Proyectos de traducción y formatos estándar: TMX, TBX, XLIFF y SRX

Antoni Oliver

P08/B0240/00289

UOC
www.uoc.edu

Índice

Objetivos 5

1. Los formatos estándar basados en XML de uso en traducción 7
1.1. TMX (Translation Memory eXchange) 7
1.2. TBX (Term Base eXchange) 7
1.3. SRX (Segmentation Rule eXchange) 7
1.4. XLIFF (Localisation Interchange File Format) 8

2. Formatos estándar y herramientas de traducción asistida 9

3. Herramientas gratuitas para trabajar con formatos estándar 10
3.1. Herramientas para trabajar con TMX 10
3.1.1. TMX Validator 10
3.1.2. CSV Converter 11
3.1.3. Olifant, de ENLASO Tools 13
3.1.4. Tumatxa 14
3.2. Herramientas para trabajar con TBX 16
3.2.1. TBX Maker 16
3.3. Herramientas para trabajar con SRX 19
3.4. Herramientas para trabajar con XLIFF 19
3.4.1. Transolution 19
3.4.2. Open Language Tools 20
3.4.3. The Translate Toolkit 22

4. Gestión de proyectos y formatos estándar 23

5. Conclusiones 24

6. Para ampliar conocimientos 25

Objetivos

1. Conocer a fondo los formatos estándar basados en XML que se utilizan en el mundo de la traducción: TMX, TBX, XLIFF y SRX.

2. Analizar la aceptación de éstos formatos por parte de las principales herramientas de traducción asistida del mercado.

3. Valorar el uso de éstos formatos en la gestión de proyectos de traducción.

1. Los formatos estándar basados en XML de uso en traducción

El lenguaje XML está perdiendo importancia día a día en la mayoría de ámbitos. En cambio, hay una gran cantidad de formatos estándar que se basan en XML. En el mundo de la traducción existen unos formatos estándar basados en XML que sirven para compartir memorias de traducción, bases de datos terminológicas, reglas de segmentación y proyectos de traducción y localización. En este apartado presentaremos brevemente cada uno de éste formatos.

1.1. TMX (Translation Memory eXchange)

El TMX (*Translation Memory eXchange*) es un formato estándar basado en XML que sirve para compartir memorias de traducción. Mediante éste formato podemos hacer servir una memoria creada por una herramienta A en una herramienta B, si se da el caso que la herramienta A dispone de una utilidad de exportación en TMX y la herramienta B dispone de una de importación del mismo formato.

1.2. TBX (Term Base eXchange)

El TBX (*Term Base eXchange)* es un formato estándar basado en XML que sirve para compartir bases de datos terminológicas.

1.3. SRX (Segmentation Rule eXchange)

El SRX (*Segmentation Rule eXchange)* es un formato estándar basado en XML que sirve para compartir reglas de segmentación. Los programas de traducción asistida utilizan una serie de reglas de segmentación para dividir el texto a traducir en segmentos y tratar y presentar cada uno de estos segmentos de forma separada. El formato SRX nos servirá para compartir estas reglas de segmentación y asegurarnos que dos herramientas de traducción diferentes dividan un mismo texto de entrada en los mismos segmentos. Eso puede ser importante si estamos utilizando una memoria de traducción que se ha generado traduciendo con una herramienta A utilizando unas reglas de segmentado A. Si ahora queremos aprovechar esta memoria con otra herramienta B nos interesará utilizar las mismas reglas de segmentación, ya que de esta manera la probabilidad de encontrar coincidencias a la memoria aumenta.

1.4. XLIFF (Localisation Interchange File Format)

El XLIFF (*Localisation Interchange File Format*) es un formato estándar basado en XML para el intercambio de proyectos de traducción y localización. Mediante éste formato se puede traducir con una herramienta B un proyecto creado con una herramienta A.

2. Formatos estándar y herramientas de traducción asistida

No todos los formatos estándar presentados en el apartado anterior disfrutan del mismo grado de integración en las herramientas de traducción asistida del mercado. De los formatos mencionados, el que disfruta de un mayor nivel de integración es TMX, que es *soportado* por la inmensa mayoría de herramientas de traducción asistida. El segundo lugar posiblemente lo ocuparía TBX, pero a mucha distancia. SRX no disfruta prácticamente de ningún tipo de integración.

Un caso especial lo constituiría XLIFF. Hay una serie de herramientas de traducción asistida que son en realidad editores de XLIFF. Por otra parte, las herramientas que no *soportan* XLIFF pero que permiten crear filtros para XML, pueden potencialmente traducir ficheros XLIFF.

3. Herramientas gratuitas para trabajar con formatos estándar

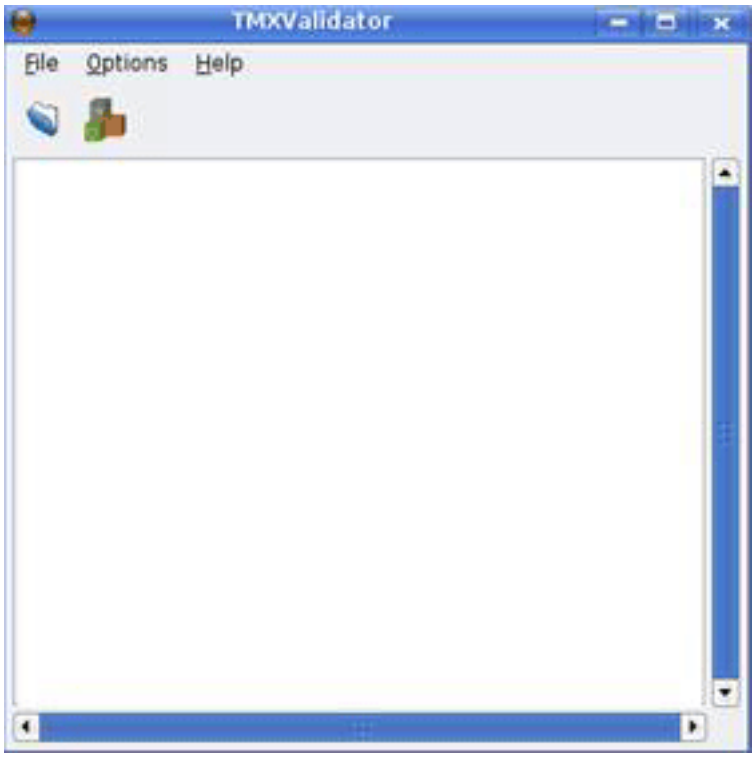
En este apartado presentaremos una serie de herramientas gratuitas que nos permitirán trabajar cómodamente con algunos de los formatos estándar.

3.1. Herramientas para trabajar con TMX

La mayoría de herramientas de traducción asistida, tanto comerciales como gratuitas, nos permiten trabajar con el formato TMX de intercambio de memorias de traducción. No tendremos mucha complicación para exportar e importar memorias de traducción en éste formato. Así pues, el intercambio de memorias de traducción entre diferentes herramientas de traducción asistida es una tarea fácil.

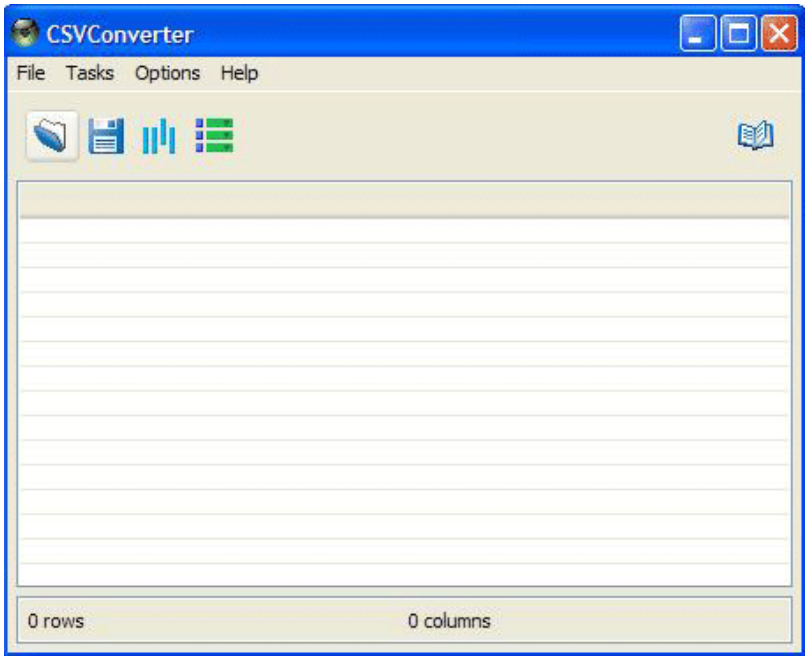
3.1.1. TMX Validator

TMX Validator es una aplicación gratuita de la empresa MaxPrograms que se puede descargar de http://www.maxprograms.com/freetools.html. Esta aplicación funciona bajo Windows y bajo Linux. Sirve para validar si un determinado archivo TMX es correcto o contiene algún tipo de error. Éste aplicación puede ser útil para comprobar que las memorias en TMX que recibimos o que enviamos sean realmente correctas. La interfaz visual es muy simple:



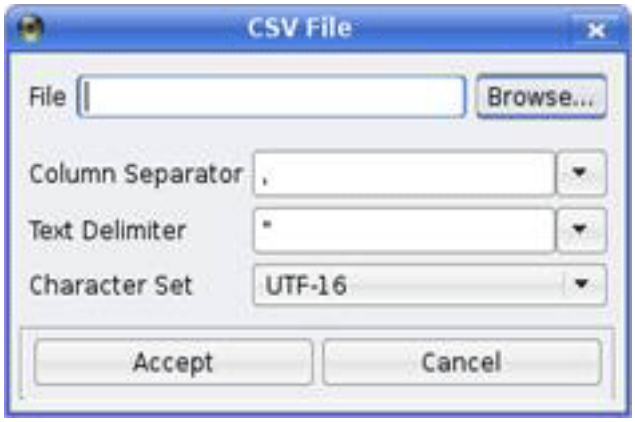
3.1.2. CSV Converter

CSV Converter es una aplicación gratuita de la empresa MaxPrograms que se puede descargar de http://www.maxprograms.com/freetools.html. Esta aplicación funciona bajo Windows y Linux. Esta utilidad permite convertir archivos CSV (ficheros de texto separados por coma o por otros separadores) en ficheros TMX. Si ejecutamos el programa nos aparece una pantalla de inicio como la siguiente:



La interfaz de usuario está en diversos idiomas, entre ellos el castellano. Si queremos cambiar el idioma de la interfaz sólo habrá que pinchar en el menú *Options* y escoger el idioma que deseemos. La explicación que sigue a continuación está realizada con la interfaz en inglés.

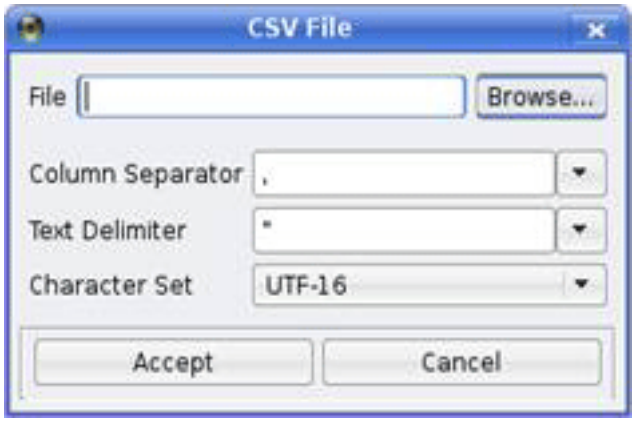
Si queremos importar un fichero CSV hay que hacer *File* > *Open CSV File* y nos aparecerá una pantalla como la siguiente:



En esta pantalla tenemos que seleccionar:

- El fichero CSV que queremos importar, con el botón *Browse*
- El separador de columnas, mediante la lista desplegable *Column separator*. Por ejemplo, si queremos importar un fichero de texto separado por tabuladores tendremos que seleccionar el separador de columnas *Tab*.
- El delimitador de texto, mediante la lista desplegable *Text delimiter*
- La codificación de caracteres mediante la lista desplegable *Character set*

Es imprescindible disponer de toda esta información del archivo que queremos importar. Si no conocemos esta información lo más aconsejable es abrir el archivo con un buen editor de textos y observarlo nosotros mismos. Si ejecutamos la importación correctamente nos aparecerá una pantalla de este estilo:



Una vez importado el archivo, si queremos eliminar alguna de las columnas, podemos utilizar este botón:

Si pinchamos sobre este botón nos aparecerá una pantalla como la siguiente:

Donde podremos seleccionar unas o más columnas por eliminar. Ahora, el último paso será indicar la lengua correspondiente a cada una de las columnas, mediante este botón.

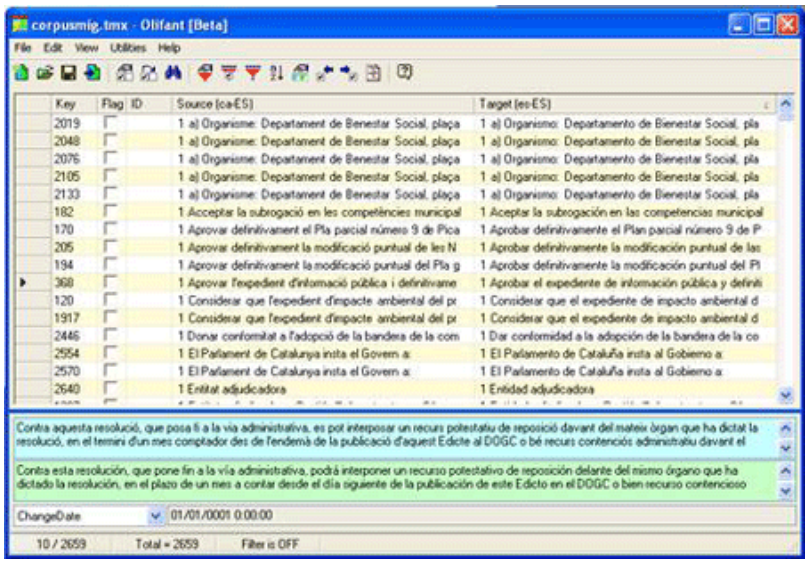
Pinchando sobre este botón nos aparecerá una pantalla como la siguiente, desde donde podremos seleccionar las lenguas.

Una vez hecho esto, podremos exportar la memoria en TMX haciendo *File > Export as TMX*.

3.1.3. Olifant, de ENLASO Tools

Olifant es una aplicación de la *suite* de herramientas gratuitas ENLASO Tools. Estas herramientas se pueden descargar de http://www.translate.com/technology/tools/. Funciona bajo Windows si tenéis instalado el entorno .NET. La herramienta permite hacer el mantenimiento de memorias de traducción, así como importar y exportar memorias desde y hacia diversos formatos.

La herramienta tiene el siguiente aspecto:



El funcionamiento de la herramienta es bastante intuitivo:

- Si queremos importar una memoria de traducción haremos *File* > *Open* y podremos escoger entre los formatos TMX, Wordfast TM File, Trados Text TM Files y Olifant TI Filas.

- Si queremos exportar la memoria que tenemos cargada simplemente tendremos que hacer *File* > *Save as* y escoger también entre uno de los formatos mencionados.

- Aparte de las funciones de abrir y guardar también disponemos de las de importación y exportación. La función de exportación permite exportar parte de la memoria, en función de unos filtros que se pueden definir.

- Podemos editar las entradas de la memoria para corregir errores o añadir información relevante.

- Podemos ejecutar búsquedas dentro de la memoria, tanto por los segmentos originales y traducidos como por el resto de información de la memoria.

3.1.4. Tumatxa

Tumatxa (www.tumatxa.com) es un gestor web de memorias de traducción desarrollado por la empresa vasca CodeSyntax y distribuida como software libre. Esta aplicación permite almacenar memorias de traducción en un repositorio web y hacer búsquedas en las memorias. Permite trabajar tanto en formato TMX como en formato PO. Se pueden realizar búsquedas sobre las memorias de traducción y seleccionar una o más memorias de traducción que descargar

a nuestro ordenador. Desde la página web de este producto se puede acceder a demostraciones. Presentamos a continuación una serie de capturas de pantalla con el objeto de explicar sus principales funcionalidades:

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 nos muestran las memorias disponibles. Desde esta pantalla podemos marcar memorias para descargar (utilizando las casillas de selección de la columna *MyTMX*) o bien hacer una búsqueda poniendo el texto que se desea buscar en el cuadro de texto y pinchando en el botón *Search here*. También podemos pinchar sobre el nombre de la memoria y nos mostrará información relevante y su contenido:

A continuación presentamos la pantalla que muestra la información de una memoria:

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 utilizamos la funcionalidad de búsqueda nos muestra todos los segmentos que contiene la cadena de busca (de todas las memorias si la búsqueda la hacemos desde la pantalla principal, o de una memoria en concreto si la hacemos desde la pantalla de presentación de una memoria):

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	

Esta aplicación puede ser de mucha utilidad en diversas situaciones:

- Para una empresa de traducciones que quiera dar acceso a los clientes a sus memorias de traducción.

- Para una empresa de traducción que quiera dar acceso a las memorias de traducción de un determinado proyecto a todos los participantes en el mismo.

- Para crear un repositorio público de memorias de traducción.

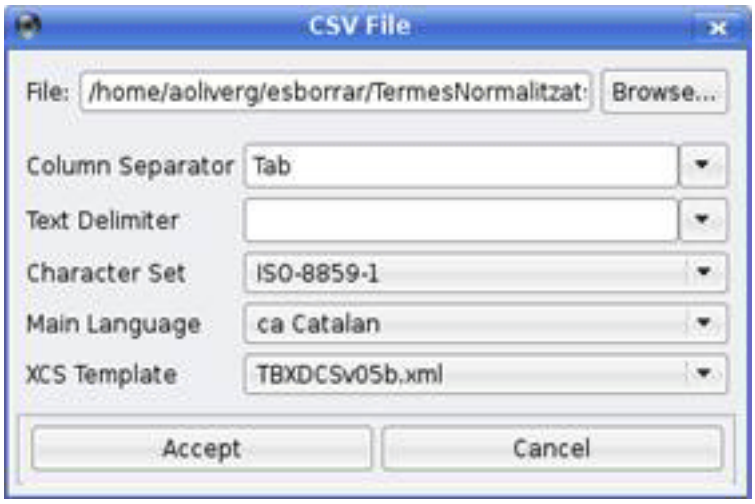
- Para localización de proyectos de software libre.

3.2. Herramientas para trabajar con TBX

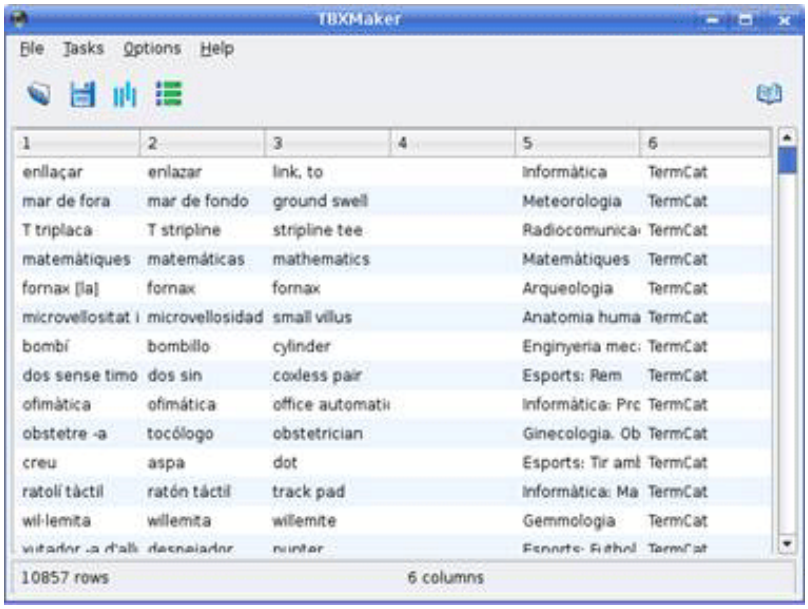
3.2.1. TBX Maker

TBX Maker es una aplicación gratuita de la empresa MaxPrograms que se puede descargar de http://www.maxprograms.com/freetools.html. Esta aplicación funciona bajo Windows y bajo Linux. El funcionamiento es muy similar al de la aplicación CSV Converter que hemos visto antes. El punto donde se diferencian básicamente las dos aplicaciones es el siguiente: en la pantalla donde cargamos el archivo que queremos transformar podemos escoger una plantilla determinada que nos marcará los atributos que podremos seleccionar para cada campo. A continuación presento una pequeña explicación del funcionamiento del programa.

Lo primero que hay que hacer es seleccionar el archivo que queremos importar haciendo *File* > *Open CSV File* y nos aparece una pantalla como la siguiente:



En esta pantalla seleccionaremos el archivo que queremos abrir, el separador de columnas, el delimitador de texto, la codificación de caracteres, la lengua principal y la plantilla XSC. Explicaremos con más detalle el tema de la plantilla más adelante; de momento podéis aceptar la que os aparezca por defecto. Si pinchamos en el botón *Accept* nos aparecerá una pantalla como la siguiente:



Ahora tendremos que eliminar, si hace falta, las columnas innecesarias, con el botón

Una vez eliminadas las columnas que no hagan falta, con el botón

asignaremos la categoría de cada columna. La cantidad de categorías que se pueden seleccionar depende de la plantilla escogida. Tendrá que aparecer una pantalla como la siguiente:

Es en esta pantalla donde tendremos que ir indicando qué información contiene cada columna. La información disponible dependerá de la plantilla que hayamos escogido en la pantalla inicial. Como la operación de seleccionar los atributos puede ser pesada, si prevemos que tenemos que tratar más ficheros iguales, podemos guardar la configuración con el botón *Save Configuration*. Si más adelante tenemos que volver a tratar un fichero igual podremos cargar esta configuración y nos ahorraremos el trabajo de ir escogiendo los atributos.

Una vez escogida la información de cada columna podemos exportar el archivo haciendo *File* > *Export as TBX* y nos aparecerá una pantalla como la siguiente que nos permitirá escoger el nombre y la ubicación del archivo TBX.

Para poder utilizar este programa con éxito es preciso conocer la estructura del fichero que queremos exportar a TBX y escoger la plantilla adecuada a la información que contiene el archivo. La documentación del programa no explica el contenido de cada plantilla y lo mejor es hacer una prueba preliminar con cada a una de las plantillas para ver cuál es la que se adapta mejor a nuestras necesidades.

3.3. Herramientas para trabajar con SRX

No hay herramientas específicas para trabajar con SRX. Algunas herramientas de traducción asistida ya soportan éste formato, entre ellas SDL, Heartsome y Trados.

3.4. Herramientas para trabajar con XLIFF

Las herramientas gratuitas que nos permitirán trabajar con el formato XLIFF las podemos dividir genéricamente en dos grandes grupos:

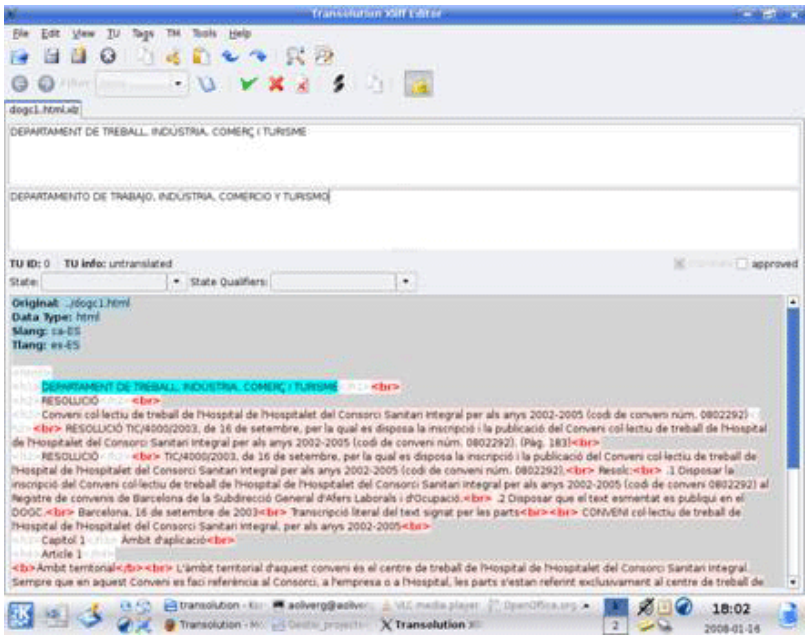
- Editores de XLIFF
- Filtros para la creación de ficheros XLIFF a partir de otros formatos y para la recuperación del formato original de la traducción a partir del fichero XLIFF traducido

Algunas de las herramientas que presentaremos a continuación proporcionan las dos funciones.

3.4.1. Transolution

Transolution es un editor de XLIFF desarrollado en Python que se distribuye bajo la licencia GNU y es, por lo tanto, software libre. Transolution se puede descargar de http://transolution.python-hosting.com

Una característica muy notable de este editor es que las memorias de traducción se consultan mediante un servidor de memorias, que se puede estar ejecutando tanto de forma local como de forma remota. El aspecto del editor es el siguiente:

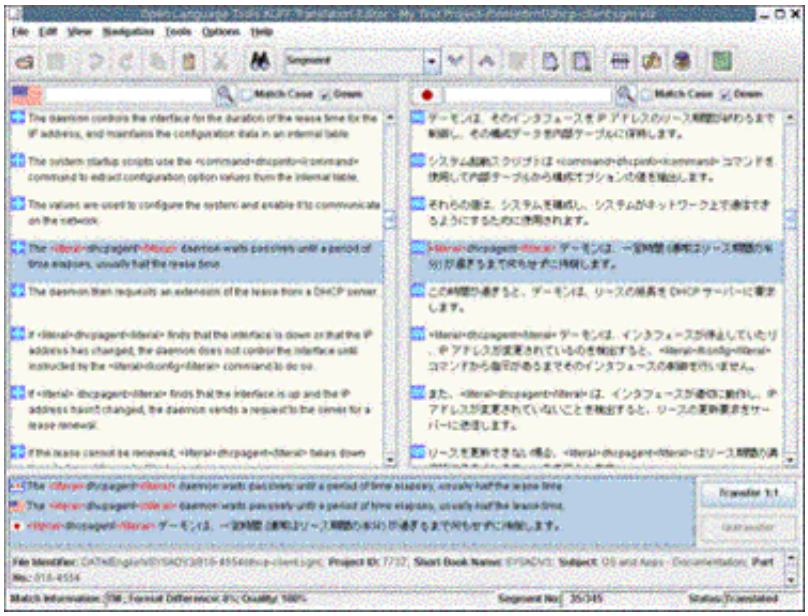


Transolution también proporciona una serie de filtros que permiten transformar html, documentos de OpenOffice (content.xml) y DocBook en XLIFF y posteriormente generar el documento traducido en el mismo formato.

3.4.2. Open Language Tools

Open Language Tools (https://open-language-tools.dev.java.net/) es un conjunto de herramientas de traducción que pretenden facilitar la tarea de traducción de documentación y de localización de software. Estas herramientas están escritas en Java y se distribuyen bajo la *Common Development and Distribution License*, que es una licencia de software libre, aunque no es compatible con la GNU GPL.

Las herramientas proporcionan un editor de XLIFF que tiene el siguiente aspecto:



También proporciona una serie de filtros que permiten tratar los siguientes formatos:

- Formatos de documentación
  - HTML
  - Docbook
  - JSP
  - XML (de forma genérica; necesita un fichero de configuración para cada tipo de XML)
  - OpenOffice.org: sxw, sxc, sxi
  - Open Documento Formado: odw, odc, odi
  - Texto simple

- Formatos de localización de software
  - PO (gettext)
  - Msg/tmsg (catgets)
  - Java .properties
  - Java ResourceBundle
  - Mozilla .DTD resource files

El funcionamiento del filtro es muy sencillo, ya que tan sólo hay que arrastrar los ficheros a transformar hasta de la pantalla del filtro, que tiene el siguiente aspecto:

3.4.3. The Translate Toolkit

Este conjunto de herramientas puede convertir entre diferentes formatos de traducción (como el Gettext PO, XLIFF, OpenOffice.org y otros). Esto permite utilizar un único formato durante todo el proceso de traducción o localización y utilizar un único editor.

Algunas de las conversiones que puede llevar a cabo son las siguientes:

- **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 texto simple a PO
- **html2po** - Conversor de HTML a PO
- **xliff2po** - Conversor de XLIFF (XML Localisation Interchange File Format) a PO
- **prop2po** - Conversor de Java property file (.properties) a PO
- **po2wordfast** - Conversor de memorias de traducción de Wordfast
- **po2tmx** - Conversor de memorias de traducción a TMX
- **csv2tbx** - Conversor de CSV a TBX

El toolkit proporciona también otras herramientas interesantes. Este toolkit se puede descargar de http://translate.sourceforge.net/wiki/toolkit/index

4. Gestión de proyectos y formatos estándar

El uso de formatos estándar puede facilitar enormemente la tarea de la gestión de proyectos de traducción. Uno de los problemas importantes a la hora de gestionar los proyectos en qué participan diversos traductores freelance es que a menudo no todos ellos disponen de la misma herramienta de traducción asistida, o incluso no disponen de ninguna. Esta situación tiene dos posibles consecuencias, si no usamos formatos estándar:

- Deberemos escoger los traductores freelance en función de la herramienta de que dispongan. Esto no es siempre una buena idea, ya que quizá el traductor ideal para un determinado proyecto no dispone de la misma herramienta de la que disponemos nosotros.

- Deberemos preparar ficheros especiales para cada traductor freelance, en función de la herramienta de traducción asistida de que disponga. Esto será así tanto para el proyecto, como para las memorias o bases de datos terminológicas. Por supuesto, este hecho hace más ardua a la ya de por sí compleja tarea de la gestión de proyectos.

El uso de los formatos estándares que hemos presentado en este capítulo nos proporcionan, pues, una serie de ventajas:

- Los ficheros de los proyecto serán válidos por trabajar con una gran cantidad de herramientas.

- Hay una buena oferta de herramientas gratuitas que nos permiten trabajar con estos formatos. De esta manera los traductores freelance pueden trabajar con herramientas de traducción asistida sin tener que hacer ninguna inversión.

- Los formatos estándar que hemos presentado están perfectamente documentados. En caso de desastre siempre será posible recuperar parte o la totalidad de los datos. En el caso de algunos formatos propietarios, esto no siempre es posible.

5. Conclusiones

En este capítulo hemos visto cómo los formatos estándar basados en XML que se utilizan en traducción (TMX, TBX, SRX y XLIFF) son de gran ayuda en la tarea de gestión de proyectos de traducción. El hecho de ser estándar y de existir muchas herramientas compatibles con ellos los convierten en una muy buena opción como formatos de trabajo en la mayoría de proyectos de traducción y localización.

6. Para ampliar conocimientos

Gómez, Josu. 2001. "Una guía al TMX", Tradumática, N° 0