TONO2TBX: Conversión del formato de archivo de Terminologia Oberta de TermCat

1.Introducción

El TERMCAT (www.termcat.cat) es el centro de terminología de la lengua catalana, creado el 1985 por la Generalitat de Catalunya y el Institut d'Estudis Catalans. El TERMCAT tiene como misión garantizar el desarrollo y la integración de la terminología catalana en los sectores especializados y en la sociedad en general, mediante la creación continua de herramientas y de recursos innovadores y de calidad, en un diálogo permanente con especialistas y usuarios. Entre los diferentes productos que desarrolla se destaca la Terminologia Oberta (http://www.termcat.cat/productes/toberta.htm), que son unas compilaciones terminológicas multilingües (normalmente incluyen catalán, castellano e inglés, y a menudo otras lenguas como el francés o el alemán) que se pueden descargar y se distribuyen bajo una licencia Creative Commons.

Los formatos en que te puedes descargar los ficheros no incluyen por ahora ni el formato estándar TBX ni un fichero de texto tabulado, cosa que dificulta el uso de los glosarios en herramientas de traducción asistida.

En este documento presentamos una sencilla aplicación que nos permite transformar el formato XML propio de la Terminologia Oberta del TermCat en ficheros TBX y texto tabulado. Con esta herramienta pretendemos fomentar el uso de estos glosarios dentro de las herramientas de traducción asistida.

2. Obtención, instalación y ejecución del programa

El programa es libre (bajo licencia GNU-GPL) y se puede descargar de http://lpg.uoc.edu/TO2TBX. El programa está escrito en Python y es por lo tanto multiplataforma. Se distribuyen 2 versiones del programa:

- Archivos fuente (TO2TBX-0.1.tar.gz): para ejecutarlo es necesario tener instalado un intérprete de Python en el sistema. Este hecho es el habitual en Linux y Mac y por lo tanto se podrá usar esta versión directamente bajo estos sistemas operativos si se dispone además de los prerequisitos. Si se quiere usar esta versión bajo Windows será neceraio instalar un intérprete de Python (versión inferior a la 3.0) que se puede descargar libremente de www.python.org
- Versión ejecutable por Windows (TO2TBX-0.1-win.zip). Esta versión se puede ejecutar bajo Windows sin necesidad de tener el intérprete de Python instalado al sistema.

El programa no hay que instalarlo, simplemente hay que descomprimir el archivo descargado y hacer doble clic sobre el archivo Wikipedia2TBX.py o Wikipedia2TBX.exe (para Windows). En los tres sistemas operativos, si se dispone del intérprete de Python, se puede ejecutar desde línea de comandos haciendo:

```
python TO2TBX.py
```

3. Funcionamiento del programa

Una vez hacemos doble clic sobre TO2TBX.py o TO2TBX.exe se abre una interfaz gráfica muy sencilla.

En esta interfaz tenemos que:

- Seleccionar el fichero xml de Terminologia Oberta del TermCat con el botón **Input file**
- El programa seleccionará automáticamente el nombre del fichero de salida (el mismo que el de entrada pero con extensión tbx y utf8) y el directorio de salida será el mismo que el de entrada. Podemos modificar esto con el botón **Output file**
- En el cuadro de texto **Available languages** aparecerán los códigos de las lenguas presentes en el fichero de entrada
- En el cuadro de texto **Selected languages** tenemos que seleccionar los códigos de las lenguas que queremos que aparezcan en el fichero de salida. Hay que tener en cuenta que si queremos un fichero tabulado para OmegaT, como que sólo puede contener dos lenguas, considerará que la lengua de partida es la correspondiente al primer código y la de llegada la correspondiente al segundo código. Si no indicamos nada en este cuadro de texto, se seleccionarán todas las lenguas presentes en Available languages.
- Si marcamos la casilla de selección OmegaT se creará un fichero de texto tabulado para OmegaT, con extensión utf8.

En la figura anterior podemos observar el programa con unas selecciones. El programa tratará el fichero TO_Videojuegos.xml de TermCat y lo transformará a TBX en un fichero llamado TO_Videojuegos.tbx. Este fichero tiene como lenguas el catalán (ca), inglés (en), alemán (de), francés (fr) y castellano (es). En Selected languages los hemos seleccionado todos pero en otro orden porque queremos también un fichero tabulado inglés-castellano.

A continuación podemos observar un fragmento de la salida en TBX:

```
<?xml version="1.0" ?>
<martif type="TBX" xml:lang="en">
  <text>
    <body>
      <termEntry id="1">
        <descrip type="subjectField">
          Dispositius de joc
        </descrip>
```

```
          <langSet xml:lang="en">
            <tig>
              <term>
                physical acceleration
              </term>
            </tig>
          </langSet>
          <langSet xml:lang="ca">
            <tig>
              <term>
                acceleració física
              </term>
            </tig>
          </langSet>
          <langSet xml:lang="es">
            <tig>
              <term>
                aceleración física
              </term>
            </tig>
          </langSet>
          <langSet xml:lang="fr">
            <tig>
              <term>
                accélération physique
              </term>
            </tig>
          </langSet>
          <langSet xml:lang="de">
            <tig>
              <term>
                Physik-Beschleunigung
              </term>
            </tig>
          </langSet>
        </termEntry>
      <termEntry id="2">
....
```

y en formado tabular por OmegaT:

```
physical acceleration      acceleració física  Dispositius de joc
hardware acceleration      acceleració per maquinari  Dispositius de joc
video game addiction       addicció als videojocs     Interacció i comunitat
addictive     addictiu -iva Interacció i comunitat
gamemaster    administrador -a de joc    Interacció i comunitat
```

4. Códigos de lengua

Los códigos de lengua que se usan son los ISO de 2 letras que podéis encontrar aquí: http://www.sil.org/iso639-3/codas.asp?order=639_1&letter =% 25

6. Conclusiones

Con TO2TBX podemos aprovechar los glosarios de Terminologia Oberta del TermCat en nuestros proyectos de traducción. Cómo que ahora podemos disponer de formatos estándar, podremos usar estos glosarios con prácticamente cualquier herramienta de traducción asistida.