Wikipedia2TBX: creation of terminological glossaries from Wikipedia

1.Introduction

Wikipedia (http:wikipedia.org) is a free encyclopedia created in a collaborative way. Wikipedia is a multilingual encyclopedia, ie, there are versions of Wikipedia for a large number of languages. For each entry in the Wikipedia crosslinguistic links are provided that lead directly to the same article in another language. Entries are also organized by categories. With all this information we can create multilingual terminological glossaries from Wikipedia.

The program we present in this document is capable of creating multilingual glossaries from Wikipedia. To create them, the program needs to know the subject area or areas of interest and the languages involved. With these data, the program will look at all the entries in the English Wikipedia which are assigned the given thematic areas and see if they have interlingual links to the required language. From the collected data the program creates a multilingual glossary of terms in TBX*(Term Base Exchange*) format. A glossary in tabular format for OmegaT can be also created. The glossaries in OmegaT format are bilingual and therefore the program only takes into account the first two given languages, the first one as the source language, and the second one as the target language. All other languages are ignored to create the glossary in tabular format.

In this document we will explain how to use the program Wikipedia2TBX. A graphical user interface is provided, so it's very easy to use the program. We will also explain which language codes are used, and how to know the subject classifications of Wikipedia.

Before using the tool, we must bear in mind the following:

- The glossaries are created based on information from Wikipedia. Remember that Wikipedia is a collaborative project where anyone can add information. The richness of the project is precisely its collaborativeness. The quality of this resource is very good, but all results must be understood as a proposal and therefore is up to the translator to use or not a specific proposal.
- The program runs against a copy of the English Wikipedia hosted on a server at the Open University of Catalonia. This copy is regularly updated with the latest versions of the English Wikipedia, but the content may differ slightly with actual Wikipedia.

2. Obtaining, installing and running the program

Wikipedia2TBX is a free and open source program (under GNU-GPL license) and can be downloaded from http://lpg.uoc.edu/Wikipedia2TBX. The program is written in Python and is therefore cross-platform. We provide 2 versions of the program:

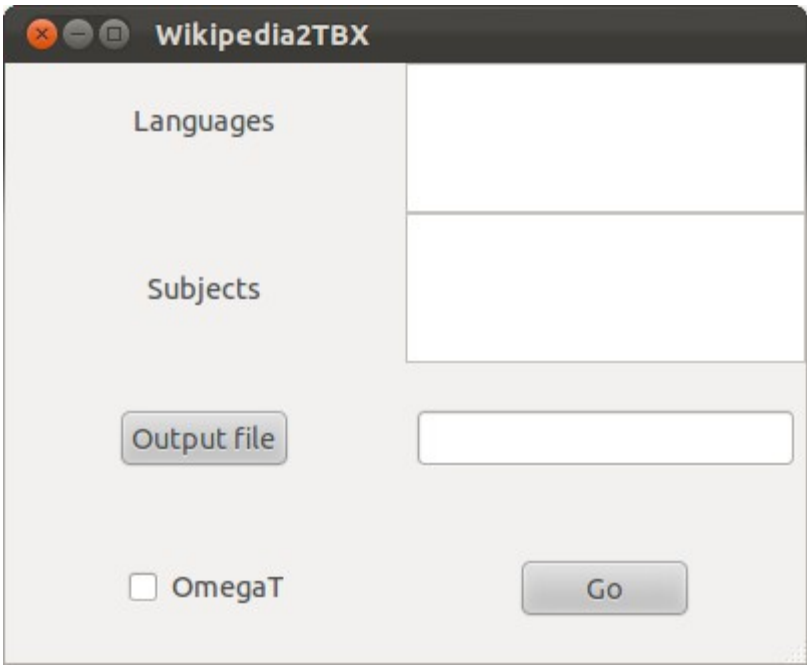
- Source files (Wikipedia2TBX-0.1.tar-gz): to run it is necessary to have a Python interpreter installed on your system. This is the standard on Linux and Mac and therefore users of these operating systems can use this version directly provided the prerequisites are installed. If you want to use this version under Windows you need to install a Python interpreter (version lower than 3.0) which can be freely downloaded from www.python.org
- Executable version for Windows (Wikipedia2TBX-0.1-win.zip). This version can run under Windows without the Python interpreter installed on your system.

The program doesn't need to be installed, simply unzip the downloaded file and double click the file or Wikipedia2TBX.py or Wikipedia2TBX.exe. In all operating systems, if the Python interpreter is installed, the program can be run writing in the command prompt:

```
python Wikipedia2TBX.py
```

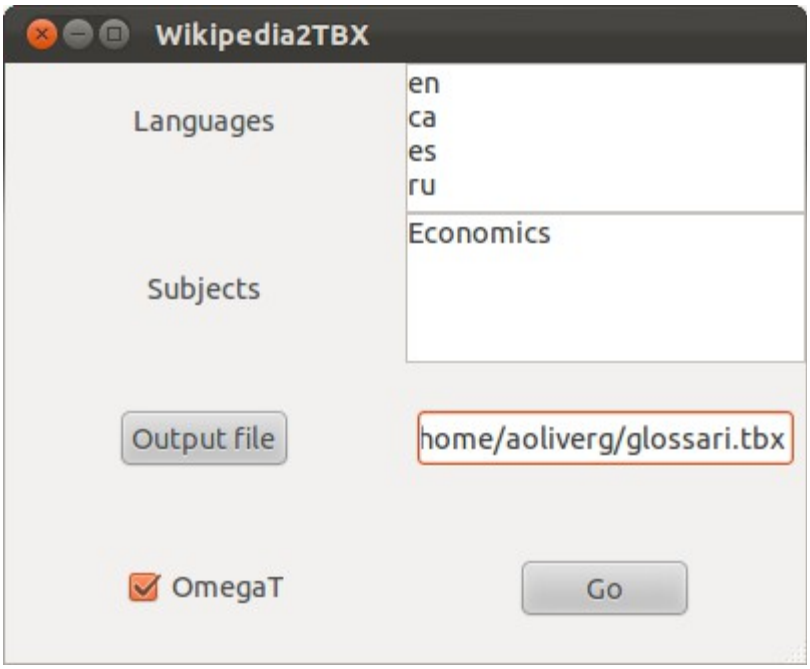
3. Running the program

When clicking on Wikipedia2TBX.py or Wikipedia2TBX.exe a simple graphical interface will open.



In this interface we have to:

- In the text box **Languages** we have to write the language codes (ISO 2-letter code, see section 4 below). At least we have to write two language codes. A code per line must be written.
- In the text box **Subjects** we have to write the subject areas (they should be expressed in English, see section 5 of this document). There must be an area by line.
- We have to select the location and name of the output file clicking at the button. **OutputFile.** It is not necessary to specify the extension of the file because the extension .tbx will automatically assigned.
- Check the selection box **OmegaT** if you want to get the output also in tabular format for OmegaT. In this case, the program will create a file in the same location and with the same name as the output file, but with. .utf8 extension.
- Once all selections have been done, click the button. **Go** to start the creation of the glossary. It might take a while and it is necessary to be connected to the Internet.



Here we can see a fragment of the output file in TBX:

```
<?xml version="1.0" ?>
<martif type="TBX" xml:lang="en">
  <text>
    <body>
      <termEntry id="1">
        <descrip type="subjectField">
          Economics
        </descrip>
           <langSet xml:lang="en">
             <tig>
               <term>
                 Game theory
               </term>
             </tig>
           </langSet>
           <langSet xml:lang="ca">
             <tig>
               <term>
                 Teoria dels jocs
               </term>
             </tig>
           </langSet>
           <langSet xml:lang="es">
             <tig>
               <term>
```

```
                Teoría de juegos
              </term>
            </tig>
          </langSet>
          <langSet xml:lang="ru">
            <tig>
              <term>
                Теория игр
              </term>
            </tig>
          </langSet>
        </termEntry>
```

and in tabular format for OmegaT:

```
Game theory  Teoria dels jocs    Economics
Human rights Drets Humans Economics
Smuggling    Contraban    Economics
Means of production Mitjans de producció      Economics
Poverty      Pobresa      Economics
Innovation   Innovació    Economics
Optimism     Optimisme    Economics
Millionaire  Milionari    Economics
Liquidity trap      Trampa de liquiditat      Economics
Break-even   Punt mort (economia)      Economics
```

4. Language Codes

Two-letter ISO language code code (ISO 639-1) must be used. These codes are available at : http://www.sil.org/iso639-3/codes.asp?order=639_1&letter =% 25

5. List of subjects

The subjects to be used are those of Wikipedia and must be expressed in English. You must keep in mind, however, that subject areas of Wikipedia are freely selected by the user, ie, any user can create a new area. In order to know which subject areas are available is useful to consult the following link: http://en.wikipedia.org/wiki/List_of_academic_disciplines

6. Conclusions

With Wikipedia2TBX we can quickly create glossaries for a large number of languages and subject areas. These glossaries are to be used with caution, but they can be a great resource to to be used along with traditional terminology resources.