TO2TBX: Conversió del format d'arxiu de Terminologia Oberta de TermCat

1.Introducció

El TERMCAT (www.termcat.cat) és el centre de terminologia de la llengua catalana, creat el 1985 per la Generalitat de Catalunya i l'Institut d'Estudis Catalans. El TERMCAT té com a missió garantir el desenvolupament i la integració de la terminologia catalana en els sectors especialitzats i en la societat en general, mitjançant la creació contínua d'eines i de recursos innovadors i de qualitat, en un diàleg permanent amb especialistes i usuaris. Entre els diferents productes que desenvolupa es destaca la Terminologia Oberta (http://www.termcat.cat/productes/toberta.htm), que són uns reculls terminològics multilingües (normalment inclouen català, castellà i anglès, i sovint altres llengües com el francès o l'alemany) que es poden descarregar i es distribueixen sota una llicència Creative Commons.

Els formats en què et pots descarregar els fitxers no inclou per ara ni el format estàndard TBX ni un fitxer de text tabulat, cosa que dificulta l'ús dels glossaris en eines de traducció assistida.

En aquest document presentem una senzilla aplicació que ens permet transformar el format XML propi de la Terminologia Oberta del TermCat en fitxers TBX i text tabulat. Amb aquesta eina pretenem fomentar l'ús d'aquests glossaris dins de les eines de traducció assistida.

2. Obtenció, instal·lació i execució del programa

El programa és lliure (sota llicència GNU-GPL) i es pot descarregar de http://lpg.uoc.edu/TO2TBX. El programa està escrit en Python i és per tant multiplataforma. Es distribueixen 2 versions del programa:

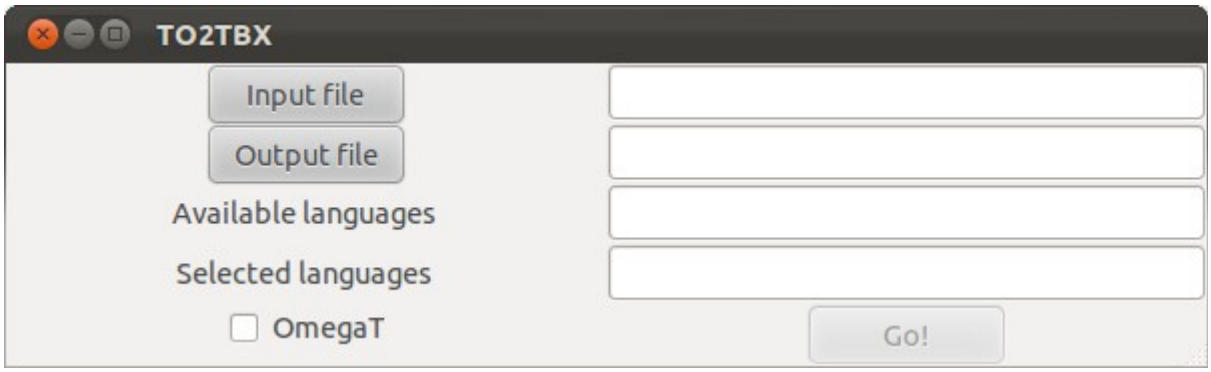
- Arxius font (TO2TBX-0.1.tar.gz): per executar-lo és necessari tenir instal·lat un intèrpret de Python en el sistema. Aquest fet és l'habitual en Linux i Mac i per tant es podrà fer servir aquesta versió directament sota aquests sistemes operatius si es disposa a més dels prerequisits. Si es vol fer servir aquesta versió sota Windows caldrà instal·lar un intèrpret de Python (versió inferior a la 3.0) que es pot descarregar lliurement de www.python.org
- Versió executable per Windows (TO2TBX-0.1-win.zip). Aquesta versió es pot executar sota Windows sense necessitat de tenir l'intèrpret de Python instal·lat al sistema.

El programa no cal instal·lar-lo, simplement cal descomprimir l'arxiu descarregat i fer doble clic sobre l'arxiu Wikipedia2TBX.py o Wikipedia2TBX.exe (per Windows). En tots tres sistemes operatius, si es disposa de l'intèrpret de Python, es pot executar des de línia de comandes fent:

```
python TO2TBX.pu
```

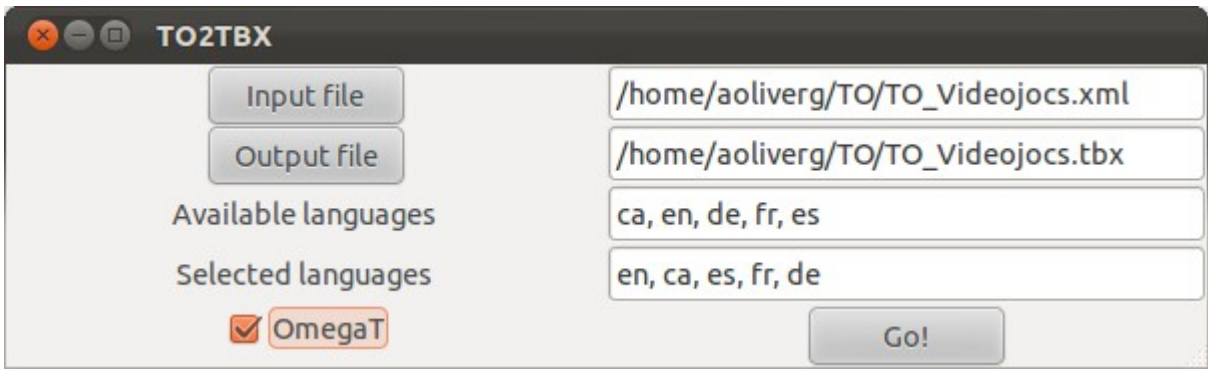
3. Funcionament del programa

Un cop fem doble clic sobre TO2TBX.py o TO2TBX.exe s'obre una interfície gràfica molt senzilla.



En aquesta interfície hem de:

- Seleccionar el fitxer xml de Terminologia Oberta del TermCat amb el botó **Input file**
- El programa seleccionarà automàticament el nom del fitxer de sortida (el mateix que el d'entrada però amb extensió tbx i utf8) i el directori de sortida serà el mateix que el d'entrada. Podem modificar això amb el botó **Output file**
- En el quadre de text **Available languages** apareixeran els codis de les llengües presents en el fitxer d'entrada
- En el quadre de text **Selected languages** hem de seleccionar els codis de les llengües que volem que apareguin en el fitxer de sortida. Cal tenir en compte que si volem un fitxer tabulat per a OmegaT, com que només pot contenir dues llengües, considerarà que la llengua de partida és la corresponent al primer codi i la d'arribada la corresponent al segon codi. Si no indiquem res en aquest quadre de text, es seleccionaran totes les llengües presents en Available languages.
- Si marquem la casella de selecció OmegaT es crearà un fitxer de text tabulat per a OmegaT, amb extensió utf8.



En la figura anterior podem observar el programa amb unes seleccions. El programa tractarà el fitxer TO_Videojocs.xml de TermCat i el transformarà a TBX en un fitxer anomenat TO_Videojocs.tbx. Aquest fitxer té com a llengües el català (ca), anglès (en), alemany (de), francès (fr) i castellà (es). En Selected languages els hem seleccionat tots però en un altre ordre perquè volem també un fitxer tabulat anglès-castellà.

A continuació podem observar un fragment de la sortida 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">
....
```

i en format tabular per 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. Codis de llengua

Els codis de llengua que es fan servir són els ISO de 2 lletres que podeu trobar aquí: http://www.sil.org/iso639-3/codes.asp?order=639_1&letter =% 25

6. Conclusions

Amb TO2TBX podem aprofitar els glossaris de Terminologia Oberta del TermCat als nostres projectes de traducció. Com que ara podem disposar de formats estàndard, podrem fer servir aquests glossaris amb pràcticament qualsevol eina de traducció assistida.