Translation of LibreOffice Guides (in two Languages in Parallel)

Miloš Šrámek and Stanislav Horáček

SOIT
Society for Open Information Technologies (Slovakia)

LibreOffice Guides

- All components covered:
  - Writer, Calc, Impress, Draw, Base, Math + Getting Started with LibreOffice
  - Keeping pace with LO development
  - In English available at:
    - https://wiki.documentfoundation.org/Documentation/Publications
  - Authors' web page: http://www.odfauthors.org/

Translations of the Guides

- Translated to a few languages:
    - Esperanto GS 3.5, 4 chapters
    - Spanish: GS 3.3, full
    - French: GS, 3.5, 4.0, 9 chapters; WG 4.0, 9 ch.; CG, 4.1 full; IG, 3.6 full; DG 4.1 full
    - Dutch: GS 3.5, 4.0 full; CG 4.0 full, Impress 3.6 full, DG 4.0 5 ch.
- A possibility to reuse the translated text in updates would be useful

Agenda

- Translating using OmegaT
- LO GUI strings in OmegaT
- Translation to language A using translation to language B
- Reusing non-OmegaT translations

OmegaT

- A Computer Aided Translation (CAT) tool
  - Java, open source, active development, large user community
  - http://omegat.org/
- Features (1)
  - Indirect translation using translation memory (TM)
    - Source (odt) split in segments (sentences)
    - Segments translated and translations stored in TM (xml file)
    - Translated document created from source and translated segments from the TM on demand
    - Advantage: Source file remains untouched

OmegaT

- Features (2)
  - Glossary of terms
    - Can hold translated GUI messages, translated chapter titles...
  - Spellchecker, grammar correction based on LanguageTool (https://languagetool.org/)
  - Similar translated segments offered for reuse
  - Machine translation possible (e. g. Google translate)
  - Collaboration of translators using git or subversion repositories (team project)
    - commits every few minutes to avoid double translations

Talk Assumptions

- The 'translate-toolkit' is installed from a repository or http://toolkit.translatehouse.org/
- The OmegaT tool is installed from http://wwww.omegat.org
    - The omegat package in Ubuntu repositories is outdated
- Python installed with lxml and goslate packages
- Examples shown for Linux
    - Perhaps they work on Mac too
    - On Windows: ??

The Basic Workflow (1)

- Install OmegaT
- Download the Guide chapters from https://wiki.documentfoundation.org/Documentation/Publications
- Start OmegaT
  - Create a new project GuideTrans: directory GuideTrans will be created
  - Set paths to spellchecker dictionaries
  - Create glossary with GUI translation
  - Import source files using the OmegaT GUI
    - Can be also copied manually to GuideTrans/source
    - Subdirectories in GuideTrans/source possible

The Basic Workflow (2)

- Start translating
    - Optionally set segment display and other preferences
- Generate translated files by choosing Project/Create translated files
    - Stored in GuideTrans/target
- Create screenshots, proofread
- Publish at the TDF wiki page and consider selling printed copies

Team Workflow with Remote Repository

- Create a *subversion* or *git* repository
  - We use code.google.com for that
- Create a project as earlier
  - Translate at least one segment (to create the TM file)
  - Delete some user specific files (more details)
  - Import it to the repository
- Translation using a team project
  - In OmegaT choose Project/Download Team Project
  - Work as usual, changes are committed periodically in background

The Problem: "Polluted" XML Code

- The XML code (content.xml) is 'polluted' by superfluous tags
  - Makes translation by OmegaT impossible
  - Solution proposed and a bug report filed
- A workaround:
  - A custom clean-up script to remove the useless tags

```
Original:
<f0>T</f0><f1>he </f1><i2/><f3>Menu bar </f3><f4>is
where</f4><f5> you </f5><f6>select</f6><f7> one of the menus
</f7><f8>and various </f8><f9> sub-menu</f9><f10>s</f10><f11>
appear </f11><f12>giving you more</f12><f13> options.
```

```
Cleaned:
The <i0/>Menu bar is where you select one of the menus and
various sub-menus appear giving you more options.
```

Cleaning the ODT Code

- The superfluous tags are in fact direct formatting tags:

  <text:span text:style-name="Txxx"> some text </text:span>

- The idea: remove direct formatting tags from the *content.xml* file
- The Guides frequently used 'useful' direct formatting
  - Manually converted to styles first
- The script:
  - Written in python using the lxml package
  - Not perfect, but usable
  - Freely available
  - Usage: `cleanodt.py -i infile.odt -o outfile.odt`
- The Getting Started 4.2 and Writer 4.2 guides available at TDF wiki have already been cleaned

Glossary with GUI Translation (1)

- Easy access to GUI translation helps to keep consistency and speeds up translation
- OmegaT glossary: a file with simple format

  source text TAB translated text

- Suggestions are displayed in a context menu
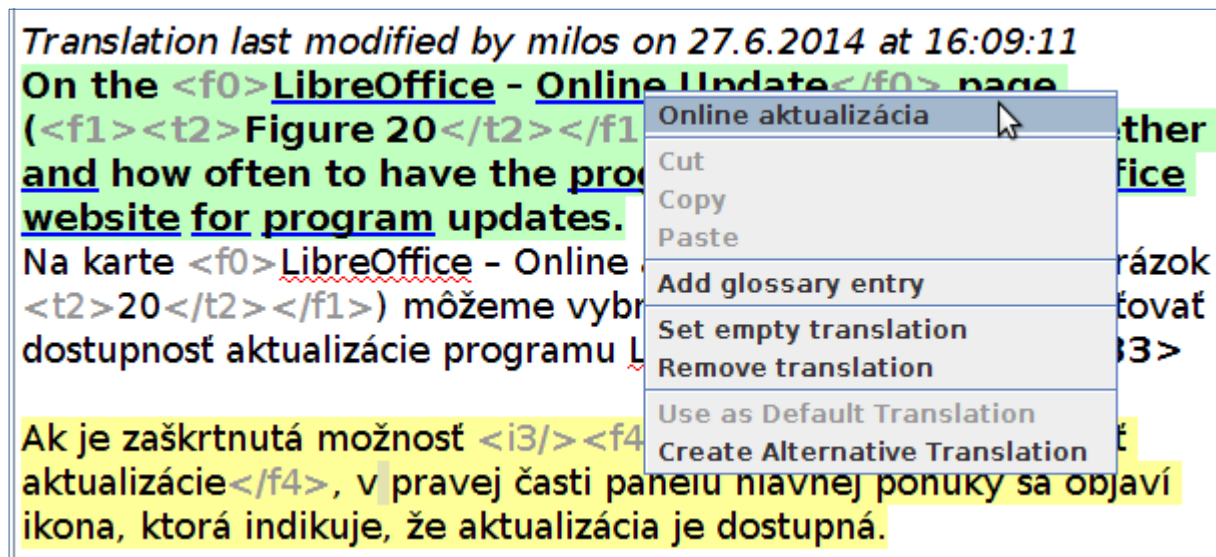
Glossary with GUI Translation (2)

- To create:
  - Download archive with GUI translation from the Pootle server at https://translations.documentfoundation.org/sk/libo_ui/ (replace 'sk' with your language code)
  - Unzip the archive into directory `podir`
    1. Make a single huge csv file:
        1. `po2csv -i podir -o csvdir`
        2. `cat `find csvdir -name *.csv` > lo.csv`
    2. Open lo.csv in LibreOffice and
        1. Delete the first column
        2. Save as 'text CSV' with tab as column separator
  - Copy the file to the GuideTrans/glossary directory
  - Optional: sort and delete long and duplicated segments

Translation Using a Third Language (1)

- OmegaT supports machine translation
  - May work poorly for your language
- Maybe a translation to a language exists, for which machine translation works better
- Tested on Czech > Slovak and Slovak > Czech
- A python script to translate tmx files written
  - Using 'goslate' package for that:

    ```
    tmxtrans -l lang -i imput.tmx -o output.tmx
    ```
  - Lang: output language code (input autodetected)

Translation Using a Third Language (2)

- Some postprocessing necessary
  - GT corrupts tag like strings
  - GT does not handle some features
- A `sed` script to correct errors in the translated text:
  - Example: <t1> 28 </ t1> </ f0>
  - Usage: `sed -e corr.sed input.tmx > output.tmx`
- A `python` script to handle features present in both texts, e.g. quotes:
  - English quotes: "text"
  - German, Slovak,... quotes: „text"
  - Usage: `tmxcorr.py -i infile -o outfile`
- Do not forget to check GUI translations using the glossary

- Direct usage of Google Translate supported by OmegaT
  - Drawbacks:
    - Corrupted tags, manual correction necessary
    - Using the API is not free (but also no expensive)
  - Indirect translation:
    - Correction of corrupted tags possible by a script
    - Free (as beer)
  - How to:
    - By pressing the Enter key copy the original text to the translated, repeat for all segments
      - Or: see OmegaT Console Mode
    - The rest: see instructions in Slide 15 and 16

Reusing Old 'Non-OmegaT' Translations (1)

- The idea: Create auxiliary TM files from the source and translated documents
  - Segment alignment necessary
- Store the TM files to the GuideTrans/tm directory
- The old translation appears as a suggestion in the 'Approximate translation' region
  - In OmegaT hit CTRL-R to use it

OmegaT tags should be preserved, so we use OmegaT for that:

- Clean formatting of both files first
- Extract sentences with OmegaT tags:

  1. Create a new OmegaT project Aux

  2. Adjust segment display to see only text

  3. Copy source and translated document to Aux/source

  4. For both files:

     1. Open the file

     2. Select all segments (only by mouse possible)

     3. Copy and paste to a new text file and save with 'txt' suffix

- Check line alignment, correct it if necessary, and export to a tmx file:

  - Use the LF_aligner tool: http://sourceforge.net/projects/aligner/

Copying Segments with Tags in OmegaT

Segments in English

Editor - en.odt

Translation last modified by milos on 25.8.2014 at 9:07:32

Getting Started Guide

Getting Started Guide****

Chapter `<f0><f1>`1 `<s2/></f1></f0><f3><br4/>`

Introducing LibreOffice`</f3>`

Copyright

This document is Copyright © 2010–2014 by the LibreOffice Documentation Team.

Contributors are listed below.

You may distribute or modify it under the terms of either the GNU General Public License (`<a0><f1>`http://www.gnu.org/licenses/gpl.html`</f1></a0>`), version 3 or later, or the Creative Commons Attribution License (`<a2><f3>`http://creativecommons.org/licenses/by/4.0/`</f3></a2>`), version 4.0 or later.

All trademarks within this guide belong to their legitimate owners.

Segments in the target language

Editor - sk.odt

Translation last modified by milos on 25.8.2014 at 9:07:32

Začíname s programom LibreOffice

Začíname s programom LibreOffice****

Kapitola `<f0><f1>`1 `<s2/></f1></f0><f3><br4/>`

Úvod do programu LibreOffice`</f3>`

Autorské práva

Tento dokument je duševným vlastníctvom dokumentačného tímu LibreOffice Copyright © 2014.

Zoznam prispievateľov a prekladateľov je uvedený nižšie.

Dokument je možné šíriť alebo upravovať v zmysle podmienok licencie GNU General Public License (`<a0><f1>`http://www.gnu.org/licenses/gpl.html`</f1></a0>`), verzia 3 alebo novšia, alebo v zmysle podmienok Creative Commons Attribution License (`<a2><f3>`http://creativecommons.org/licenses/by/4.0/`</f3></a2>`), verzia 4.0 alebo novšia.

Všetky ochranné známky uvedené v tejto príručke patria ich vlastníkom.

LibreOffice

Translation of the Getting Started Guide to Slovak and Czech (1)

- **Translation to Slovak:**
  - **Started with LO40 guide in August 2013**
    - **5 translated chapters for LO 3.5 existed**
  - **Status:**
    - **13 from 16 chapters published, 3 need proofreading**
    - **3 chapters translated using translation from Czech**
      - Speeds up translation by 75 %
    - **Screenshots: a 2 step process:**
      - Screenshots stored in and odg file (in repository)
      - Transfer of images from the odg file to chapter text document

- The team:
  - Translation: 2 persons
  - Screenshots: 2 persons
  - Proofreading (2-3 readings): 5 persons
  - Coordination, repository administration, final touches: 1 person
  - Translation to Czech: a few chapters translated now

Thank you … for considering translation of LO Guides!

- The scripts:
- The google code project (Slovak only): https://code.google.com/p/sk-libreoffice-guides/
- Help always online: milos.sramek (at) soit.sk