Projectes de traducció i formats estàndard: TMX, TBX, XLIFF i SRX

Antoni Oliver

P08/C0240/00289

UOC
www.uoc.edu

Índex

Objectius

1. Conèixer a fons els formats estàndard basats en XML que es fan servir en el món de la traducció: TMX, TBX, XLIFF i SRX.

2. Analitzar l'acceptació d'aquests formats per part de les principals eines de traducció assistida del mercat.

3. Valorar l'ús d'aquests formats en la gestió de projectes de traducció.

1. Els formats estàndard basats en XML d'ús en traducció

El llenguatge XML està prenent cada dia més importància en la majoria d'àmbits. Hi ha una gran quantitat de formats estàndard que es basen en XML. En el món de la traducció existeixen uns formats estàndard basats en XML per compartir memòries de traducció, bases de dades terminològiques, regles de segmentació i projectes de traducció i localització. En aquest apartat presentarem breument cada un d'aquest formats.

1.1. TMX (Translation Memory eXchange)

El TMX (*Translation Memory eXchange*) es un format estàndard basat en XML que serveix per compartir memòries de traducció. Mitjançant aquest format podem fer servir una memòria creada per una eina A en un eina B, si és que l'eina A disposa d'una utilitat d'exportació a TMX i l'eina B en disposa d'una d'importació.

1.2. TBX (Term Base eXchange)

El TBX (*Term Base eXchange)* és un format estàndard basat en XML que serveix per compartir bases de dades terminològiques.

1.3. SRX (Segmentation Rule eXchange)

L'SRX (*Segmentation Rule eXchange)* és un format estàndard basat en XML que serveix per compartir regles de segmentació. Els programes de traducció assistida fan servir una sèrie de regles de segmentació per a dividir el text a traduir en segments i tractar i presentar cada un d'aquests segments de forma separada. El format SRX ens servirà per compartir aquestes regles de segmentació i assegurar-nos que dues eines de traducció diferents divideixin un mateix text d'entrada en els mateixos segments. Això pot ser important si estem fent servir una memòria de traducció que s'ha generat traduint amb una eina A fent servir unes regles de segmentat A. Si ara volem aprofitar aquesta memòria amb una altra eina B ens interessarà fer servir les mateixes regles de segmentació, ja que d'aquesta manera la probabilitat de trobar coincidències a la memòria augmenta.

1.4. XLIFF (Localisation Interchange File Format)

L'XLIFF (*Localisation Interchange File Format*) és un format estàndard basat en XML per a l'intercanvi de projectes de traducció i localització. Mitjançant aquest format es pot traduir amb una eina B un projecte creat amb una eina A.

2. Formats estàndard i eines de traducció assistida

No tots els formats estàndard presentats a l'apartat anterior gaudeixen del mateix grau d'integració a les eines de traducció assistida del mercat. Dels formats esmentats, el que gaudeix d'un més alt nivell d'integració és el TMX, que és suportat per la immensa majoria de les eines de traducció assistida. El segon lloc possiblement l'ocuparia el TBX, però a molta distància. L'SRX no gaudeix pràcticament de cap mena d'integració.

Un cas especial el constituiria l'XLIFF. Hi ha una sèrie d'eines de traducció assistida que són en realitat editors d'XLIFF. Per una altra banda, les eines que no tenen suport per XLIFF però que permeten crear filtres per XML potencialment poden traduir fitxers XLIFF.

3. Eines gratuïtes per treballar amb formats estàndard

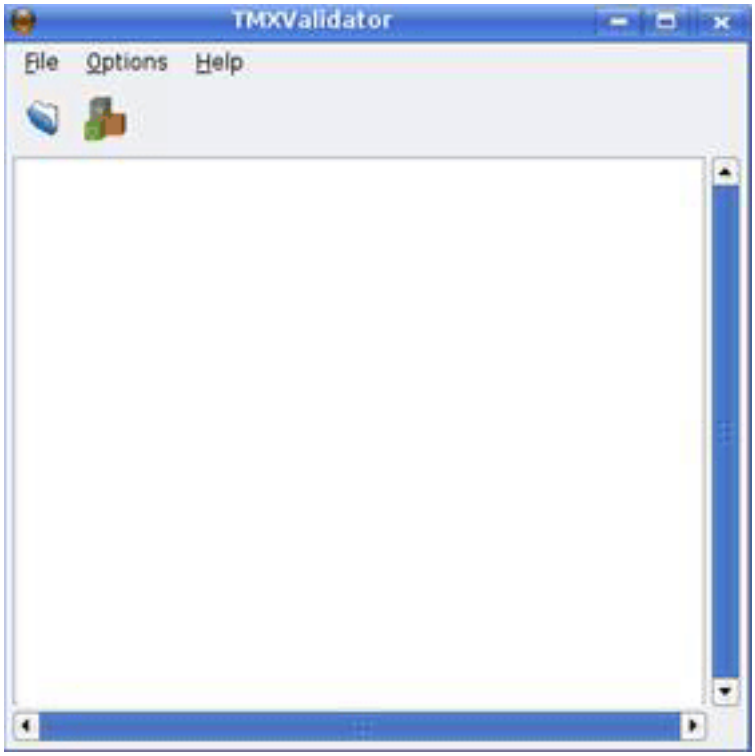
En aquest apartat presentarem una sèrie d'eines gratuïtes que ens permetran treballar còmodament amb alguns dels formats estàndard.

3.1. Eines per treballar amb TMX

La majoria d'eines de traducció assistida, tant comercials com gratuïtes, ens permeten treballar amb el format TMX d'intercanvi de memòries de traducció. No tindrem gaire complicació per exportar i importar memòries de traducció en aquest format. D'aquesta manera l'intercanvi de memòries de traducció entre diferents eines de traducció assistida és una tasca fàcil.

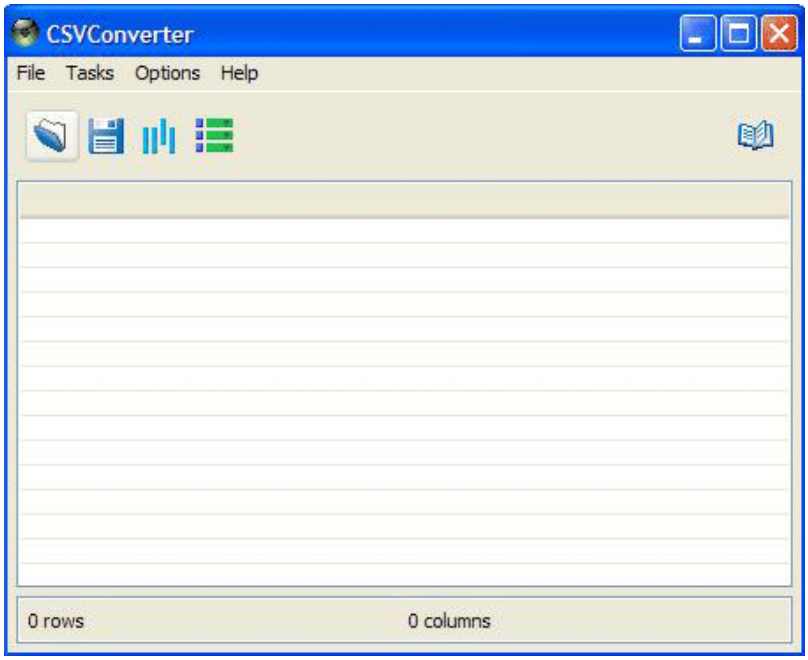
3.1.1. TMX Validator

TMX Validator és una aplicació gratuïta de l'empresa MaxPrograms que es pot descarregar de http://www.maxprograms.com/freetools.html. Aquesta aplicació funciona sota Windows i sota Linux. Aquesta aplicació serveix per validar si un determinat arxiu TMX és correcte o conté algun tipus d'error. Aquest aplicació pot ser útil per comprovar que les memòries en TMX que rebem o que enviem siguin realment correctes. La interfície visual és molt simple:



3.1.2. CSV Converter

CSV Converter és una aplicació gratuïta de l'empresa MaxPrograms que es pot descarregar de http://www.maxprograms.com/freetools.html. Aquesta aplicació funciona sota Windows i sota Linux. Aquesta utilitat permet convertir arxius CSV (fitxers de text separats per coma o per altres separadors) a fitxers TMX. Si executem el programa ens apareix una pantalla d'inici com la següent:



La interfície d'usuari està en diversos idiomes, entre ells el castellà. Si volem canviar l'idioma de la interfície només caldrà que fem clic al menú *Options* i triem l'idioma que vulguem. L'explicació que presento a continuació la faré amb la interfície en anglès.

Si volem importar un fitxer CSV hem de fer *File* > *Open CSV File* i ens apareixerà una pantalla com la següent:

En aquesta pantalla hem de seleccionar:

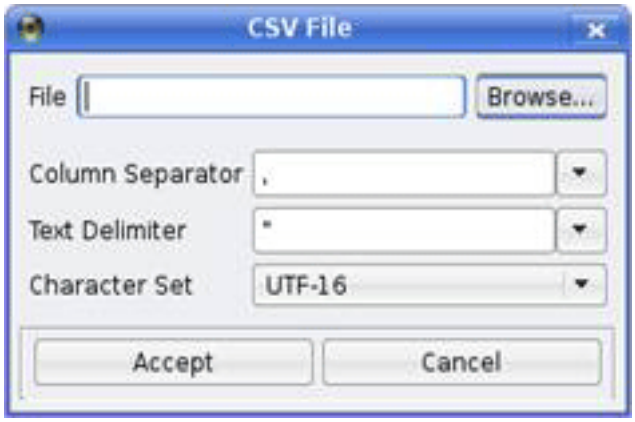
- El fitxer CSV que volem importar, amb el botó *Browse*

- El separador de columnes, mitjançant la llista desplegable *Column separator*. Per exemple, si volem importar un fitxer de text separat per tabuladors haurem de seleccionar el separador de columnes *Tab*.

- El delimitador de text, mitjançant la llista desplegable *Text deliminter*

- La codificació de caràcters mitjançant la llista desplegable *Character set*

És imprescindible disposar de tota aquesta informació de l'arxiu que volem importar. Si no coneixem aquesta informació el més aconsellable és obrir aquest arxiu amb un bon editor de textos i observar-lo nosaltres mateixos. Si fem la importació correctament ens apareixerà una cosa del estil:



Un cop importat l'arxiu, si volem eliminar alguna de les columnes, podem fer servir aquest botó:

Si fem clic sobre aquest botó ens apareixerà una pantalla com la següent:

On podrem seleccionar una o més columnes per eliminar. Ara, el darrer pas serà indicar la llengua corresponent a cada una de les columnes, mitjançant aquest botó:

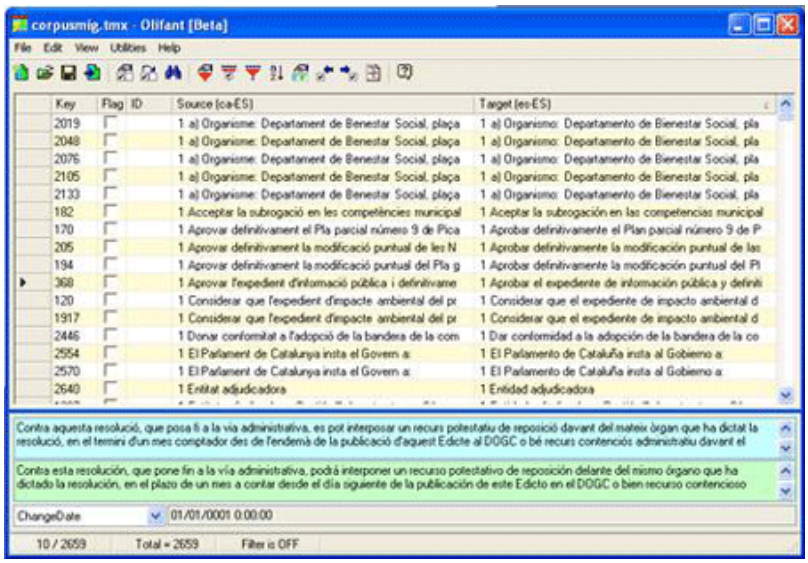
Fent clic sobre aquest botó ens apareixerà una pantalla com la següent, des d'on podrem seleccionar les llengües.

Un cop fet això, podrem exportar la memòria en TMX fent *File* > *Export as TMX*.

3.1.3. Olifant d'ENLASO Tools

Olifant és una aplicació del conjunt d'eines gratuïtes ENLASO Tools. Aquestes eines es poden descarregar de http://www.translate.com/technology/tools/ . Aquesta eina funciona sota Windows si teniu l'entorn .NET instal·lat. L'eina permet fer el manteniment de memòries de traducció, així com importar i exportar memòries des de i cap a diversos formats.

L'eina té el següent aspecte:



El funcionament de l'eina és força intuïtiu:

- Si volem importar una memòria de traducció farem *File* > *Open* i podrem triar entre els formats TMX, Wordfast TM File, Trados Text TM Files i Olifant TI Files

- Si volem exportar la memòria que tenim carregada simplement haurem de fer *File* > *Save as* i triar també entre un dels formats esmentats

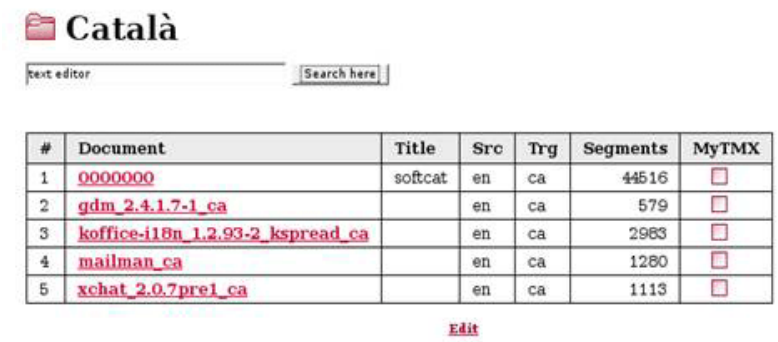
- A part de les funcions d'obrir i guardar també disposem de les d'importació i exportació. La funció d'exportació permet exportar part de la memòria, en funció d'uns filtres que es poden definir

- Podem editar les entrades de la memòria per corregir errors o afegir informació rellevant

- Podem fer cerques dins de la memòria, tant pels segments originals i traduïts com per la resta d'informació de la memòria

3.1.4. Tumatxa

Tumatxa (www.tumatxa.com) és un gestor web de memòries de traducció desenvolupat per l'empresa basca CodeSyntax i distribuït com a programari lliure. Aquesta aplicació permet emmagatzemar memòries de traducció en un repositori web i fer cerques en les memòries. Permet treballar tant en format TMX com en format PO. Es poden fer cerques sobre les memòries de traducció i seleccionar una o més memòries de traducció per descarregar al nostre ordinador. Des de la plana web d'aquest producte es pot accedir a unes demostracions. Presento a continuació una sèrie de captures de pantalla que intenten explicar les principals funcionalitats:



Català

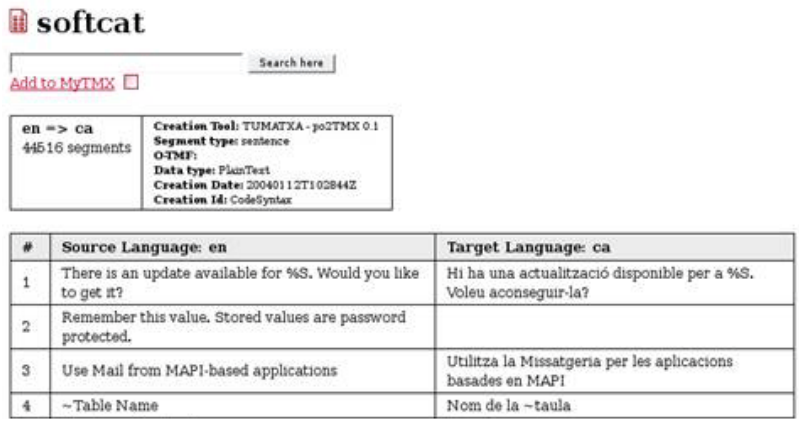
text editor Search here

#	Document	Title	Src	Trg	Segments	MyTMX
1	0000000	softcat	en	ca	44516	☐
2	gdm_2.4.1.7-1_ca		en	ca	579	☐
3	koffice-i18n_1.2.93-2_kspread_ca		en	ca	2983	☐
4	mailman_ca		en	ca	1280	☐
5	xchat_2.0.7pre1_ca		en	ca	1113	☐

Edit

En la pantalla inicial se'ns mostra les memòries disponibles. Des d'aquesta pantalla podem marcar memòries per descarregar (fent servir les caselles de selecció de la columna *MyTMX*) o bé fer una cerca posant el text a cercar en el quadre de text i fent clic al botó *Search here*. També podem fer clic sobre el nom de la memòria i ens mostrarà informació rellevant i el seu contingut:

A continuació presentem la pantalla que mostra la informació d'una memòria:



softcat

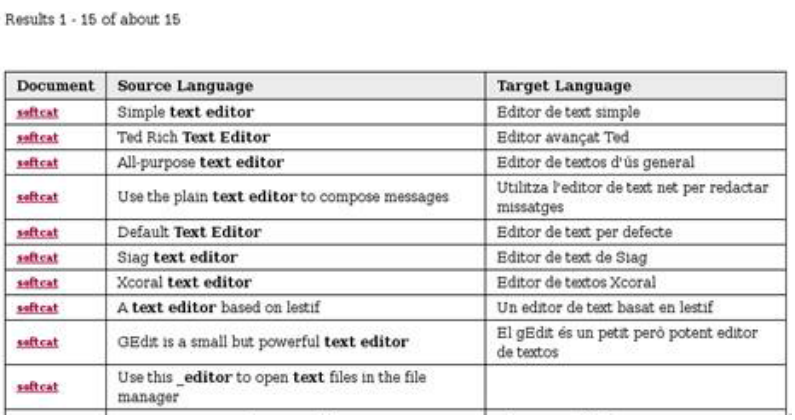
Search here

Add to MyTMX ☐

en => ca 44516 segments	**Creation Tool:** TUMATXA - po2TMX 0.1 **Segment type:** sentence **O-TMF:** **Data type:** PlainText **Creation Date:** 20040112T102844Z **Creation Id:** CodeSyntax

#	Source Language: en	Target Language: ca
1	There is an update available for %S. Would you like to get it?	Hi ha una actualització disponible per a %S. Voleu aconseguir-la?
2	Remember this value. Stored values are password protected.	
3	Use Mail from MAPI-based applications	Utilitza la Missatgeria per les aplicacions basades en MAPI
4	~Table Name	Nom de la ~taula

Si fem servir la funcionalitat de cerca ens mostra tots els segments (de totes les memòries si la cerca la fem des de la pantalla principal, o d'una memòria en concret si la cerca la fem des de la pantalla de presentació d'una memòria) que conté la cadena de cerca:



Results 1 - 15 of about 15

Document	Source Language	Target Language
softcat	Simple **text editor**	Editor de text simple
softcat	Ted Rich **Text Editor**	Editor avançat Ted
softcat	All-purpose **text editor**	Editor de textos d'ús general
softcat	Use the plain **text editor** to compose messages	Utilitza l'editor de text net per redactar missatges
softcat	Default **Text Editor**	Editor de text per defecte
softcat	Siag **text editor**	Editor de text de Siag
softcat	Xcoral **text editor**	Editor de textos Xcoral
softcat	A **text editor** based on lestif	Un editor de text basat en lestif
softcat	GEdit is a small but powerful **text editor**	El gEdit és un petit però potent editor de textos
softcat	Use this **_editor** to open **text** files in the file manager	

Aquesta aplicació pot ser de molta utilitat en diverses situacions:

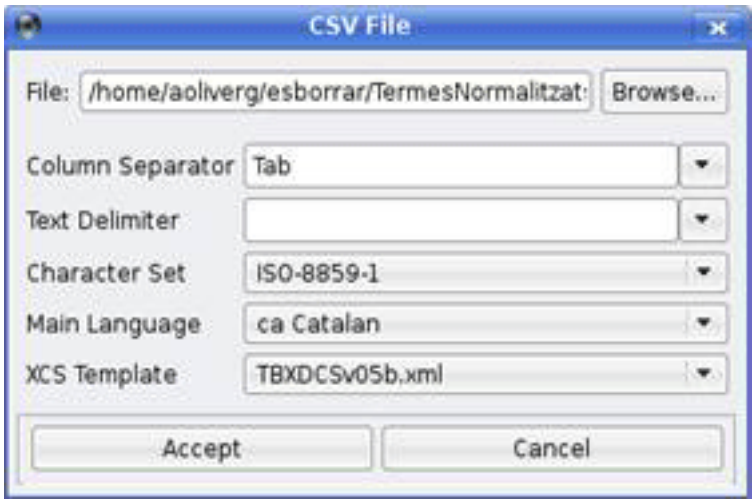
- Per a una empresa de traduccions que vulgui donar accés als clients a les seves memòries de traducció
- Per a una empresa de traducció que vulgui donar accés a les memòries de traducció d'un determinat projecte a tots els participants en aquest projecte
- Per crear un repositori públic de memòries de traducció
- Per a localització de projectes de programari lliure

3.2. Eines per treballar amb TBX

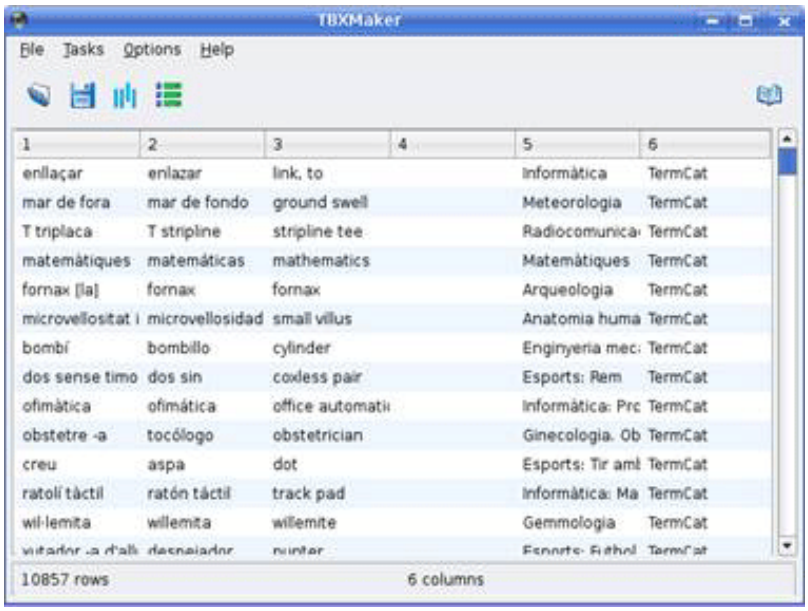
3.2.1. TBX Maker

TBX Maker és una aplicació gratuïta de l'empresa MaxPrograms que es pot descarregar de http://www.maxprograms.com/freetools.html. Aquesta aplicació funciona sota Windows i sota Linux. El funcionament és molt similar al de l'aplicació CSV Converter que hem explicat abans. El punt on es diferencien bàsicament les dues aplicacions és el següent: en la pantalla on carreguem l'arxiu que volem transformar podem triar una plantilla determinada que ens marcarà els atributs que podrem seleccionar per a cada camp. A continuació presento una petita explicació del funcionament del programa.

El primer que cal fer és seleccionar l'arxiu que volem importar fent *File* > *Open CSV File* i ens apareix una pantalla com la següent:



En aquesta pantalla seleccionarem l'arxiu que volem obrir, el separador de columnes, el delimitador de text, la codificació de caràcters, la llengua principal i la plantilla XSC. Explicarem amb més detall el tema de la plantilla més endavant, ara podeu acceptar la que us aparegui per defecte. Si fem clic en el botó *Accept* ens apareixerà una pantalla com la següent:



Ara haurem d'eliminar, si cal, les columnes innecessàries, amb el botó.

Un cop eliminades les columnes que no calguin, amb el botó

assignarem la categoria de cada columna. La quantitat de categories que es poden seleccionar depèn de la plantilla triada. Haurà d'aparèixer una pantalla com la següent:

És en aquesta pantalla on haurem d'anar indicant quina informació conté cada columna. La informació disponible dependrà de la plantilla que haguem triat a la pantalla inicial. Com que l'operació de seleccionar els atributs pot ser pesada, si preveiem que hem de tractar més fitxers iguals, podem guardar la configuració amb el botó *Save Configuration*. Si més endavant hem de tornar a tractar un fitxer igual podrem carregar aquesta configuració i ens estalviarem la feina d'anar triant els atributs.

Un cop triada la informació de cada columna podem exportar l'arxiu fent *File > Export as TBX* i ens apareixerà una pantalla com la següent que ens permetrà triar el nom i la ubicació de l'arxiu TBX.

Per poder utilitzar aquest programa amb èxit cal conèixer l'estructura del fitxer que volem exportar a TBX i triar la plantilla adequada a la informació que conté l'arxiu. La documentació del programa no explica el contingut de cada plantilla i el millor és fer una prova preliminar amb cada una de les plantilles per veure quina és la que s'adapta millor a les nostres necessitats.

3.3. Eines per treballar amb SRX

No hi ha eines específiques per treballar amb SRX. Algunes eines de traducció assistida ja suporten aquest format, entre elles SDL, Heartsome i Trados.

3.4. Eines per treballar amb XLIFF

Les eines gratuïtes que ens permetran treballar amb el format XLIFF les podem dividir genèricament en dos grans grups:

- Editors d'XLIFF

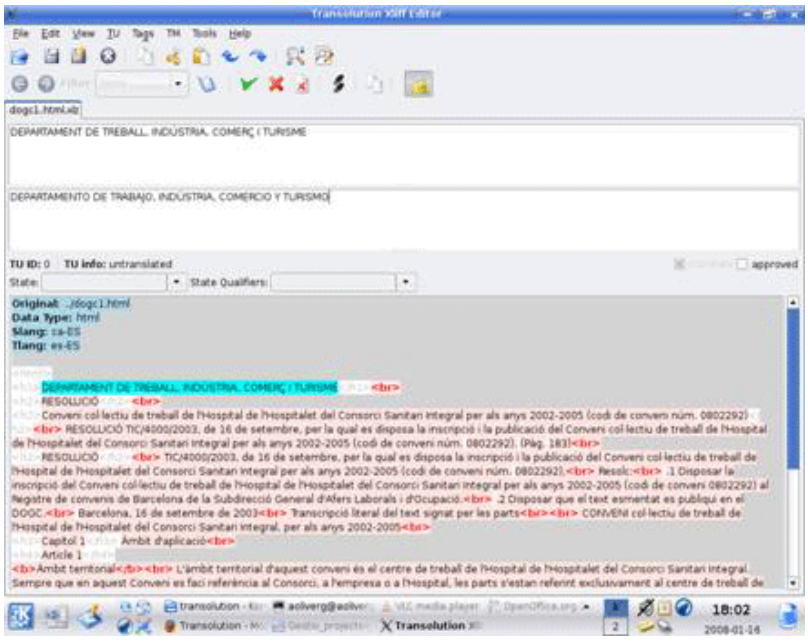
- Filtres per a la creació de fitxers XLIFF a partir d'altres formats i per a la recuperació del format original de la traducció a partir del fitxer XLIFF traduït

Algunes de les eines que presentarem proporcionen les dues funcions.

3.4.1. Transolution

Transolution és un editor de XLIFF desenvolupat en Python que es distribueix sota la llicència GNU i és, per tant, programari lliure. Transolution es pot descarregar de http://transolution.python-hosting.com.

Una característica molt notable d'aquest editor és que les memòries de traducció es consulten mitjançant un servidor de memòries, que es pot estar executant tant de forma local com de forma remota. L'aspecte de l'editor és el següent:

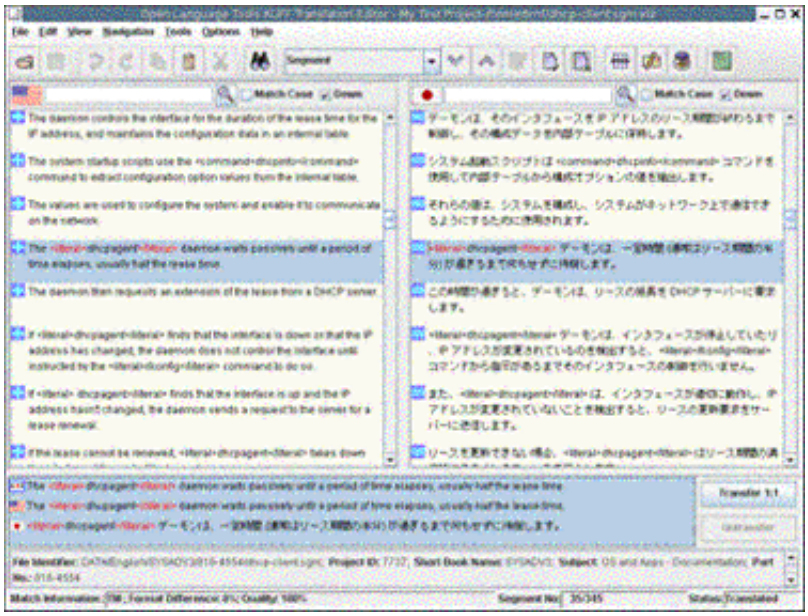


Transolution també proporciona una sèrie de filtres que permeten transformar html, documents d'OpenOffice (el content.xml) i DocBook a XLIFF i posteriorment generar el document traduït amb el mateix format.

3.4.2. Open Language Tools

Open Language Tools (https://open-language-tools.dev.java.net/) és un conjunt d'eines de traducció que pretenen facilitar la tasca de traducció de documentació i de localització de programari. Aquestes eines estan escrites en Java i es distribueixen sota la *Common Development and Distribution License,* que és una llicència de programari lliure, tot i que no és compatible amb la GNU GPL.

Les eines proporcionen un editor d'XLIFF que té el següent aspecte:



També proporciona una sèrie de filtres que permeten tractar els següents formats:

- Formats de documentació
  - HTML
  - Docbook
  - JSP
  - XML (de forma genèrica; necessita un fitxer de configuració per a cada tipus d'XML)
  - OpenOffice.org : sxw, sxc, sxi
  - Open Document Format : odw, odc, odi
  - Text pla

- Formats de localització de programari
  - PO (gettext)
  - Msg/tmsg (catgets)
  - Java .properties
  - Java ResourceBundle
  - Mozilla .DTD resource files

El funcionament del filtre és molt senzill, ja que només cal arrossegar els fitxers a transformar dins de la pantalla del filtre, que té el següent aspecte:

3.4.3. The Translate Toolkit

Aquest conjunt d'eines pot convertir entre diferents formats de traducció (com el Gettext PO, XLIFF, OpenOffice.org i d'altres). Això permet fer servir un únic format durant tot el procés de traducció o localització i fer servir un únic editor.

Algunes de les conversions que pot portar a terme són:

- **oo2po** – Conversor d'OpenOffice.org a PO
- **oo2xliff** – Conversor d'OpenOffice.org a XLIFF
- **csv2po** – Conversor de Comma Separated Value (CSV) a PO
- **php2po** – Conversor de PHP localisable string arrays a PO
- **txt2po** – Conversor de text pla a PO
- **html2po** – Conversor d'HTML a PO
- **xliff2po** – Conversor d'XLIFF (XML Localisation Interchange File Format) a PO
- **prop2po** – Conversor de Java property file (.properties) a PO
- **po2wordfast** – Conversor de memòries de traducció de Wordfast
- **po2tmx** – Conversor de memòries de traducció en TMX
- **csv2tbx** – Conversor de CSV a TBX

El toolkit proporciona també altres eines interessants. Aquest toolkit es pot descarregar de http://translate.sourceforge.net/wiki/toolkit/index.

4. Gestió de projectes i formats estàndard

El ús de formats estàndard pot facilitar enormement la tasca de la gestió de projectes de traducció. Un dels problemes importants a l'hora de gestionar els projectes on participen diversos traductors freelance és que sovint no tots ells disposen de la mateixa eina de traducció assistida, o fins i tot no en disposen de cap. Això té dues possibles conseqüències, si no fem ús de formats estàndard:

- Haurem de triar els traductors freelance en funció de l'eina que tinguin. Això no és sempre una bona idea, ja que potser el traductor ideal per un determinat projecte no disposa de la mateixa eina que nosaltres.

- Haurem de preparar fitxers especials per a cada traductor freelance, en funció de l'eina de traducció assistida que tingui. Això serà tant pel projecte, per les memòries, bases de dades terminològiques, etc. Això implica afegir molta feina a la ja per si complexa tasca de gestió de projectes.

L'ús dels formats estàndards que hem presentat en aquest capítol ens proporcionen, doncs, un seguit d'avantatges:

- Els fitxers dels projecte seran vàlids per treballar amb una gran quantitat d'eines

- Hi ha una bona oferta d'eines gratuïtes que ens permeten treballar amb aquests formats. D'aquesta manera els traductors freelance poden treballar amb eines de traducció assistida sense haver de fer cap inversió

- Els formats estàndard que hem presentat estan perfectament documentats. En cas de desastre sempre serà possible recuperar part o la totalitat de les dades. En el cas d'alguns formats propietaris, això no sempre és possible.

5. Conclusions

En aquest capítol hem presentat com els formats estàndard basats en XML que es fan servir a traducció (TMX, TBX, SRX i XLIFF) són de gran ajuda en la tasca de gestió de projectes de traducció. El fet de ser estàndard i d'existir moltes eines compatibles fan que siguin una molt bona opció com a formats de treball en la majoria de projectes de traducció i localització.

6. Per ampliar coneixements

Gómez, Josu. 2001. "Una guía al TMX", Tradumática, N° 0
http://www.fti.uab.es/tradumatica/revista/num0/articles/jgomez/art.htm