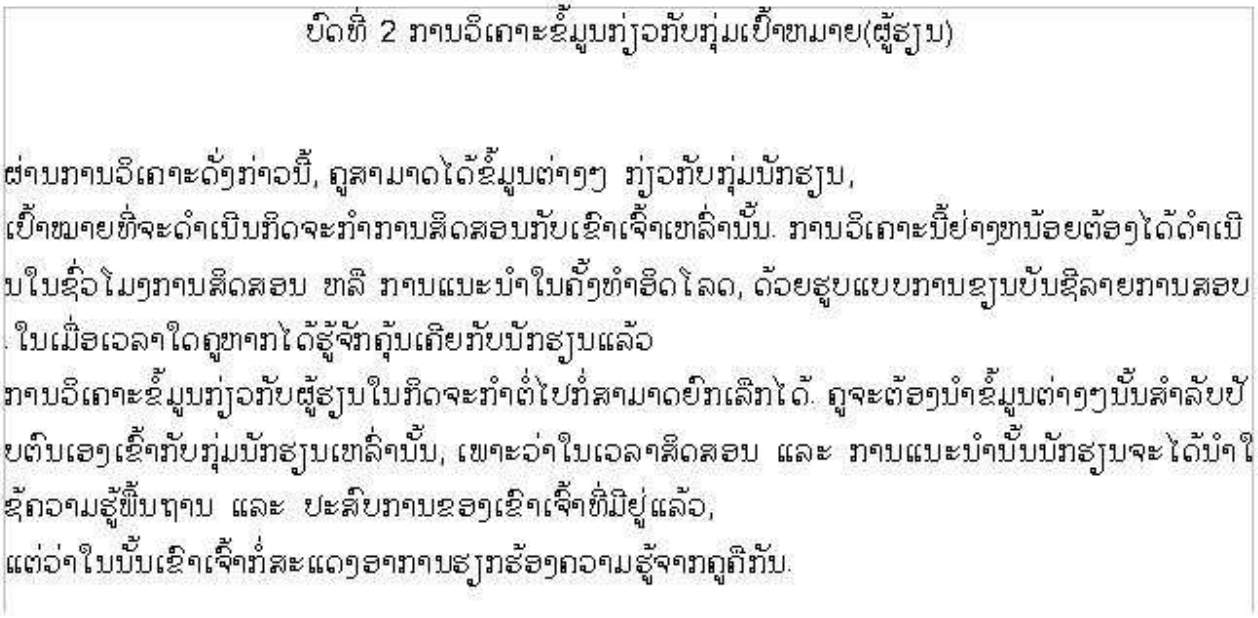
Figure 8

After Installing the plug-ins the option to line break the doument is available under the AddOn Menu of OpenOffice.org Writer as shown in Figure 9.

Figure 9: LineBreaking plug-in in Open Office Writer

4.2 Results of Line Breaking Plug-in

Figure 10 and 11 and 12 show the results before and after the line breaking respectively.


ບົດທີ 2 ການວິເຄາະຂໍ້ມູນກ່ຽວກັບກຸ່ມເປົ້າຫມາຍ(ຜູ້ຮຽນ)

ຜ່ານການວິເຄາະດັ່ງກ່າວນີ້, ຄູສາມາດໄດ້ຂໍ້ມູນຕ່າງໆ ກ່ຽວກັບກຸ່ມນັກຮຽນ,
ເປົ້າໝາຍທີ່ຈະດຳເນີນກິດຈະກຳການສິດສອນກັບເຂົາເຈົ້າເທລົ່ານັ້ນ. ການວິເຄາະນີ້ຢ່າງຫນ້ອຍຕ້ອງໄດ້ດຳເນີ
ນໃນຊົ່ວໂມງການສິດສອນ ຫລື ການແນະນຳໃນຄັ້ງທຳອິດໂລດ, ດ້ວຍຮູບແບບການຂຽນບັນຊີລາຍການສອບ
ໃນເມື່ອເວລາໃດຄູຫາກໄດ້ຮູ້ຈັກຄຸ້ນເຄີຍກັບນັກຮຽນແລ້ວ
ການວິເຄາະຂໍ້ມູນກ່ຽວກັບຜູ້ຮຽນໃນກິດຈະກຳຕໍ່ໄປກໍ່ສາມາດຍົກເລີກໄດ້. ຄູຈະຕ້ອງບຳຂໍ້ມູນຕ່າງໆນັ້ນສຳລັບປັ
ບຕົນເອງເຂົ້າກັບກຸ່ມນັກຮຽນເທລົ່ານັ້ນ, ເພາະວ່າໃນເວລາສິດສອນ ແລະ ການແນະນຳນັ້ນນັກຮຽນຈະໄດ້ນຳໃ
ຊ້ຄວາມຮູ້ພື້ນຖານ ແລະ ປະສົບການຂອງເຂົາເຈົ້າທີ່ມີຢູ່ແລ້ວ,
ແຕ່ວ່າໃນນັ້ນເຂົາເຈົ້າກໍ່ສະແດງອາການຮຽກຮ້ອງຄວາມຮູ້ຈາກຄູຄືກັນ.

Figure 10: Text in Open Office Writer without line breaking

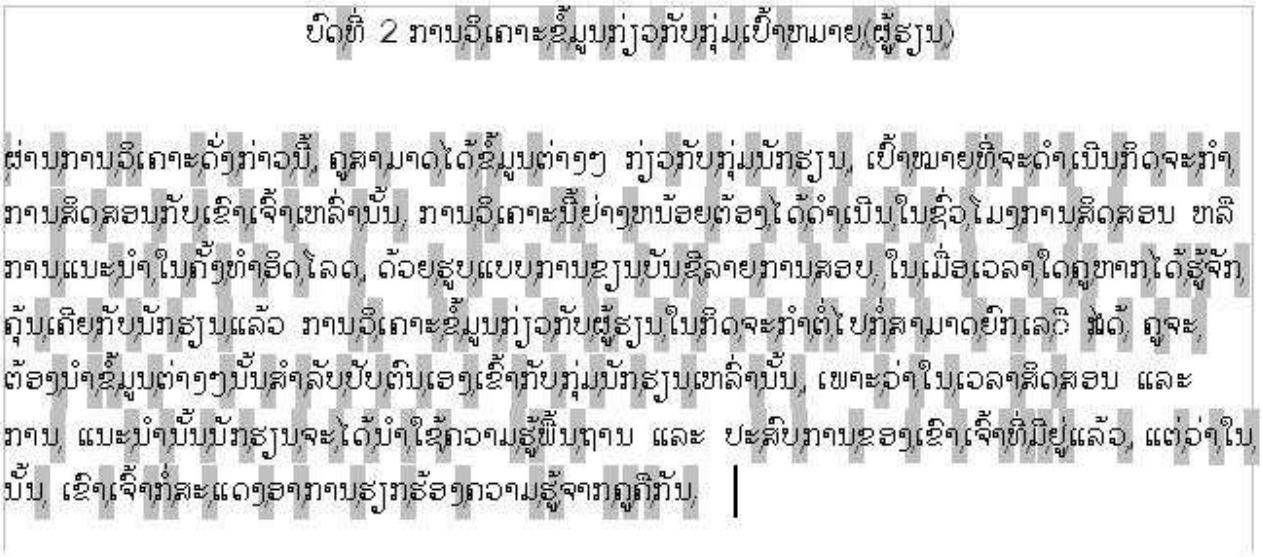


Figure 11: Text in Open Office Writer after line breaking

To remove unwanted shading characters, deselect the option (Go to “View **> Field Shadings** “)

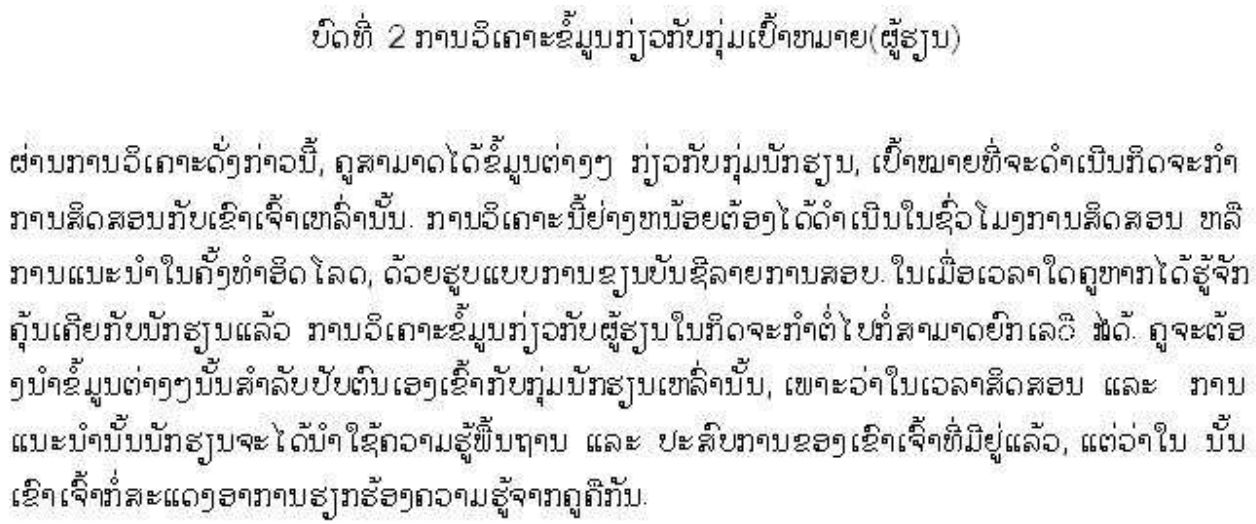


Figure 12: Text after deselected Field Shadings

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.

[3]: Technical Report for Microsoft Office Plug-in by Atif GULZAR.