Universitat Oberta de Catalunya

Trabajo Final de Máster en Traducción Especializada

Creación de una base de datos terminológicos especializada del Eurocódigo 7 inglés-español

Autora: Lara Caballero Freyer

Tutor académico: Antoni Oliver

Fecha de entrega: 18 de junio de 2018

Ciudad: Las Palmas de Gran Canaria

Índice

Creación de una base de datos terminológicos especializada del Eurocódigo 7 inglés-español

Creating an English-Spanish specialized terminology database of Eurocode 7

Autora: Lara Caballero Freyer
Tutor académico: Antoni Oliver
Universidad Oberta de Catalunya
lkfreyer@gmail.com

Resumen: La informática proporciona la posibilidad de contar con un instrumento de trabajo que permite una búsqueda rápida y eficaz de términos especializados: la base de datos terminológicos. El hecho de disponer de una base de datos que ofrezca terminología especializada del Eurocódigo 7, que además se ha unificado a nivel europeo, es un recurso muy valioso para utilizar en traducciones de textos técnicos dentro del ámbito de la edificación e ingeniería civil, subámbito de geotecnia. El presente trabajo final de máster describe una propuesta de obtención de base de datos terminológicos bilingüe inglés-español en el formato de intercambio de datos terminológicos (TBX), que enumera los términos especializados que recoje la norma europea *Eurocódigo 7: Proyecto geotécnico*, para utilizarla como recurso de traducción. Esta normativa consta de tres partes y es posible disponer de las mismas en formato de documento portátil (PDF). Por un lado se obtienen los documentos en texto plano (TXT) mediante el conversor en línea *Convert PDF To Word .net*, que se retocan a mano con el editor de textos *Notepad++*. Una vez están limpios los documentos TXT, se unen todas las partes de la normativa, tanto para el inglés como para el español, se segmentan y alinean con la ayuda de la herramienta *LF-Aligner* para obtener un corpus paralelo en formato de intercambio de memorias de traducción (TMX). Por otro lado se crea un glosario de referencia en formato de texto separado por tabuladores a partir de una búsqueda y selección de glosarios técnicos bilingües relevantes. Para la búsqueda de candidatos a término se utiliza *TBXTools (Oliver and Vàquez 2015)*, una herramienta que se puede ejecutar en cualquier sistema operativo que tenga un intérprete de Python instalado y que utiliza el sistema de gestión de bases de datos relacional *SQLite* compatible con *ACID*. Con *TBXTools* se extrae de forma automática la terminología especializada con tres tareas distintas: a) extracción de terminología monolingüe mediante una estrategia estadística con la ayuda de una lista de palabras vacías como filtro; b) extracción de terminología monolingüe mediante una estrategia lingüística con la ayuda del corpus paralelo de todo el Eurocódigo 7 y un glosario de referencia; c) detección de equivalentes de traducción dentro de ese corpus paralelo con la utilización de la lista de palabras vacías para obtener mejores resultados. Una vez llevado a cabo el proceso de extracción, se evaluan los resultados y se seleccionan los candidatos a término obtenidos con las tareas de extracción automática de terminología que se presentan. Finalmente se crea la base de datos terminológicos en el formato estándar TBX.
Palabras clave: Eurocódigo 7, extracción automática de terminología, corpus paralelo, base de datos terminológicos, TBXTools.

Abstract: Computer science provides the possibility of having a working tool, letting us search specialized terms quickly and effectively: the terminology database. Having a database that offers specialized terms of Eurocode 7 that has also been unified at European Union level is a very valuable resource to use in translations of technical texts within the field of building and civil engineering, geotechnical subarea. This master's degree paper describes a proposal for getting a bilingual English-Spanish terminology database in TermBase eXchange format (TBX), listing specialised terms of the European standard *Eurocode 7: Geotechnical project* to be used as a translation resource. This standard consists of three parts, which are available in Portable Document Format (PDF). On the one hand, documents are obtained in plain text (TXT) using the Convert PDF to Word .net online converter. They are manually retouched with the Notepad++ text editor. Once TXT documents are cleaned, all the parts of the standard are joined together, for both English and Spanish. Then they are

segmented and aligned with the help of the LF-aligner tool in order to obtain a parallel corpus in Translation Memory eXchange format (TMX). On the other hand, a reference glossary is created in tab-separated text format from a search and selection of relevant bilingual technical glossaries. TBXTools (Oliver and Vàquez 2015) is used for searching for term candidates, a tool that can be run on any operating system that has a Python interpreter installed and uses the ACID-compatible database management system SQLite. With TBXTools, the specialized terminology is extracted automatically with three different tasks: a) monolingual terminology extraction using a statistical strategy with the help of a list of stop-words as a filter; b) monolingual terminology extraction using a linguistic strategy with the help of the parallel corpus of the whole Eurocode 7 and a reference glossary; c) automatic translation equivalent detection in that parallel corpus using the list of stop-words for best results. Once the extraction process is carried out, results are evaluated. Term candidates, obtained with the automatic terminology extraction tasks given, are selected. Finally, the terminology database in standard TBX format is created.
Keywords: Eurocode 7, automatic terminology extraction, terminology database, parallel corpus, TBXTools.

1. Introducción

Hoy en día es posible el procesamiento de un corpus representativo y su representación en bases de datos terminológicos mediante operaciones asistidas por ordenador.

El presente artículo explica y describe en detalle una propuesta de creación de base de datos terminológicos especializada mediante el uso de la tecnología.

Aunque trata un subámbito específico como es la geotecnia, la propuesta que ofrece es aplicable y válida para cualquier otro campo especializado.

Tiene como objetivo servir de ayuda al profesional que aún no dispone de experiencia en la elaboración de este tipo de recursos de traducción, así como servir de inspiración para aquel que aún teniendo experiencia le gustaría incrementar su productividad con la utilización de las herramientas y metodología propuestas.

La extracción automática de terminología en inglés y la detección automática de equivalentes de traducción en español se llevan a cabo mediante TBXTools (Oliver and Vàquez 2015), una herramienta rápida y flexible, con clases de Python configurables que se puede ejecutar en cualquier sistema operativo que tenga un intérprete de Python instalado y que utiliza el sistema de gestión de bases de datos relacional SQLite compatible con ACID.

Los candidatos a término se extraen de forma automática mediante varias tareas con estrategias estadísticas y lingüísticas, con la ayuda de una lista de palabras vacías, un corpus paralelo y un glosario de referencia.

En los siguientes apartados se describen los requisitos necesarios y se enumeran las distintas fases que se proponen seguir para el cumplimiento del objetivo de este trabajo, enumeradas de forma cronológica de acuerdo con el avance del trabajo.

Se presenta brevemente la norma europea «Eurocódigo 7: Proyecto geotécnico», para conocer el tipo de terminología que se planea extraer para formar parte de esta base de datos.

Luego se expone la fase preliminar: la necesidad de que toda la documentación esté disponible como archivo de texto, donde se describe la disponibilidad de documentación y las estrategias de conversión de formatos.

Los siguientes apartados explican la creación de un corpus paralelo de todo el Eurocódigo 7, la elaboración de un glosario de referencia que ayude a componer un patrón lingüístico aprendido, la extracción automática de candidatos a término en inglés mediante estrategias estadísticas y lingüísticas, así como la detección de equivalentes de traducción en español dentro del corpus.

Por último, se describe la creación de una base de datos terminológicos especializada del Eurocódigo 7 inglés-español para utilizarla en formato de intercambio de datos terminológicos *(TermBase eXchange).*

La razón por la cual se ha escogido el formato TBX es debido a su formato estándar de código abierto para datos terminológicos, basado en XML, que facilita el intercambio entre bases terminológicas con la posibilidad de importación y exportación en paquetes de software que incluyan bases de datos terminológicos, además de estar basado en tres normativas ISO: ISO 12620, ISO 12200 y ISO 16642 (Vasiljevs, Rirdance y Liedskalnins, 2008).

2. Eurocódigo 7

El Eurocódigo 7 es una norma europea para la edificación e ingeniería civil que pretende unificar

procedimientos que garanticen la seguridad y funcionalidad de una estructura.

Esta normativa se encuentra dividida en tres partes, donde la primera describe las reglas generales, la segunda los ensayos de laboratorio y la tercera los ensayos de campo.

Los principios y reglas que recoge se ponen en práctica para confeccionar un proyecto geotécnico que se incluye en el proyecto de edificación o ingeniería civil.

La finalidad de cada proyecto geotécnico es la descripción del comportamiento del terreno frente a la naturaleza y a las cargas, del lugar donde yace o se va a edificar una estructura y de una solución de cimentación para esa estructura.

Los términos especializados que constituyen la base de datos terminológicos objetivo de este trabajo están relacionados con el terreno donde se puede ubicar una estructura y los ensayos que demuestran el comportamiento de la estructura sobre el terreno.

3. Materiales

En este apartado se describen los recursos materiales utilizados, su disponibilidad y características de formato.

El recurso material principal que se utiliza en este trabajo es el Eurocódigo 7, en sus versiones inglesa y española.

Es posible obtener la versión inglesa en formato de documento portátil (PDF) en este enlace de la Unión Europea.

En cuanto a la versión española, dos calculistas de estructuras de hormigón, acero y madera en edificación permiten bajar de su página la parte 1, parte 2 y parte 3 en formato PDF.

Otro recurso material que se utiliza en este trabajo es el informe preparado por el Comité ACI 116, que constituye un glosario para la tecnología del cemento y el hormigón, repleto de términos que también se utilizan habitualmente en los proyectos geotécnicos.

Así mismo se utiliza el Diccionario Técnico Vial de la A.I.P.C.R. De la Asociación Técnica de Carreteras, en especial los siguientes capítulos:

9 01 Materiales
9 02 Suelos y áridos
11 Cualidades, defectos y ensayos

4. Herramientas de trabajo

Este apartado enumera las herramientas utilizadas en la creación de la base de datos terminológicos con una breve descripción de su función en este trabajo:

- Notepad++: editor de textos para la edición de documentos de texto y apertura de documentos para su evaluación
- IATE: base de datos interactiva con terminología inter-institucional de la Unión Europea para la extracción de glosarios
- LF-Aligner: programa de alineación automática, que facilita el uso de Hunalign mediante un diálogo textual, para la segmentación de documentos, su alineación y creación del corpus paralelo
- ApSIC Xbench: programa de gestión terminológica y control de calidad para la conversión de TBX a TXT
- Apache OpenOffice Calc: hoja de cálculo del paquete ofimático Apache OpenOffice para el borrado de la información adicional del glosario de referencia y para la conversión de TXT a CSV
- TBXTools: herramienta ejecutable con Python, que utiliza el sistema de gestión de bases de datos relacional SQLite compatible con ACID y realiza tareas relacionadas con la extracción automática de terminología, para la extracción automática de terminología especializada que constituye la base de datos terminológicos
- Python: lenguaje de programación multiplataforma para la interpretación de TBXTools
- NLTK: conjunto de bibliotecas y programas para el procesamiento del lenguaje natural simbólico y estadístico para el lenguaje de programación de Python
- FreeLing: biblioteca en C++ con funcionalidades de análisis de idiomas para el etiquetado del texto
- Anchovy de MaxPrograms: editor de glosarios multiplataforma para la conversión de CSV a GlossML y de GlossML a TBX
- OmegaT: herramienta de memoria de traducción para comprobar la funcionalidad de la base de datos en archivo TBX

5. Archivos de texto

La obtencion de archivos de texto varía según la disponiblidad de documentos.

Por un lado, hay cinco documentos PDF:

- UNE-EN 1997-1: 2010
- UNE-ENV 1997-2: 2001
- UNE-ENV 1997-3: 2002
- Informe preparado por el Comité ACI 116
- Diccionario Técnico Vial de la A.I.P.C.R. De la Asociación Técnica de Carreteras

Que son las tres partes de la versión española del Eurocódigo 7 y dos de los documentos que se necesitan para crear el glosario de referencia.

Se requiere que estos documentos tengan extensión de archivo TXT, para lo cual se opta por la utilización del conversor en línea Convert PDF To Word .net.

El uso de este conversor es sencillo: en primer lugar se elige en la parte superior de la página `PDF to TXT`, después se sube el archivo que se desea convertir mediante un clic en el botón `[Examinar...]` y una vez aparece el mensaje sobre el éxito de conversión *(File converted successfully!),* el archivo de descarga se obtiene tras hacer clic en el botón `[Download]` (véase la Figura 1).

Figura 1: Conversión de PDF a TXT mediante *Convert PDF To Word.net*

Es posible encontrar el documento convertido con la extensión TXT en la carpeta `Descargas` del equipo.

Por otro lado, las tres partes de la versión inglesa del Eurocódigo 7 se obtienen de varias formas: la primera parte puede bajarse al equipo como archivo de texto a través de la biblioteca digital Internet Archive (véase la Figura 2).

Figura 2: Opción de descarga de la versión inglesa de la parte 1 del Eurocódigo 2 como archivo de texto

Aunque también es posible descargar el archivo de texto de la parte 2 de esta biblioteca, tan sólo para la versión más actualizada del año 2007 (EN 1997-2: 2007) y no para la versión del año 2000 (ENV 1997-2: 2000) que equivale plenamente a la versión española del año 2001 (UNE-ENV 1997-2: 2001).

La parte 2 de la versión inglesa (ENV 1997-2: 2000) que equivale plenamente a la parte 2 de la versión española está disponible en PDF y se trata de convertir a TXT mediante el conversor, sin embargo la conversión no es satisfactoria, ya que el conversor no es capaz de convertir todo el documento a archivo de texto. Tras un intento de desbloqueo del PDF mediante el software PDF Unlocker sin éxito, se opta finalmente por escribir las 107 páginas de este documento a mano en Notepad++ y se guarda en TXT, lo cual supone una carga de trabajo importante.

La parte 3 no se encuentra en esta biblioteca digital, sin embargo, puede visualizarse en el sitio web geotechnicaldesign.info y mediante un copia y pega, haciendo clic en cada uno de sus apartados, es posible crear un archivo TXT limpio de esta parte.

Una vez se dispone de estos ocho archivos de texto, es necesario retocarlos: primero se abre cada documento en Notepad++, luego se borran los saltos de línea en cada línea, excepto los saltos de párrafo, se borran las cabeceras y por último los pies de página.

Aunque esta fase de retoque es tediosa, es importante una buena ejecución de la misma para obtener mejores resultados en fases posteriores.

6. *Corpus paralelo del Eurocódigo 7*

El objetivo de este apartado es la obtención de un corpus paralelo de todo el Eurocódigo 7.

El primer paso consiste en unir en un único documento los archivos de texto de las tres partes del Eurocódigo 7 seguidas, una detrás de otra, para el inglés y en otro documento para el español.

El siguiente paso consiste en utilizar el programa de alineación automática LF-Aligner.

Este programa se ejecuta haciendo doble clic en `LF_aligner_4.1.exe`.

Primero aparece una pantalla de elección del tipo de archivo de entrada, aquí se marca `txt (UTF-8!)`.

En la siguiente pantalla se especifican los idiomas: *English* para *Language 1* y *Spanish* para *Language 2*.

Después pide el archivo correspondiente para cada una de las lenguas, se recomienda que la ruta sea corta y no contenga espacios.

Se ejecuta la alineación y aparece la pantalla que pregunta si se desean utilizar las versiones

segmentadas por párrafos o bien las segmentadas por oraciones, a lo cual se responde marcando la primera opción, el uso de la versión segmentada por oraciones.

A continuación se marca la opción de revisión de resultados de alineación y aparece la pantalla de la Figura 3.

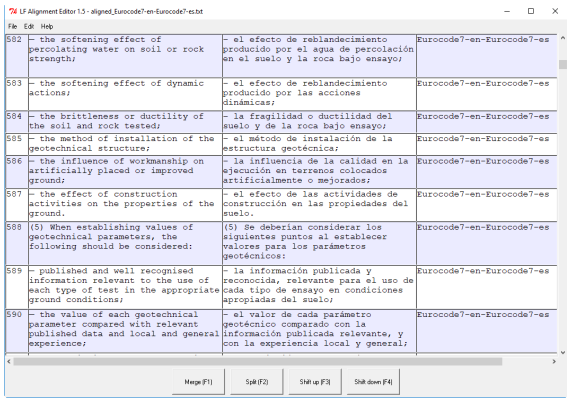

Figura 3: Editor gráfico de los documentos alineados

Mientras se revisa es posible la unión, la separación y el desplazamiento de párrafos mediante el uso de los botones que se encuentran en la parte inferior de la pantalla.

Una vez se termina la revisión se cierra esta pantalla y se marca la casilla de generación de memorias de traducción en TMX.

Se acepta la propuesta que ofrece la configuración para la memoria y aparece una pantalla que confirma que la alineación se ha realizado correctamente con información sobre la segmentación.

La memoria de traducción en TMX se genera en el mismo directorio donde están los archivos originales, dentro de una carpeta que se designa como `align_yyyy.mm.dd_hh.mm.ss`, y el archivo recibe el nombre de `Eurocode7-en-Eurocode7-es.tmx`.

7. *Glosario de referencia*

Se realiza una búsqueda de glosarios relevantes para elaborar un glosario de referencia del cual se pueda obtener un patrón de referencia para utilizarlo en la extracción lingüística con patrón aprendido.

Se han seleccionado los siguientes glosarios relevantes con terminología común a los proyectos geotécnicos:

- Informe preparado por el Comité ACI 116
- Diccionario Técnico Vial de la A.I.P.C.R. De la Asociación Técnica de Carreteras
- IATE-4816-eng-spa
- IATE-6831-eng-spa

En el apartado 5 ya se enumeran dos de los glosarios escogidos, donde se explica que están disponibles en PDF y se convierten a archivos de texto.

Los otros dos glosarios se extraen de la IATE siguiendo las instrucciones que facilita en su página.

Tras la consulta de los códigos que proporciona la IATE, se extraen los siguientes glosarios en TBX:

`4816 - Land transport`

`6831 - Building and public works`

Que contienen terminología de interés para enriquecer el contenido del glosario de referencia.

Mediante ApSIC Xbench se realiza la conversión de TBX a archivo de texto delimitado por tabuladores (`Tab-delimited Text File`).

Al abrir estos archivos TBX mediante Notepad++ es posible visualizar información adicional sin interés para el glosario (véase la Figura 4), la cual se borra con rapidez pegando el texto del documento en una hoja de cálculo de OpenOffice Calc y eliminando las columnas que contienen esa información.

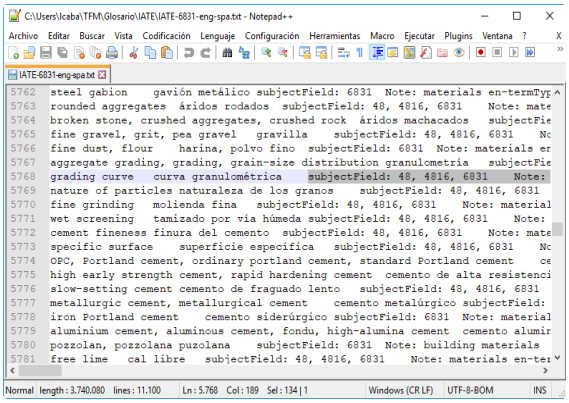

Figura 4: Información adicional sin interés resaltada en color gris

La información de interés se guarda como archivo de texto: `IATE-4816-eng-spa.txt` e `IATE-6831-eng-spa.txt`

Estos cuatro glosarios guardados como archivos de texto delimitado por tabuladores se copian seguidos uno detrás de otro en un único documento de texto para obtener el glosario de referencia con nombre `glosario-en-es.txt`.

Para disponer únicamente de los términos del glosario en inglés se copia todo el texto de `glosario-en-es.txt` en una hoja de cálculo de OpenOffice Calc, desde donde se copia la columna de términos en inglés y se pega en un nuevo documento de texto con nombre `glosario-en.txt`, se retoca de forma que aparezca un término por línea, ya que a veces

hay dos o tres términos separados por coma y se guarda como `glosario-mod-eng.txt`.

Este glosario de referencia `glosario-mod-eng.txt` con términos en inglés es el que se utiliza para crear un patrón aprendido en el apartado 11.

8. Extracción estadística

En esta fase se extraen candidatos a término en inglés mediante una estrategia estadística con la ayuda de una lista de palabras vacías como filtro.

Se requiere que en un directorio del equipo (p.ej.: `C:\TBXTools\STE`) se guarden los siguientes documentos, archivos y programas:

- `estadistic1.py`
- `Eurocode7.en`
- `Eurocode7.es`
- `stop-eng.txt`
- `TBXTools.py`

Que en definitiva se trata, por orden de aparición: del programa que extrae términos en inglés, de los dos archivos de texto plano que constituyen el corpus del Eurocódigo 7, de la lista de palabras vacías en inglés y del programa principal.

Una vez se ejecuta `estadistic1.py` en `Símbolo del sistema`, se genera una lista de candidatos que se designa como `candidates-eng.txt` (véase la Figura 5).

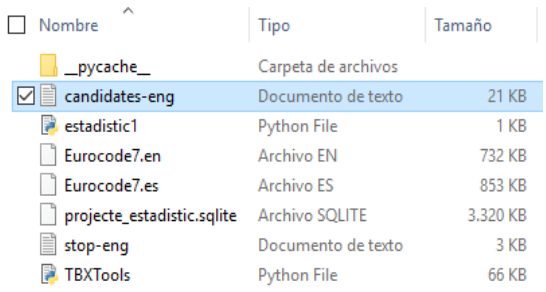

Figura 5: Tras ejecutar `estadistic1.py` se genera la lista de candidatos a término en inglés

Se abre la lista de candidatos con Notepad++ para ver los resultados (véase la Figura 6).

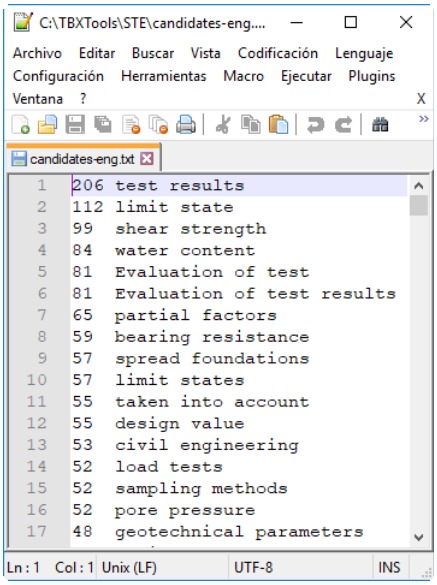

Figura 6: Lista de candidatos a término mediante una estrategia estadística abierta con Notepad++

Esta lista está formada por 971 candidatos a término en inglés enumerados en orden decreciente según la frecuencia con la que aparecen en el Eurocódigo 7 y con el número de repeticiones delante de cada candidato.

9. Extracción lingüística

Se trata de extraer candidatos a término en inglés mediante una estrategia lingüística con la ayuda del corpus paralelo de todo el Eurocódigo 7.

Para hacer la extracción lingüística, primero hay que etiquetar el texto.

Se requiere que en un directorio del equipo (p.ej.: `C:\TBXTools\LTE`) se guarden los siguientes archivos y programas:

- `etiquetatge.py`
- `Eurocode7.en`
- `Eurocode7.es`
- `TBXTools.py`

Que en definitiva se trata, por orden de aparición: del programa que hace de etiquetador, de los dos archivos de texto plano que constituyen el corpus del Eurocódigo 7 y del programa principal.

Una vez se ejecuta `etiquetatge.py` en Símbolo del sistema, este etiquetador crea un proyecto con el corpus del Eurocódigo 7 que designa `projecte_linguistic.sqlite` (véase la Figura 7), lo etiqueta y queda en el proyecto de `TBXTools`.

Si bien no saca nada específico por pantalla, cada token tiene lema, forma y etiqueta separada por «|» de la siguiente manera:

```
The|the|DT wall|wall|NN of|of|IN the|
```

```
the|DT borehole|borehole|NN should|
should|MD be|be|VB checked|check|VBN
for|for|IN fissures|fissure|NNS ,|,|Fc
voids|void|NNS and|and|CC rock|rock|NN
fragments|fragment|NNS ,|,|Fc that|
that|WDT might|may|MD damage|damage|VB
the|the|DT flexible|flexible|JJ
membrane|membrane|NN or|or|CC trap|
trap|NN the|the|DT probe|probe|NN .|.|
Fp
```

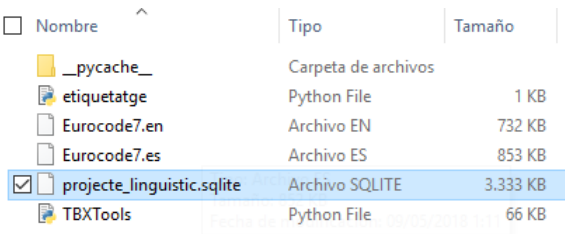

Figura 7: Tras ejecutar `etiquetatge.py` se genera `project_linguistic.sqlite` (el proyecto con el corpus)

La extracción lingüística se realiza con la siguiente serie de patrones lingüísticos que se guardan en el documento de texto `patterns.txt`:

```
#||NN |#|NN
#||JJ |#|NN
#||NN |#|NN
#||NN #||NN |#|NN
#||JJ |#|NN
#||JJ #||JJ |#|NN
#||JJ |#|NN |#|NN
```

El símbolo # se utiliza para expresar el campo que ocupa el lugar del término (Oliver, 2018), para entender el resto del patrón se presentan algunos ejemplos de patrones morfológicos junto con su explicación en la Tabla 1.

`#||NN.? |#|NN.?`	Un nombre (que adopta forma de palabra) seguido de un nombre (que adopta el lema)
`#||JJ.? |#| NN.?`	Un adjetivo (que adopta forma de palabra) seguido de un nombre (que adopta el lema)
`|#||NN.? #|of|IN |#|NN.?`	Un nombre (que adopta forma de palabra) seguido de `of` seguido de un nombre que adopta el lema

Tabla 1: Ejemplos de patrones morfológicos (extraída de Oliver, 2018)

El siguiente paso consiste en extraer los candidatos a término con el patrón lingüístico `patterns.txt` mediante el programa `linguistic-extract.py`, para lo cual se requiere que se guarden ambos en `C:\TBXTools\LTE` (véase la Figura 8).

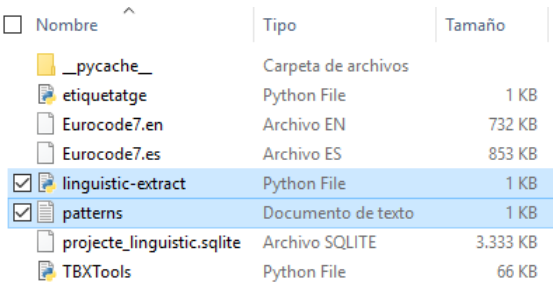

Figura 8: Se copian los patrones lingüísticos y el programa que extrae los candidatos en `C:\TBXTools\LTE`

Una vez se ejecuta el extractor de candidatos lingüísticos en Símbolo del sistema, aparece en el directorio la lista `candidates-linguistic-eng.txt` (véase la Figura 9).

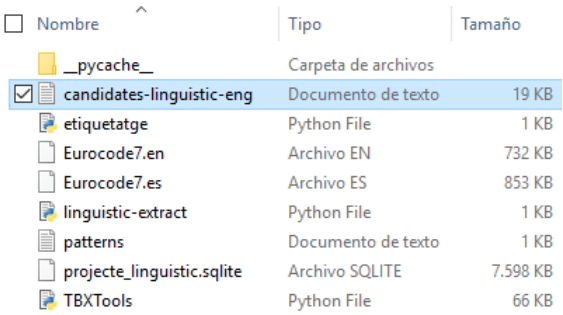

Figura 9: Tras ejecutar `linguistic-extract.py` se genera la lista de candidatos

Cuyo contenido se visualiza mediante Notepad++ en la Figura 10.

Esta lista que ha creado `linguistic-extract.py` tiene la característica de que ahora los candidatos se limitan a 1000, se enumeran en orden decreciente según la frecuencia con la que aparecen en el Eurocódigo 7 y con el número de repeticiones delante de cada candidato.

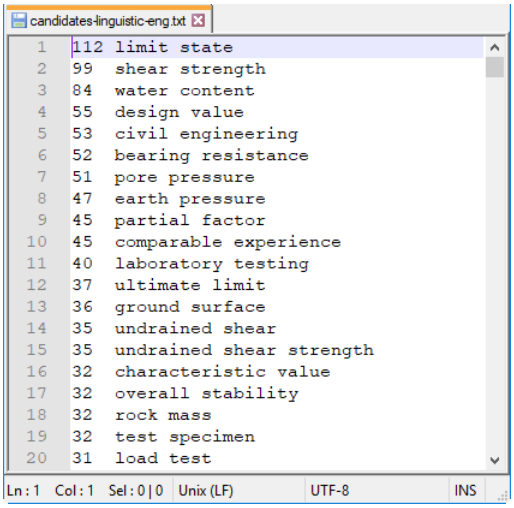

Figura 10: Lista de candidatos a término mediante una estrategia lingüística

10. Aprendizaje de patrones lingüísticos

Se trata de aprender patrones lingüísticos a partir del glosario de referencia `glosario-mod-eng.txt` obtenido en el apartado 7.

Primero se crea un duplicado de la subcarpeta `\LTE`, a partir de ahora con nombre `\LTE-learnt-patterns` para que no se sobrescriban los datos nuevos de `candidates-linguistic-eng` del directorio anterior.

En `C:\TBXTools\LTE-learnt-patterns` se guarda el programa `learn-patterns.py` y el glosario de referencia `glosario-mod-eng.txt` (véase la Figura 11) de forma que el directorio contenga el corpus, el corpus etiquetado y el programa que mira si en el corpus etiquetado hay alguno de los términos de la lista y si es que sí, aprende el patrón.

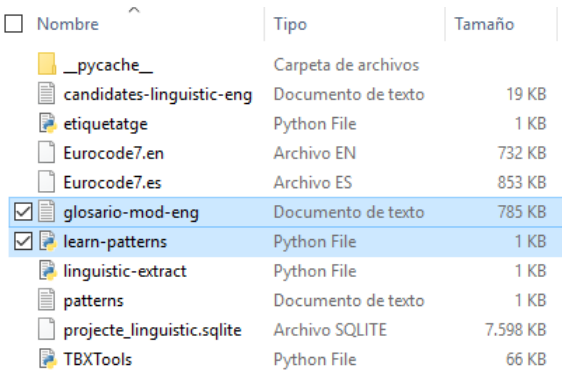

Figura 11: Se guardan `learn-patterns.py` y `glosario-mod-eng.txt` en `C:\TBXTools\LTE-learnt-patterns`

Una vez se ejecuta el programa `learn-patterns.py` en Símbolo del sistema, aparecen los patrones que ha aprendido en pantalla (véase la Figura 12) y como documento de texto en `C:\TBXTools\LTE-learnt-patterns` (véase la Figura 13) con nombre `learnt-patterns.txt`.

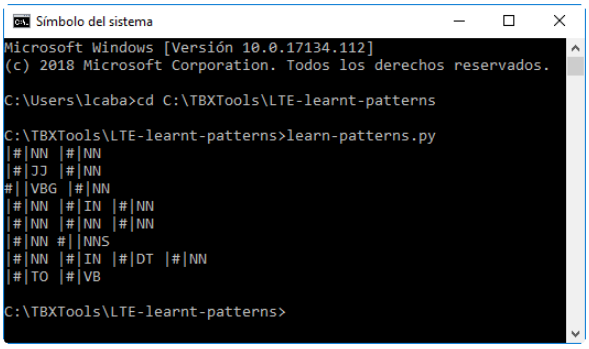

Figura 12: Creación de patrón aprendido y extracción de términos

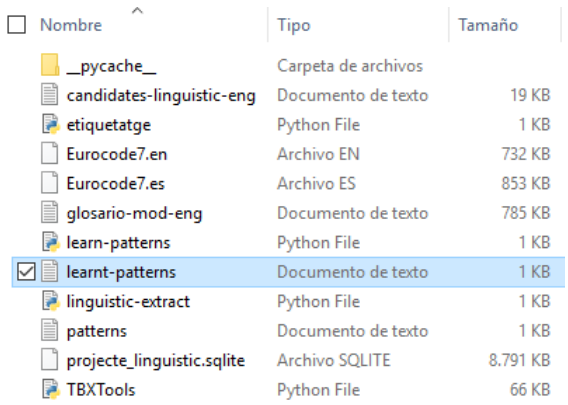

Figura 13: Al ejecutar `learn-patterns.py` se crea el documento con los patrones aprendidos `learnt-patterns.txt`

11. Extracción lingüística con patrones aprendidos

Ahora es posible hacer una nueva extracción lingüística de candidatos utilizando estos patrones aprendidos.

Hay que guardar en `C:\TBXTools\LTE-learnt-patterns` el programa `linguistic-extract.py` ligeramente modificado, ahora es capaz de cargar de nuevo el glosario inglés modificado y si algún candidato coincide lo marca con un «*» delante.

Se ejecuta el programa `linguistic-extract.py` en Símbolo del sistema y se genera un nuevo documento de texto que recoge los candidatos extraídos en `candidates-linguistic-eng.txt` (véase la Figura 14).

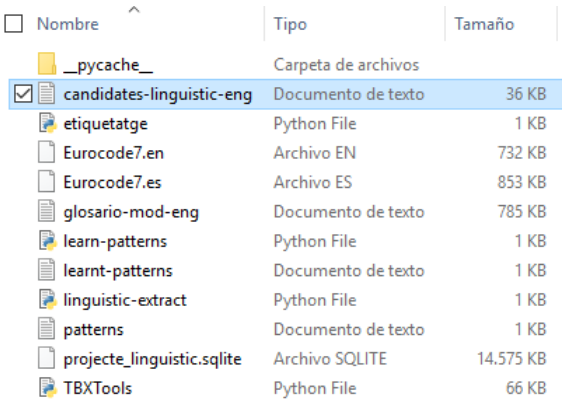

Figura 14: Extracción de candidatos a término con un patrón aprendido

Se abre la lista de candidatos en Notepad++ y mediante la función de búsqueda del editor se realiza un recuento de candidatos marcados con «*».

La lista está constituida por 2128 candidatos de los cuales 167 aparecen marcados con «*».

En el cuarto puesto de la lista se puede ver el primer candidato a término marcado (véase la Figura 15).

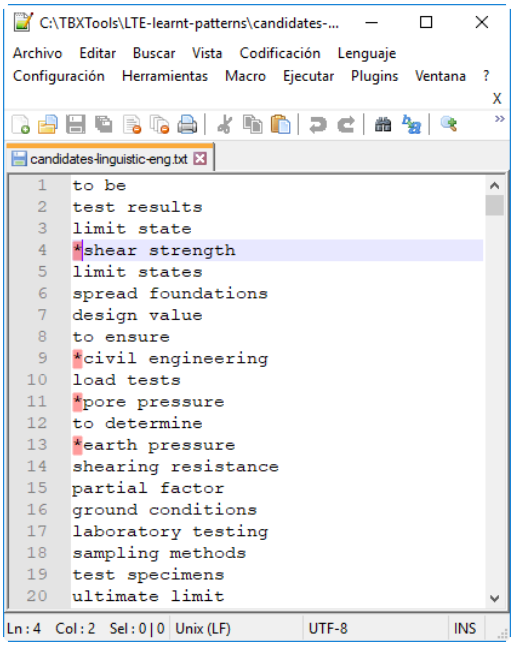

Figura 15: Lista de candidatos a término con un patrón aprendido

Se revisan los primeros 1000 candidatos, desmarcando aquellos que no son términos especializados sobre proyectos geotécnicos y marcando los que sí lo son con la marca «*» delante de cada candidato.

Se han desmarcado los siguientes candidatos, considerados más generales de la ingeniería civil y no tan específicos de los proyectos geotécnicos: *to check, water content, base area, to take, construction process, construction materials, pile foundations, structural design, construction supervision, finite element, strength properties, to serve y construction material.*

De la misma forma, se han marcado algunos candidatos más como *shearing resistance* o *ground conditions*, entre otros.

A este documento de texto editado se le designa como `candidates-linguistic-mod-eng.txt`.

12. Detección en corpus paralelo

El punto de partida de esta fase es la lista modificada de candidatos a término en inglés marcados con «*» delante (`candidates-linguistic-mod-eng.txt`).

El objetivo aquí es buscar dentro del corpus paralelo de todo el Eurocódigo 7 los equivalentes de traducción en español de los candidatos a término en inglés de esa lista modificada.

A fin de obtener mejores resultados en la búsqueda, se utilizan dos listas de palabras vacías, una para el inglés `stop-eng.txt` y otra para el español `stop-spa.txt`.

Se requiere que en un directorio del equipo (p.ej.: `C:\TBXTools\BTE`) se guarden los siguientes documentos, archivos y programas:

- `candidates-linguistic-mod-eng.txt`
- `Eurocode7.en`
- `Eurocode7.es`
- `stop-eng.txt`
- `stop-spa.txt`
- `TBXTools.py`
- `tond1.py`

Que en definitiva se trata, por orden de aparición, de la lista modificada de candidatos a término marcados, de los dos archivos de texto plano que constituyen el corpus del Eurocódigo 7, de las dos listas de palabras vacías, del programa principal que tiene implementada la búsqueda en corpus paralelos y del programa que busca el equivalente a cada término de la lista.

Al ejecutar `tond1.py` en Símbolo del sistema, comienzan a aparecer en pantalla todos los candidatos con su traducción más probable al lado, delimitados estos mediante tabulador seguidos de hasta 10 traducciones alternativas separadas por «:» (véase la Figura 16).

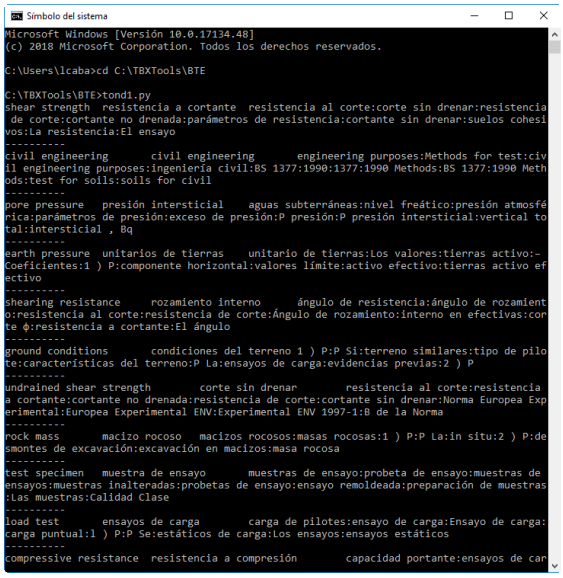

Figura 16: Al ejecutar `tond1.py` aparece en pantalla los equivalentes de traducción

Una vez se finaliza la búsqueda, aparece en `C:\TBXTools\BTE` la lista que recoge los candidatos con sus equivalentes con nombre `candidates-eng-spa.txt` (véase la Figura 17).

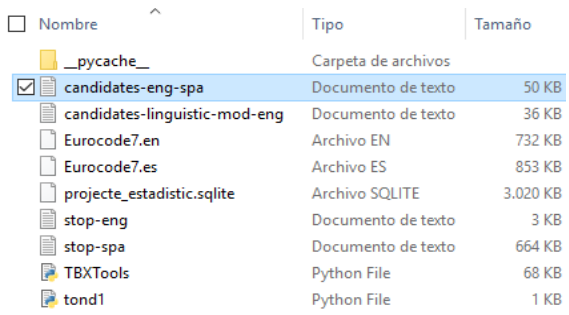

Figura 17: Tras ejecutar `tond1.py` aparece en el directorio la lista de candidatos con sus equivalentes

Se abre `candidates-eng-spa.txt` con Notepad++ para visualizar todos los candidatos a término que aparecían marcados con «*» en `candidates-linguistic-mod-eng.txt` con su equivalente al lado, separados mediante tabulador, seguidos de más propuestas de equivalentes delimitadas por «:» (véase la Figura 18).

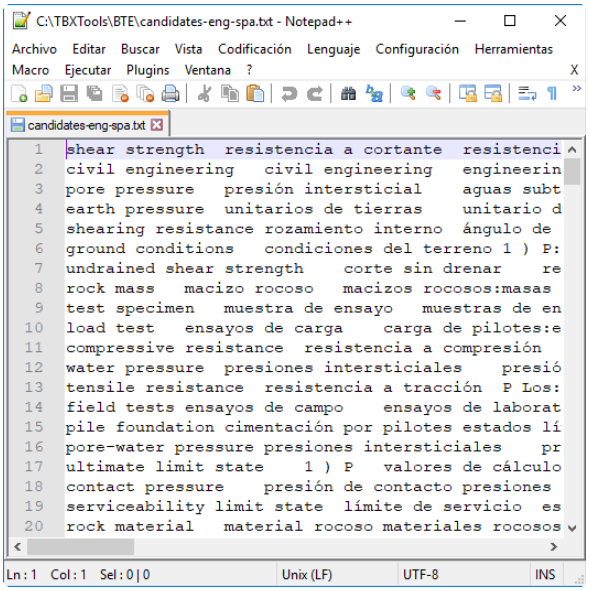

Figura 18: Lista de candidatos con su equivalente y más propuestas

Se comprueba la traducción de cada candidato de la lista `candidates-eng-spa.txt`, poniendo la traducción correcta en el segundo campo (en caso de que no sea la correcta).

Se ha sustituido el plural por el singular, se han quitado algunos candidatos (*general arrangement, time period, cross-sectional, cross section, visual signal, volume change*...) por no considerarse específicos únicamente de los proyectos geotécnicos.

También se han añadido candidatos que ya aparecían en `candidates-linguistic-mod-eng.txt` pero sin la marca «*» delante: *filter tip, galvanic cell, high air entry value filter* (éste aparecía marcado como *air entry*), *open system*...

A este documento de texto editado se le designa como `candidates-mod-eng-spa.txt`.

Después de revisar la lista y editarla, en caso necesario, se copia su contenido en una hoja de cálculo de OpenOffice Calc, se borran las columnas con los equivalentes adicionales, se ordenan alfabéticamente y se guarda como `terms-A-Z-eng-spa.ods`.

13. Base de datos terminológicos

En este apartado se explica la creación de una base de datos terminológicos con los términos recogidos en la hoja de cálculo `terms-A-Z-eng-spa.ods`.

Esta hoja de cálculo se guarda como Texto CSV en `Archivo > Guardar como` y en el siguiente mensaje que aparece se elige `[Mantener el formato actual]`, opción que al pulsar hace que se abra el cuadro de diálogo

`Exportar a un archivo de texto` que se completa como se indica en la Figura 19.

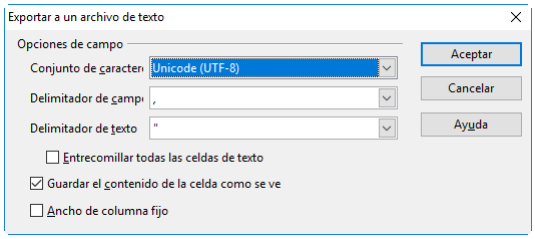

Figura 19: Exportación a CSV mediante OpenOffice Calc

Se abre el archivo de valores separados por comas `terms-A-Z-eng-spa.csv` en Notepad++ para comprobar que se ha exportado con éxito (véase la Figura 20).

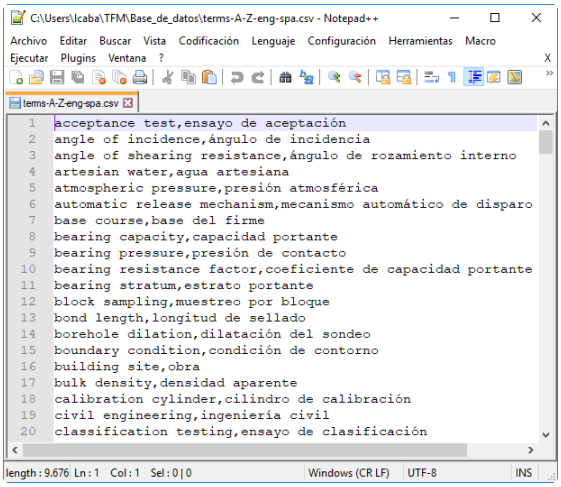

Figura 20: Archivo CSV

Para el siguiente paso se utiliza la herramienta Anchovy de MaxPrograms, donde se transforma el archivo `terms-A-Z-eng-spa.csv` a GlossML.

En primer lugar se hace clic en `Task > Convert CSV File to GlossML Format` para abrir el cuadro el diálogo `Convert to GlossML`, que se completa como se indica en la Figura 21.

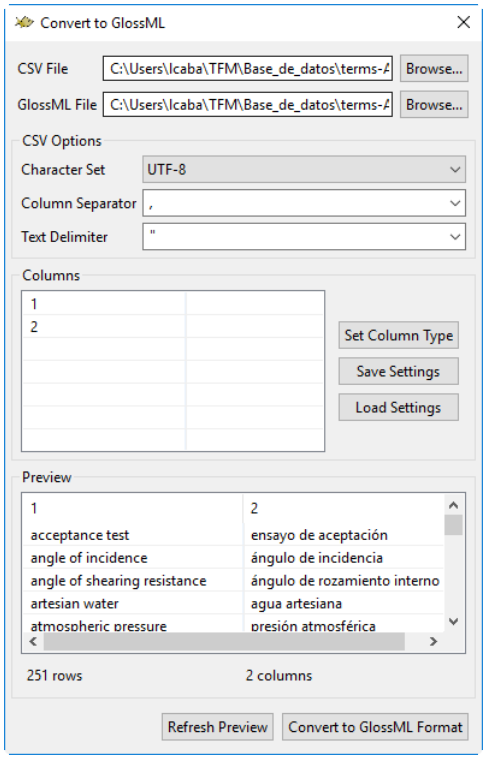

Figura 21: Conversión de CSV a GlossML con Anchovy de MaxPrograms

Se fija el tipo de las columnas 1 y 2 de la sección `Columns` haciendo clic en el botón `[Set Column Type]` para abrir el cuadro de diálogo `Column Type` (véase la Figura 22).

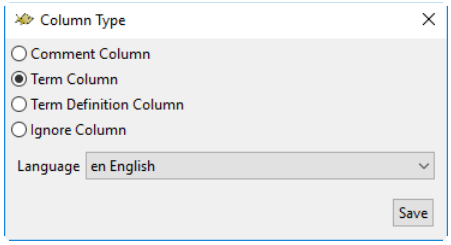

Figura 22: Fijación del tipo de columna 1; para la columna 2 se procede igual a excepción del idioma que se elige es *Spanish*

Después se abre con Anchovy el documento GLS que aparece en el directorio que se había indicado en `GlossML File` y se exporta a TBX con un clic en `File > Export as TBX` (véase la Figura 23).

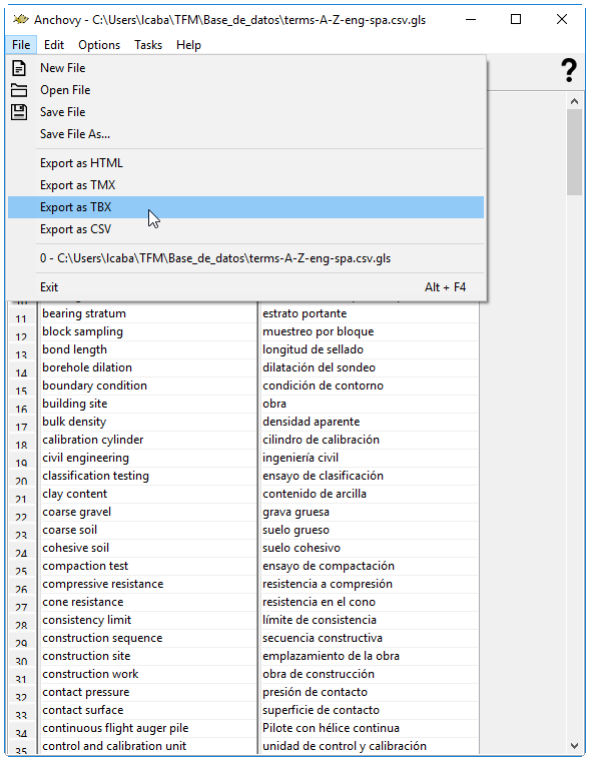

Figura 23: Exportación a TBX con Anchovy de MaxPrograms

Es posible comprobar en el mismo directorio que se ha creado correctamente y abrir el archivo TBX con Notepad++ (véase la Figura 24).

```
<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE martif SYSTEM "TBXcoreStructV02.dtd" >
<martif type="TBX" xml:lang="en">
  <martifHeader>
    <fileDesc>
      <sourceDesc>
        <p>Generated with Swordfish  3.3-20</p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <p type="XCSURI">http://www.ttt.org/oscarstar
    </encodingDesc>
  </martifHeader>
  <text>
    <body>
      <termEntry>
        <langSet xml:lang="en">
          <tig>
            <term>acceptance test</term>
          </tig>
        </langSet>
        <langSet xml:lang="es">
          <tig>
            <term>ensayo de aceptación</term>
          </tig>
        </langSet>
      </termEntry>
      <termEntry>
        <langSet xml:lang="en">
          <tig>
            <term>angle of incidence</term>
          </tig>
        </langSet>
        <langSet xml:lang="es">
          <tig>
            <term>ángulo de incidencia</term>
          </tig>
        </langSet>
      </termEntry>
```

Figura 24: Base de datos terminológicos como archivo TBX

Finalmente se realiza una prueba en OmegaT para comprobar la funcionalidad del archivo `terms-A-Z-eng-spa.tbx`, en este caso se ha creado un proyecto de OmegaT llamado `Muestra`, constituido por el archivo de texto `Eurocode7-en.txt` (véase la Figura 25).

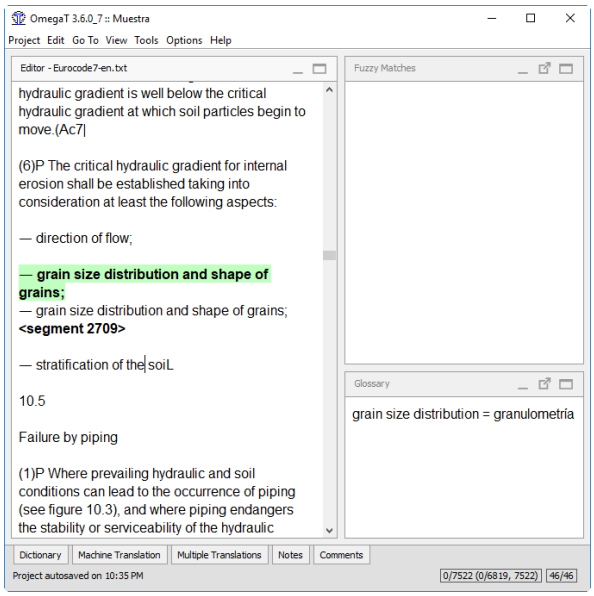
OmegaT 3.6.0_7 :: Muestra
Project Edit Go To View Tools Options Help
Editor - Eurocode7-en.txt
hydraulic gradient is well below the critical hydraulic gradient at which soil particles begin to move.(Ac7|
(6)P The critical hydraulic gradient for internal erosion shall be established taking into consideration at least the following aspects:
— direction of flow;
— grain size distribution and shape of grains;
— grain size distribution and shape of grains;
<segment 2709>
— stratification of the soiL
10.5
Failure by piping
(1)P Where prevailing hydraulic and soil conditions can lead to the occurrence of piping (see figure 10.3), and where piping endangers the stability or serviceability of the hydraulic
Fuzzy Matches
Glossary
grain size distribution = granulometría
Dictionary Machine Translation Multiple Translations Notes Comments
Project autosaved on 10:35 PM
0/7522 (0/6819, 7522) 46/46

Figura 25: Comprobación del funcionamiento del archivo TBX en OmegaT

Es posible comprobar que el archivo TBX funciona correctamente cuando se va a traducir un segmento que contiene un término de la base de datos, en el ejemplo: *grain size distribution*.

Por último, se hace una copia del archivo TBX con un nombre más corto: `terms-eng-spa.tbx`

El resultado final: una base de datos terminológicos con 251 términos del Eurocódigo 7 que se adjunta en versión electrónica como Anexo 1.

14. Evaluación y resultados

La terminología se extrae mediante varias estrategias: con la estadística se obtiene una lista de 977 candidatos, con la lingüística una lista límite de 1000 donde se comprueba la coincidencia de candidatos con una mejora de precisión respecto a la anterior.

Después se extraen términos con una estrategia lingüística con patrones aprendidos de un glosario de referencia y se obtienen 2128 candidatos de los cuales 167 aparecen marcados como favoritos, que se revisan y desmarcan aquellos que no son términos especializados sobre proyectos geotécnicos y se marcan los que sí lo son.

Tras la detección de equivalentes de traducción en español en el corpus del

Eurocódigo 7 se obtiene una lista de candidatos que se edita, cambiando plurales por singulares y evaluando si cada miembro es apto para ser incorporado como valor de la misma.

La combinación de tareas de extracción de terminología cumple su cometido realizando una extracción satisfactoria de todos los posibles candidatos a término que se encuentran dentro del Eurocódigo 7.

Finalmente se obtiene una base de datos de de 251 candidatos que se exporta como archivo TBX y está lista para utilizar como recurso de traducción.

15. Conclusiones

La base de datos terminológicos se ha creado con éxito.

Todas las herramientas que se han utilizado facilitan notablemente las distintas tareas de manera sencilla y eficaz, su uso es intuitivo y son de libre acceso.

Es una gran ventaja que herramientas como TBXTools existan para facilitar la tarea del profesional que necesita crear un recurso tan valioso como es la base de datos terminológicos especializada.

16. Bibliografía

Cabré, M. (1999) *Theory, methods and applications.* Johm Benjamins B.V., 160-193.

Comité Español de la A.I.P.C.R. (2002) *Diccionario Técnico Vial de la A.I.P.C.R.* Asociación Técnica de Carreteras, 124-139.

Comité Técnico CEN/TC 250 Eurocódigos estructurales (2010) *UNE-EN 1997-1:2010 Eurocódigo 7: Proyecto geotécnico. Parte 1: Reglas generales.* Asociación Española de Normalización y Certificación.

Comité Técnico CEN/TC 250 Eurocódigos estructurales (2001) *UNE-ENV 1997-2:2001 Eurocódigo 7: Proyecto geotécnico. Parte 2: Proyecto asistido por ensayos de laboratorio.* Asociación Española de Normalización y Certificación.

Comité Técnico CEN/TC 250 Eurocódigos estructurales (2002) *UNE-ENV 1997-3:2002 Eurocódigo 7: Proyecto geotécnico. Parte 3: Proyecto asistido por ensayos de campo.* Asociación Española de Normalización y Certificación.

De Irazazabal, A. Schwarz, E. (Sin fecha) *Las bases de datos terminológicas como ayuda al traductor.*

Gómez, A. Vargas, Ch. (2004) *Aspectos metodológicos para la elaboración de diccionarios especializados bilingües destinados al traductor.* Esletra, 365-398. Las palabras del traductor.

Mather, B. Babcock, H. Bollin, G. Dodson, V. Fiala, D. Gibbe, K. Henry, R. Hogan, M. Isabelle, H. Libby, J. Lorman, W. Mielenz, R. Morgan, A. Payne, E. Richards, O. Rutenbeck, T. Senbetta, E. Tuthill, L. (Sin fecha) *Terminología del cemento y el hormigón.*

Morais, J. L. (2013) *Cómo convertir de PDF a DOC.* La Linterna del Traductor, 8, 21-28.

Oliver, A. (2018) *Teaching and Researching Automatic Terminology Extraction with TBXTools.*

Oliver, A. Vàzquez, M. (2015) *TBXTools: A Free, Fast and Flexible Tool for Automatic Terminology Extraction.*

Pazienza, M.T. Pennacchiotti, M. Zanzotto, F. M. (2005) *Terminology Extraction: An Analysis of Linguistic and Statistical Approaches.* StudFuzz, 185, 255-279. Springer-Verlag Berlin Heidelberg.

Seghiri, M. (2017) *Corpus e interpretación biosanitaria: extracción terminológica basada en bitextos del campo de la Neurología para la fase documental del intérprete.* Panace, 18 (46), 123-132.

Varga, D. Halácsy, P. Kornai, A. Nagy, V. László, N. Trón, V. (2003) Parallel corpora for medium density languages.

Vasiljevs, A. Rirdance, S. Liedskalnins, A. (2008) *EuroTermBank: Towards Greater Interoperability of Dispersed Multilingual Terminology Data.* The First International Conference on Global Interoperability for Language Resources, 213-220.

A Anexo 1: Base de datos terminológicos como archivo TBX

Se adjunta el archivo `terms-eng-spa.tbx`