revista.tradumàtica
tecnologies de la traducció

608-622

TMX: intercambio de memorias de traducción.

Antoni Oliver
Director del máster en Traducción Especializada
Universitat Oberta de Catalunya (UOC)

RESUMEN

En este artículo presentamos TMX (Translation Memory eXchange), el formato estándar de intercambio de memorias de traducción. Repasaremos el concepto de memoria de traducción y sus usos, que las convierten en uno de los principales recursos para el traductor. Veremos las estrategias para recuperar de manera rápida los segmentos más similares al que estamos traduciendo y los mecanismos para ordenar los segmentos recuperados según su similitud con el segmento a traducir. Se analizarán los formatos internos de las memorias de traducción en las principales herramientas de traducción asistida y se verá la importancia de disponer de un formato de intercambio que sea estándar, versátil y que permita su evolución para adaptarse a las nuevas necesidades. Presentaremos brevemente las especificaciones del formato TMX y sus diferentes niveles y analizaremos el grado de aceptación de este formato entre las herramientas de traducción asistida. Finalmente presentaremos algunas de las propuestas de futuro para este formato.

Palabras clave: TMX, memoria de traducción, herramientas TAO, estándar

RESUM *(TMX: intercanvi de memòries de traducció)*

En aquest article presentem TMX (Translation Memory eXchange), el format estàndard d'intercanvi de memòries de traducció. Repassarem el concepte de memòria de traducció i els seus usos, que les converteixen en un dels principals recursos per al traductor. Veurem les estratègies per recuperar de manera ràpida els segments més similars als que estem traduint i els mecanismes per ordenar els segments recuperats segons la seva similitud amb el segment a traduir. Presentarem breument les especificacions del format TMX i els seus diferents nivells i analitzarem el grau d'acceptació d'aquest format entre les eines de traducció assistida.

Paraules clau: TMX, memories de traducció, eines TAO, estàndard.

ABSTRACT *(TMX: Translation memories interchange)*

In this paper the standard format for translation memories interchange (TMX) is presented. We review the concept of translation memory and its uses. We also present strategies for quick access to the most similar segments to the one being translated and the ways to sort the retrieved segments according to similarity. The specifications of the TMX format and its levels will be presented. We analyze the degree of implementation of this format in CAT tools.

Keywords: TMX, translation memory, CAT tools, standard

1. Las memorias de traducción

1.1. ¿Qué es una memoria de traducción?

Una memoria de traducción es un repositorio de segmentos de texto en una determinada lengua con las correspondientes traducciones a una o más lenguas.

Las memorias de traducción proporcionan una relación directa entre los segmentos de texto en una lengua y sus traducciones a otra lengua. Los segmentos de texto acostumbran a ser oraciones, pero como el proceso de segmentación se realiza a partir de un conjunto de reglas basadas en expresiones regulares que tienen en cuenta las secuencias de ciertos caracteres, los segmentos no siempre coinciden con oraciones desde el punto de vista gramatical. Las memorias de traducción no relacionan unidades más grandes como, por ejemplo, párrafos porque la probabilidad de encontrar párrafos iguales o similares en un texto es muy baja. Tampoco se relacionan unidades más pequeñas como, por ejemplo, palabras, ya que el traductor humano no trabaja tratando de manera aislada estas unidades.

1.2. ¿Para qué sirve una memoria de traducción?

Las memorias de traducción habitualmente se utilizan en las herramientas de traducción asistida. De esta manera, cuando la herramienta presenta un nuevo segmento a traducir busca en la memoria o memorias asignadas al proyecto los segmentos iguales o similares y presenta las correspondientes traducciones. El traductor humano podrá aceptar alguna de estas propuestas, modificándola si es necesario. Una vez traducido el nuevo segmento, el segmento original y el traducido se pueden incorporar a una memoria de traducción. De este modo, las memorias de traducción se van ampliando a medida que se va traduciendo y con el tiempo este recurso va siendo más valioso.

Las memorias de traducción pueden tener también otras utilidades, entre las que se pueden destacar:

- La búsqueda de ciertas expresiones para ver cómo se han traducido anteriormente.
- Relacionada con la anterior, la búsqueda automática de equivalentes de traducción de unidades terminológicas.

Dado que una memoria de traducción es también un corpus paralelo, especialmente cuando estas son de gran tamaño, se pueden utilizar en aplicaciones típicas de los corpus paralelos, como pueden ser:

- Tareas de extracción automática de terminología.
- Entrenamiento de sistemas de traducción automática estadísticos.

Existen diversas páginas web que permiten hacer búsquedas en memorias de traducción. Linguee (www.linguee.com) permite buscar expresiones y muestra información que proviene de bases de datos terminológicas y diccionarios junto a los segmentos de la memoria de traducción que contienen el término de búsqueda. Otro servicio muy similar es Reverso Context (http://context.reverso.net/). En la literatura científica, la tarea de buscar la traducción

de un determinado término o expresión en una memoria de traducción, especialmente si esta búsqueda la realiza la herramienta automáticamente, se conoce como translation spotting (Simard, 2003).

1.3. Recuperación de segmentos: indexación de memorias de traducción

Cuando una herramienta de traducción asistida busca si existe un segmento igual o similar al que estamos traduciendo, esta búsqueda no se puede llevar a cabo de una manera secuencial, recuperando todos los segmentos y comparándolos con el segmento a traducir. La búsqueda de segmentos exactamente iguales se puede realizar de una manera muy rápida utilizando la estructura de datos adecuada, pero la búsqueda de segmentos similares requeriría mucho más tiempo. Incluso para memorias pequeñas y medianas, la búsqueda de segmentos similares requeriría tanto tiempo que la herramienta en lugar de agilizar el proceso de traducción, lo entorpecería.

Para acceder de manera rápida a la información contenida en una memoria de traducción se tiene que utilizar algún tipo de base de datos y se tendrá que llevar a cabo algún tipo de indexación.

La indexación de una memoria de traducción consiste en construir un índice inverso de las palabras (o de fragmentos de palabras, o de alguna de las palabras) en la lengua de partida que aparecen en la memoria. El índice nos da el identificador de todos los segmentos en los que aparece una determinada palabra.

Veamos el siguiente ejemplo (extraído de Oliver 2015)
Imaginemos que tenemos una memoria de traducción con los siguientes segmentos:

ID	Segmento original	Segmento traducido
1	Search the Legal framework	Buscar en Marco jurídico
2	Legal framework of the ESCB	Régimen jurídico del SEBC
3	ECB institutional provisiones	Disposiciones institucionales del BCE
4	Monetary policy and Operations	Política monetaria y operaciones
5	Payment and settlement systems	Sistemas de pago y Liquidación
6	Banknotes and coins, means of payment and currency matters	Billetes de banco y monedas, medios de pago y Cuestiones de moneda
7	Foreign exchange and Foreign reservas	Divisas y reservas exteriores
8	The approach was successful: in volume terms, close to 80% of the initial Banknote demand and 97% of the total coin needs for the changeover had been distributed before 1 January 2002.	Resultando un planteamiento acertado ya que, en términos de volumen, casi el 80% de la demanda inicial de billetes y el 97% de las monedas necesarias para la introducción del euro habían sido distribuidos antes del 1 de enero de 2002.

9	Employment, conduct, fraud prevention and transparency	Contratación, conducta, prevención del fraude y transparencia
10	Financial market stability	Estabilidad de los Mercados Financieros

El índice inverso tendría el siguiente aspecto:

and 4:5:6:7:8:9 approach 8 banknote 8 banknotes 6 been 8 before 8 changeover 8 close 8 coin 8 coins 6 conduct 9 currency 6 demand 8 distributed 8 ecb 3 employment 9 escb 2 exchange 7	financial 10 for 8 foreign 7 framework 1:2 fraud 9 had 8 in 8 initial 8 institutional 3 january 8 legal 1:2 market 10 matters 6 means 6 monetary 4 needs 8 of 2:6:8:8 operations 4	payment 5:6 policy 4 prevention 9 provisions 3 reserves 7 search 1 settlement 5 stability 10 successful 8 systems 5 terms 8 the 1:2:8 to 8 total 8 transparency 9 volume 8 was 8

Esta tabla nos proporciona información sobre en qué segmentos aparecen cada una de las palabras. Por ejemplo, *payment* aparece en los segmentos 5 y 6 y *market* en el 10.

Imaginemos que queremos encontrar segmentos parecidos a este:

Banknotes, coins, and types of payment

Tomaríamos los índices de los segmentos y el índice que apareciera más veces sería probablemente el más parecido, ya que contendría más palabras comunes. Dependiendo del algoritmo de cálculo de similitud entre segmentos, el orden de las palabras nos puede jugar malas pasadas así que, a menudo, se toma no sólo el más parecido, sino también los primeros más parecidos y se calcularía la similitud, hasta que ésta estuviera por debajo de la similitud mínima dada por el usuario. Sobre el cálculo de similitud, hablaremos en el siguiente subapartado. Según esto, quedaría:

banknotes 6
coins 6
types
payment 5:6

y, por tanto, el segmento más parecido sería el 6 *(Banknotes and coins, means of payment and currency matters)*, ya que tiene 3 palabras que coinciden. El segundo segmento más parecido sería el 5 *(Payment and settlement systems).*

La explicación que hemos presentado corresponde a una estrategia muy básica que puede mejorarse en muchos aspectos. El algoritmo de indexación puede llegar a ser un secreto comercial de las herramientas de traducción asistida. A continuación, presentamos algunos aspectos básicos de mejora de este algoritmo:

- Cada lengua tiene una serie de palabras funcionales que tienden a aparecer con mucha frecuencia en los textos (en nuestro ejemplo *and* y *the*). Muy a menudo estas palabras funcionales son palabras cortas y, por este motivo, es habitual no tener en cuenta las palabras formadas por pocos caracteres en el momento de calcular los índices[1].
- En muchas lenguas las palabras correspondientes a categorías abiertas están sometidas a flexión. Por este motivo es habitual en el proceso de indexación eliminar los últimos caracteres de las palabras más largas[2].
- En general las herramientas de traducción asistida trabajan con poca información lingüística específica para una determinada lengua. Si se utiliza información lingüística es posible mejorar el proceso de indexación utilizando técnicas de *stemming* (eliminación de los sufijos morfológicos, como el algoritmo de Porter, 1980). Si añadimos aún más recursos lingüísticos específicos para una lengua podemos utilizar analizadores morfológicos y lematizadores como, por ejemplo, *Freeling*[3] (Padró y Stanilovsky 2012).

1.4. Selección y reordenamiento de los segmentos: cálculo de las similitudes

En el apartado anterior hemos visto cómo podemos recuperar de una manera rápida los segmentos de la memoria de traducción que son más parecidos al segmento que estamos traduciendo. Es posible que esta búsqueda nos devuelva una serie de segmentos y es importante que la herramienta presente este resultado ordenado según su similitud al que estamos traduciendo. Evidentemente, si hay un segmento exactamente igual, este se presentará en primer lugar, pero ¿cómo podemos valorar la similitud de los segmentos parecidos?

Una primera aproximación podría medir la similitud entre dos segmentos contando cuántas palabras tienen en común. Según esta medida, si dos segmentos tienen todas las palabras iguales la similitud sería del 100%. Esta aproximación, no obstante, puede fallar por diferencias en el orden la de las palabras. Vamos a aplicar esta sencilla medida a los segmentos del ejemplo presentado en el apartado anterior. Si tenemos en cuenta el número de palabras que coinciden:

[1] Esta estrategia puede no ser válida para algunas lenguas y, muy especialmente, para lenguas como el chino, japonés y coreano.
[2] De nuevo esta estrategia no es válida para todas las lenguas.
[3] http://nlp.lsi.upc.edu/freeling/

Segmento que buscamos: Banknotes, coins, and types of payment

1º segmento encontrado: Banknotes and coins, means of payment and currency matters

Palabras coincidentes: 3 Total palabras (segmento a buscar): 6 Similitud: 50%

2º segmento encontrado: Payment and settlement systems

Palabras coincidentes: 1 Total palabras (segmento a buscar): 6 Similitud: 16.6%

Ahora aplicamos una pequeña variación y tendremos en cuenta el número de caracteres de las palabras que son iguales respecto el número total de caracteres:

Segmento que buscamos: Banknotes, coins, and types of payment

1º segmento encontrado: Banknotes and coins, means of payment and currency matters

Caracteres coincidentes: 22 Total caracteres (segmento a buscar): 34 Similitud: 64.7%

2º segmento encontrado: Payment and settlement systems

Caracteres coincidentes: 7 Total caracteres (segmento a buscar): 34 Similitud: 20.6%

Esta medida de similitud tan simple puede no representar el esfuerzo real de edición que supondría modificar la traducción del segmento recuperado hasta convertirlo en la traducción deseada del segmento original. Por este motivo será necesario utilizar medidas más complejas.

Una posibilidad es utilizar la distancia de edición o “distancia de Levenshtein” (1966), que proporciona el número mínimo de operaciones de edición (inserción, supresión o substitución de un carácter) para poder transformar una cadena de caracteres en otra.

La distancia de edición nos puede dar una idea muy aproximada del esfuerzo real que puede suponer editar una coincidencia parcial de una memoria de traducción. Si aplicamos esta medida a nuestro ejemplo obtenemos los siguientes valores:

Segmento que buscamos: Banknotes, coins, and types of payment

1º segmento encontrado: Banknotes and coins, means of payment and currency matters

Distancia de edición: 31

2º segmento encontrado: Payment and settlement systems

Distancia de edición: 29

Este resultado es algo sorprendente, ya que el segundo segmento recuperado, aunque tiene una distancia de edición menor probablemente ofrezca menos información útil para el traductor. En Sommers (2003) podemos encontrar otros ejemplos de cómo la distancia de edición no siempre funciona correctamente.

1.5. Tipos de coincidencia: exacta y parciales

Las coincidencias que se recuperan de la memoria de traducción pueden ser exactas (exact matches) cuando el texto del segmento que recuperamos de la memoria de traducción es exactamente igual al texto del segmento que buscamos; o parciales (fuzzy matches): cuando el texto del segmento que recuperamos de la memoria de traducción no es exactamente igual al texto del segmento que buscamos, pero su índice de similitud es superior o igual al índice fijado por el usuario. Estas definiciones no explican qué pasa cuando las coincidencias sólo difieren en algunas cifras u otras cadenas alfanuméricas que muchas herramientas son capaces de sustituir automáticamente, o bien en aspectos relacionados con el formato. En este sentido, Bowker (2002:98) introduce la distinción entre coincidencia exacta (exact match) y coincidencia completa (full match). Según la autora:

> *Una coincidencia exacta es 100% idéntica al segmento que está traduciendo el traductor tanto desde el punto de vista lingüístico, como desde el punto de vista del formato. Esto significa que las dos cadenas tienen que ser idénticas en todos los sentidos, incluyendo la ortografía, la puntuación, la flexión, cifras e incluso el formato (cursiva, negrita, etc.)*
>
> *Una coincidencia total se produce cuando el segmento que se está traduciendo difiere del segmento almacenado en la memoria de traducción sólo en lo que se denominan elemento variables, que a menudo en inglés reciben el nombre de* placeables*.*

Cuando trabajamos con una herramienta de traducción asistida es importante establecer la similitud mínima para recuperar segmentos de la memoria. Si seleccionamos una similitud muy alta, por ejemplo del 95%, será muy difícil encontrar coincidencias en la memoria y el programa mostrará muy pocas sugerencias. En cambio, si establecemos una similitud muy baja de, por ejemplo, el 10%, el sistema nos mostrará muchas sugerencias, pero probablemente serán de poca utilidad. Un buen compromiso puede ser establecer la similitud mínima entre el 65 y el 85%.

1.6. Combinación de unidades subsegmentales

En muchos casos en la memoria de traducción no tenemos ningún segmento completo similar al que estamos traduciendo pero, en cambio, disponemos de uno o más fragmentos de segmentos que contienen información interesante. Algunas herramientas de traducción asistida son capaces de buscar en la memoria coincidencias a nivel subsegmental, es decir, a nivel de fragmentos de un segmento y deducir qué parte del segmento traducido corresponde a la traducción de un determinado subsegmento. A partir de estos

subsegmentos traducidos algunas herramientas son también capaces de construir la traducción total del segmento.

2. Necesidad de un formato de intercambio para las memorias de traducción

En las secciones anteriores hemos podido observar que para poder utilizar una memoria de traducción de manera eficiente se ha tenido que llevar a cabo un proceso de indexación. La memoria de traducción se almacenará de manera interna en algún tipo de base de datos o estructura informática similar. Cada herramienta realiza el proceso de indexación de manera diferente y utiliza estructuras y bases de datos totalmente diferentes. Este hecho hace que, sin la ayuda de algún formato de intercambio eficiente, sea imposible compartir memorias de traducción entre usuarios de diferentes herramientas de traducción asistida.

Supongamos, por ejemplo, que dispongo de una memoria de traducción entre el catalán y castellano, de 10.000 segmentos, compilada durante diversos meses de trabajo y que proviene de diferentes proyectos de traducción. Si únicamente me interesase compartir el segmento original y el segmento traducido, podría generar un fichero de texto separado por algún carácter específico (comas, tabuladores, etc.) y enviárselo a mi colaborador. A continuación, vemos un segmento de ejemplo en este formato (entre los diferentes campos hay un tabulador que no se aprecia en el texto):

> EDICTE de 14 de febrer de 2000, sobre un acord de la Comissió d'Urbanisme de Tarragona referent al municipi de Reus. EDICTO de 14 de febrero de 2000, sobre un acuerdo de la Comisión de Urbanismo de Tarragona referente al municipio de Reus.

Ahora bien, si me interesase también enviar la información sobre cuándo se ha generado cada segmento, el usuario que lo ha creado, etc., aunque sería también posible utilizar algún tipo de fichero de texto CSV o tabulado, este ya sería más complicado. Podría tener la estructura del siguiente ejemplo:

> EDICTE de 14 de febrer de 2000, sobre un acord de la Comissió d'Urbanisme de Tarragona referent al municipi de Reus. EDICTO de 14 de febrero de 2000, sobre un acuerdo de la Comisión de Urbanismo de Tarragona referente al municipio de Reus. 20140508T150609Z aoliverg

Imaginémonos que además la memoria es multilingüe, la estructura del fichero tabulado se complicaría cada vez más. Además, para poder importar la memoria de traducción correctamente necesitaríamos información adicional sobre la estructura del fichero. Esta información se podría especificar en la primera línea del archivo y la importación se podría realizar de manera más o menos automática. Por ejemplo, el fichero podría quedar del siguiente modo:

> CA ES changedate changeid
>
> EDICTE de 14 de febrer de 2000, sobre un acord de la Comissió d'Urbanisme de Tarragona referent al municipi de Reus. EDICTO de 14 de febrero de 2000, sobre un acuerdo de la

Comisión de Urbanismo de Tarragona referente al municipio de Reus. 20140508T150609Z aoliverg

Para evitar los problemas que pueden surgir a la hora de compartir memorias de traducción, un grupo de expertos del grupo de interés OSCAR (Open Standards for Container/Content Allowing Re-use) de LISA (Localization Industry Standards Association) desarrolló el formato TMX en 1997.

3. El estándar TMX

3.1. Procesadores de texto

Como ya hemos comentado, el objetivo del formato TMX es proporcionar un método estándar para la descripción de los datos contenidos en una memoria de traducción con el fin de poder compartir las memorias entre diferentes usuarios y herramientas sin que se produzca una pérdida de datos importante durante el proceso.

El TMX se define en dos partes:

- Una especificación del formato del contenedor, es decir, los elementos de nivel superior que proporcionan información sobre el archivo en conjunto y sobre las entradas. En TMX una entrada consistente en segmentos alineados de texto en dos o más lenguas se denomina *unidad de traducción* (el elemento <tu>).
- Una especificación para el formato de meta-marcado de bajo nivel para el contenido de un segmento de texto de la memoria de traducción. En TMX, un segmento individual del texto de la memoria de traducción en una lengua determinada se denota con el elemento <seg>.

A continuación podemos ver un ejemplo de memoria de traducción en formato TMX que consiste en un único segmento en castellano y catalán (el mismo ejemplo que hemos ofrecido en la explicación del formato tabulado o CSV):

```
<?xml version="1.0" encoding="UTF-8"?>
    <!DOCTYPE tmx SYSTEM "tmx11.dtd">
    <tmx version="1.1">
     <header creationtool="OmegaT" o-tmf="OmegaT TMX" adminlang="EN-US"
    datatype="plaintext" creationtoolversion="2.6.3" segtype="sentence"
    srclang="CA"/>
     <body>
      <tu>
       <tuv lang="CA">
        <seg>EDICTE de 14 de febrer de 2000, sobre un acord de la Comissió
    d'Urbanisme de Tarragona referent al municipi de Reus.</seg>
       </tuv>
```

```
<tuv lang="ES" changeid="aoliverg" changedate="20140508T150609Z">
<seg>EDICTO de 14 de febrero de 2000, sobre un acuerdo de la Comisión de Urbanismo de Tarragona referente al municipio de Reus.</seg>
</tuv>
</tu>
</body>
</tmx>
```

El TMX es un formato basado en XML y, por lo tanto, cumple todas sus especificaciones. Además, al ser XML puede beneficiarse de todas las herramientas estándar para el procesado de este formato. Entre estas herramientas cabe destacar la de verificación de la validez de un fichero, que se puede realizar de manera automática comparándola con el DTD (Document Type Definition) del propio TMX. Dado que el formato XML distingue entre mayúsculas y minúsculas, todos los nombres de elementos y atributos del formato TMX se definen en minúsculas.

Los archivos TMX utilizan siempre la codificación Unicode, ya sea UTF-16, UTF-8 o ISO-646 (es decir US-ASCII, que es de hecho un subconjunto del UTF-8). Dado que sólo se permiten estas codificaciones, no es necesario incluir la declaración de codificación en la cabecera del XML.

A diferencia del HTML y, como ocurre con cualquier otro XML, en TMX sólo se permiten las siguientes referencias a entidades: & (&), < (<),> (>), ' (‘) y " (“).

Los archivos TMX tienen la siguiente estructura general:

- El elemento raíz, que engloba todo el documento TMX, es <tmx> y contiene dos elementos:
    - <header>: que contiene metadatos sobre el documento
    - <body>: que contiene la colección de unidades de traducción, los elementos <tu>. A su vez, este elemento contiene variantes de unidades de traducción para una lengua determinada, que se define con el atributo xml:lang, en los elementos <tuv>, y que puede contener:
        - <seg>: que contiene el texto del segmento dado.
        - <note>: que se utiliza para añadir comentarios.
        - <prop>: permite definir propiedades del elemento padre (el elemento que contiene a <prop>). Estas propiedades no están definidas por el estándar y pueden utilizarse para cualquier propósito.

No es el objeto de este artículo repasar todas las especificaciones de TMX por lo que remitimos al lector interesado a las propias especificaciones del estándar en Savourel (2005). Si nos fijamos en el ejemplo anterior, el formato TMX es bastante claro y se puede deducir de manera fácil la información que contiene.

3.2. Niveles

El TMX puede tener dos niveles de implementación:

- Nivel 1. Únicamente texto plano: soporte sólo para el contenedor. Los datos entre los elementos <seg> contienen únicamente información textual, sin marcas de formato.
- Nivel 2. Marcado del contenido: soporte tanto para el contenedor como para el contenido. Se utiliza el marcado de contenido propio del TMX para permitir que otras herramientas que sean compatibles con TMX nivel 2 puedan recrear la versión traducida de un documento original usando únicamente el archivo TMX.

El ejemplo anterior de TMX correspondería a un Nivel 1. Si el segmento original tuviera el siguiente formato:

> **EDICTO** de 14 de febrero de 2000, sobre un acuerdo de la Comisión de Urbanismo de Tarragona referente al municipio de Reus.

la memoria en TMX de Nivel 2 tendría el siguiente aspecto:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
 <header creationtool="OmegaT" o-tmf="OmegaT TMX" adminlang="EN-US"
datatype="plaintext" creationtoolversion="2.6.3" segtype="sentence" srclang="CA"/>
 <body>
<!-- Default translations -->
   <tu>
    <tuv xml:lang="CA">
      <seg><bpt i="0" x="0">&lt;f0&gt;</bpt>EDICTO<ept i="0">&lt;/f0&gt;</ept><bpt
i="1" x="1">&lt;f1&gt;</bpt> de 14 de febrero de 2000, sobre un acuerdo de la <ept
i="1">&lt;/f1&gt;</ept><bpt i="2" x="2">&lt;f2&gt;</bpt>Comissió d'Urbanisme<ept
i="2">&lt;/f2&gt;</ept><bpt i="3" x="3">&lt;f3&gt;</bpt> de Tarragona referent al
municipi de Reus.<ept i="3">&lt;/f3&gt;</ept></seg>
    </tuv>
    <tuv xml:lang="ES" changeid="aoliverg" changedate="20140508T150609Z">
```

<seg><bpt i="0" x="0"><f0></bpt>EDICTO<ept i="0"></f0></ept><bpt i="1" x="1"><f1></bpt> de 14 de febrero de 2000, sobre un acuerdo de la <ept i="1"></f1></ept><bpt i="2" x="2"><f2></bpt>Comisión de Urbanismo<ept i="2"></f2></ept><bpt i="3" x="3"><f3></bpt> de Tarragona referente al municipio de Reus.<ept i="3"></f3></ept></seg>

</tuv>

</tu>

<!-- Alternative translations -->

</body>

</tmx>

Si nos fijamos, este nivel incluye información sobre el formato del segmento. El nivel 2 de TMX es muy útil para traducir documentación con formato variado (negritas, colores, etc.), ya que en muchos casos podrá recuperar también las marcas de formato y ahorrará tiempo de edición al traductor.

En Gómez (2001) encontramos una explicación muy detallada de cómo se puede recuperar el formato en el TMX de nivel 2. Como explica el autor, los diferentes formatos utilizan diferentes etiquetas para indicar la información de formato. Estas etiquetas suelen estar formadas por un "inicio de etiqueta" y un "fin de etiqueta" y aplica el formato indicado a todo el texto que está entre el inicio y el final. Para mantener esta información, el TMX dispone de dos elementos:

- <bpt>: *begin paired tag* o inicio de etiqueta emparejada
- <ept>: *end paired tag* o fin de etiqueta emparejada

Si tomamos como ejemplo una línea de HTML como la siguiente:

Esto es un ejemplo de texto en negrita.

En TMX se expresaría como[4]

<seg>Esto es un ejemplo de texto en <bpt></bpt>negrita<ept></ept>.</seg>

En cambio, si la misma frase estuviese en formato RTF sería:

Esto es un ejemplo de texto en {\b negrita}.

Que en TMX quedaría:

[4] Fijémonos que las marcas i se han modificado por y ya que si no, el fragmento resultante sería un XML no válido.

Esto es un ejemplo de texto en <bpt>{\b</bpt> negrita<ept>}</ept>.

En el artículo citado se pueden consultar otros ejemplos interesantes.

3.3. Grado de aceptación entre las herramientas TAO

El formato TMX es probablemente el más extendido de los formatos XML utilizados en el mundo de la traducción, por delante del TBX (Term Base eXchange) y el XLIFF *(XML Localization Interchange File Format)*. También está mucho más extendido que el formato SRX *(Segmentation Rule eXchange),* que se utiliza para el intercambio de reglas de segmentación, aspecto importante, aunque no imprescindible, si se pretenden compartir memorias de traducción generadas mediante distintas herramientas de traducción. Se puede afirmar que la totalidad de las herramientas de traducción asistida por ordenador admiten el formato TMX. Ahora bien, no todas ellas admiten los dos niveles de implementación, aunque se prevé que lo hagan en un futuro próximo.

Aunque, como ya hemos dicho, la mayoría de herramientas son compatibles con el TMX, todavía existen algunos problemas de compatibilidad (Raya, 2007):

- Algunas herramientas no utilizan analizadores de XML estándar sino algoritmos propios, por lo que es posible que no acepten algunos TMX perfectamente válidos. Un ejemplo sería un elemento *header* sin elementos hijos escrito como <header/> en lugar de <header></header>.
- Algunas herramientas generan archivos TMX que no son documentos XML válidos. El caso más frecuente es la inclusión de caracteres de control no permitidos en el estándar XML, en lugar de entidades.
- Algunas herramientas sólo son compatibles con versiones antiguas de XML. La versión actual es la 1.4b, pero muchas herramientas todavía funcionan con la versión 1.1.
- Otras herramientas no admiten documentos TMX multilingües y restringen el número de lenguas permitidas a dos por archivo.

Existen diversas herramientas que nos permiten verificar y corregir archivos TMX. A modo de ejemplo, enumeraremos algunas de ellas:

- *CheckMate* de *Okapi Tools*[5].
- *TMXValidator* de *MaxPrograms*[6].
- *TMX Editor* de *Heartsome*[7].

[5] http://www.opentag.com/okapi/wiki/index.php?title=CheckMate
[6] http://www.maxprograms.com/products/tmxvalidator.html
[7] https://github.com/heartsome/tmxeditor8

3.4. Propuestas de futuro

En el año 2007 se publicó un borrador de las especificaciones de la nueva versión 2.0 de TMX, pero no se ha avanzado más en esta versión. En marzo de 2011 LISA fue declarada insolvente y todas sus especificaciones han pasado a publicarse bajo una licencia *Creative Commons Attribution 3.0* que permite su reutilización y la creación de trabajos derivados.

Aunque por el momento no parece que haya una actividad importante en la creación de nuevas versiones del estándar TMX sí que podemos encontrar diversas propuestas para la extensión de sus funcionalidades, entre las que podemos destacar la de Forcada (2014) que propone la manera de representar unidades subsegmentales mediante TMX utilizando las etiquetas "bpt" y "ept".

4. Conclusiones

En este artículo se ha pretendido dar una visión general del formato TMX para el intercambio de memorias de traducción. Primero se ha presentado el concepto de memoria de traducción y el proceso de indexación para entender la importancia de disponer de un formato de intercambio estándar.

El formato TMX está bien introducido en todas las herramientas de traducción asistida aunque, como hemos visto, no siempre son compatibles con la última versión o con sus dos niveles. Todavía existen algunos problemas aislados de compatibilidad con alguna herramienta de traducción asistida.

Por último hemos visto que no se han producido nuevas versiones del estándar en mucho tiempo pero que existen propuestas de ampliación de sus funcionalidades.

Bibliography

Bowker, L. (2002). Computer-Aided Translation Technology. A Practical Introduction. Ottawa: University of Ottawa Press.

Forcada, M. (2014) “On the annotation of tmx translation memories for advanced leveraging in computer-aided translation” Proceedings of the LREC’14. Reykjavik, Iceland

Gomez, Josu (2001) “Una guía al TMX”. Revista Tradumàtica. Núm. 0. http://www.fti.uab.es/tradumatica/revista

Levenshtein, Vladimir I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals". Soviet Physics Doklady 10 (8): 707–710.

Oliver, Antoni (2015). Herramientas tecnológicas para traductores. Editorial UOC.

Padró, L.; Stanilovsky, E. (2012) “FreeLing 3.0: Towards Wider Multilinguality” Proceedings of the International Conference on Language Resources and Evaluation. "LREC2012". Istanbul: 2012.

Porter, M. (1980) “An algorithm for suffix stripping”. Program 14.3 (1980): 130-137

Raya, Rodolfo M. (2007) Reduce translation time and effort with the aid of XML standards http://www.maxprograms.com/articles/tmx.html

Savourel, Yves (2005). TMX 1.4g Specification. Localization Industry Standards Associacion. http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

Simard, M. (2003) “Translation spotting for translation memories”. Proceedings of NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, p. 65–72, Edmonton, AB, Canada

Somers, H. (ed.) (2003) Computers and translation: a translator’s guide. John Benjamins Publishing Company. Philadelphia, PA, USA. ISBN 9789027296696