

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

1.Introducción	4
2.Eurocódigo 7	4
3.Materiales	5
4.Herramientas de trabajo	5
5.Archivos de texto	5
6.Corpus paralelo del Eurocódigo 7	6
7.Glosario de referencia	7
8.Extracción estadística	8
9.Extracción lingüística	8
10.Aprendizaje de patrones lingüísticos	10
11.Extracción lingüística con patrones aprendidos	10
12.Detección en corpus paralelo	11
13.Base de datos terminológicos	12
14.Evaluación y resultados	14
15.Conclusiones	15
16.Bibliografía	15
A Anexo 1: Base de datos terminológicos como archivo TBX	15

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 Ikfreyer@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 Vaguez 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 <u>este</u> 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 <u>parte 1</u>, <u>parte 2</u> y <u>parte 3</u> en formato PDF.

Otro recurso material que se utiliza en este trabajo es el <u>informe</u> 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 <u>Diccionario Técnico</u> <u>Vial de la A.I.P.C.R.</u> 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 <u>Convert</u> <u>PDF To Word .net</u>.

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 <u>Internet Archive</u> (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 <u>PDF</u> <u>Unlocker</u> 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 <u>geotechnicaldesign.info</u> 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.

74 UF	Alignment Editor 1.5 - aligned_Eurocode?-en-Eurocode?-es.bt		- 🗆 🗙
File E	ldt Help		
582	 the softening effect of percolating water on soil or rock strength; 	 el efecto de reblandecimiento producido por el agua de percolación en el suelo y la roca bajo ensayo; 	Eurocode7-en-Eurocode7-es
583	 the softening effect of dynamic actions; 	- el efecto de reblandecimiento producido por las acciones dinámicas;	Eurocode7-en-Eurocode7-es
584	 the brittleness or ductility of the soil and rock tested; 	- la fragilidad o ductilidad del suelo y de la roca bajo ensayo;	Eurocode7-en-Eurocode7-es
585	 the method of installation of the geotechnical structure; 	- el método de instalación de la estructura geotécnica;	Eurocode7-en-Eurocode7-es
586	 the influence of workmanship on artificially placed or improved ground; 	 la influencia de la calidad en la ejecución en terrenos colocados artificialmente o mejorados; 	Eurocode7-en-Eurocode7-es
587	 the effect of construction activities on the properties of the ground. 	 el efecto de las actividades de construcción en las propiedades del suelo. 	Eurocode7-en-Eurocode7-es
588	(5) When establishing values of geotechnical parameters, the following should be considered:	(5) Se deberían considerar los siguientes puntos al establecer valores para los parámetros geotécnicos:	Eurocode7-en-Eurocode7-es
589	 published and well recognised information relevant to the use of each type of test in the appropriate ground conditions; 	 la información publicada y reconocida, relevante para el uso de cada tipo de ensayo en condiciones apropiadas del suelo; 	Eurocode7-en-Eurocode7-es
590	 the value of each geotechnical parameter compared with relevant published data and local and general experience; 	- el valor de cada parámetro geotécnico comparado con la información publicada relevante, y con la experiencia local y general;	Eurocode7-en-Eurocode7-es
<			>
	Herps (F1)	Split (F2) Shift up (F3) Shift down (F4)	

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-Eurocode7es.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 <u>instrucciones</u> 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.

📔 C:\Us	sers\lcaba\TFM\Glosario\IATE\IATE-6831-eng-spa.txt - Notepad++ X
Archivo	Editar Buscar Vista Codificación Lenguaie Configuración Herramientas Macro Eiecutar Plugins Ventana ? X
R 🗐 🛛	8 8 8 8 4 9 6 9 C 8 4 9 9 8 8 5 1 2 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
IATE-6	331-eng-spatht 🛛
5762	steel gabion gavión metálico subjectField: 6831 Note: materials en-termTyp ^
5763	rounded aggregates áridos rodados subjectField: 48, 4816, 6831 Note: mate
5764	broken stone, crushed aggregates, crushed rock áridos machacados subjectFie
5765	fine gravel, grit, pea gravel gravilla subjectField: 48, 4816, 6831 Nc
5766	fine dust, flour harina, polvo fino subjectField: 6831 Note: materials er
5767	aggregate grading, grading, grain-size distribution granulometría subjectFie
5768	grading curve curva granulométrica subjectField: 48, 4816, 6831 Note:
5769	nature of particles naturaleza de los granos subjectField: 48, 4816, 6831
5770	fine grinding molienda fina subjectField: 48, 4816, 6831 Note: material
5771	wet screening tamizado por vía húmeda subjectField: 48, 4816, 6831 Note:
5772	cement fineness finura del cemento subjectField: 48, 4816, 6831 Note: mate
5773	specific surface superficie específica subjectField: 48, 4816, 6831 Nc
5774	OPC, Portland cement, ordinary portland cement, standard Portland cement ce
5775	high early strength cement, rapid hardening cement cemento de alta resistenci
5776	slow-setting cement cemento de fraguado lento subjectField: 48, 4816, 6831
5777	metallurgic cement, metallurgical cement cemento metalúrgico subjectField:
5778	iron Portland cement cemento siderúrgico subjectField: 6831 Note: material
5779	aluminium cement, aluminous cement, fondu, high-alumina cement cemento alumir
5780	pozzolan, pozzolana puzolana subjectField: 6831 Note: building materials
5781 <	free lime cal libre subjectField: 48, 4816, 6831 Note: materials en-ter >
5774 5775 5776 5777 5778 5779 5780 5781	OPC, Fortland sement, ordinary portland cement, standard Portland cement ce high early strength cement, rapid hardening cement cemento de alta resistenci slow-setting cement cemento de fraquedo lento subjectFieldt (84, 4816, 6831 metallurgic cement, metallurgical cement cemento metalúrgico subjectFieldt iron Portland cement cemento siderúrgico subjectFieldt (831 Note: material aluminium cement, aluminous cement, fondu, high-alumina cement cemento alumir pozzolan, pozzolana puzolana subjectFieldt (831 Note: material aluminium cement, aluminous cement, fondu, high-alumina cement cemento alumir pozzolan, pozzolana puzolana subjectFieldt (831 Note: bullding materials free lime cal libre subjectFieldt (84, 4816, 6831 Note: materials enter * emb.)2000m [me:1100]=:5786 (de):180.110

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-modeng.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 candidateseng.txt (véase la Figura 5).

Nombre	Tipo	Tamaño	
pycache	Carpeta de archivos		
🗹 📄 candidates-eng	Documento de texto	21 KB	
🗟 estadistic1	Python File	1 KB	
Eurocode7.en	Archivo EN	732 KB	
Eurocode7.es	Archivo ES	853 KB	
projecte_estadistic.sqlite	Archivo SQLITE	3.320 KB	
stop-eng	Documento de texto	3 KB	
TBXTools	Python File	66 KB	

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
```

Nombre ^	Тіро	Tamaño
pycache	Carpeta de archivos	
📄 etiquetatge	Python File	1 KB
📄 Eurocode7.en	Archivo EN	732 KB
Eurocode7.es	Archivo ES	853 KB
🗹 📄 projecte_linguistic.sqlite	Archivo SQLITE	3.333 KB
🗟 TBXTools	Python File	66 KB

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		
#	ככ	#	NN		
#	NN	#	NN		
#	NN	#	NN	#	NN
#	ככ	#	NN		
#	ככ	#	ככ	#	NN
#	ככ	#	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).

Nombre	Тіро	Tamaño
pycache	Carpeta de archivos	
📑 etiquetatge	Python File	1 KB
Eurocode7.en	Archivo EN	732 KB
Eurocode7.es	Archivo ES	853 KB
🗹 🛃 linguistic-extract	Python File	1 KB
patterns	Documento de texto	1 KB
projecte_linguistic.sqlite	Archivo SQLITE	3.333 KB
📄 TBXTools	Python File	66 KB

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-linguisticeng.txt (véase la Figura 9).

Тіро	Tamaño
Carpeta de archivos	
Documento de texto	19 KB
Python File	1 KB
Archivo EN	732 KB
Archivo ES	853 KB
Python File	1 KB
Documento de texto	1 KB
Archivo SQLITE	7.598 KB
Python File	66 KB
	Tipo Carpeta de archivos Documento de texto Python File Archivo EN Archivo ES Python File Documento de texto Archivo SQLITE Python File

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 linguisticextract.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.

📄 candi	idates-lir	nguistic-eng.txt 🗵	
1	112	limit state	^
2	99	shear strength	
3	84	water content	
4	55	design value	
5	53	civil engineering	
6	52	bearing resistance	
7	51	pore pressure	
8	47	earth pressure	
9	45	partial factor	
10	45	comparable experience	
11	40	laboratory testing	
12	37	ultimate limit	
13	36	ground surface	
14	35	undrained shear	
15	35	undrained shear strength	
16	32	characteristic value	
17	32	overall stability	
18	32	rock mass	
19	32	test specimen	
20	31	load test	¥
.n:1 (Col:1	Sel:0 0 Unix (LF) UTF-8 INS	

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-modeng.txt obtenido en el apartado 7.

Primero se crea un duplicado de la subcarpeta \LTE, a partir de ahora con nombre \LTElearnt-patterns para que no se sobrescriban los datos nuevos de candidates-linguisticeng 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.

Nombre	Tipo	Tamaño
pycache	Carpeta de archivos	
candidates-linguistic-eng	Documento de texto	19 KB
潯 etiquetatge	Python File	1 KB
Eurocode7.en	Archivo EN	732 KB
Eurocode7.es	Archivo ES	853 KB
🗹 📄 glosario-mod-eng	Documento de texto	785 KB
🗹 🝺 learn-patterns	Python File	1 KB
🗟 linguistic-extract	Python File	1 KB
patterns	Documento de texto	1 KB
projecte_linguistic.sqlite	Archivo SQLITE	7.598 KB
🛃 TBXTools	Python File	66 KB

Figura 11: Se guardan learn-patterns.py y glosario-mod-eng.txt en C:\TBXTools\LTElearnt-patterns Una vez se ejecuta el programa learnpatterns.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.

🖼 Símbolo del sistema	-		×
Microsoft Windows [Versión 10.0.17134.112] (c) 2018 Microsoft Corporation. Todos los derechos	reser	vados	^
C:\Users\lcaba>cd C:\TBXTools\LTE-learnt-patterns			
C:\TBXTools\LTE-learnt-patterns>learn-patterns.py # NN # NN # JJ # NN # VBG # NN # NN # NN			
# NN # NN # NN # NN # NNS # NN # IN # DT # NN # TO # VB			
C:\TBXTools\LTE-learnt-patterns>			

Figura 12: Creación de patrón aprendido y extracción de términos

Nombre	Тіро	Tamaño
pycache	Carpeta de archivos	
candidates-linguistic-eng	Documento de texto	19 KB
🝺 etiquetatge	Python File	1 KB
Eurocode7.en	Archivo EN	732 KB
Eurocode7.es	Archivo ES	853 KB
📄 glosario-mod-eng	Documento de texto	785 KB
🝺 learn-patterns	Python File	1 KB
learnt-patterns	Documento de texto	1 KB
🗟 linguistic-extract	Python File	1 KB
patterns	Documento de texto	1 KB
projecte_linguistic.sqlite	Archivo SQLITE	8.791 KB
🛃 TBXTools	Python File	66 KB

Figura 13: Al ejecutar learn-patterns.py se crea el documento con los patrones aprendidos learntpatterns.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\LTElearnt-patterns el programa linguisticextract.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 linguisticextract.py en Símbolo del sistema y se genera un nuevo documento de texto que recoge los candidatos extraídos en candidateslinguistic-eng.txt (véase la Figura 14).

Nombre ^	Тіро	Tamaño
pycache	Carpeta de archivos	
Candidates-linguistic-eng	Documento de texto	36 KB
📑 etiquetatge	Python File	1 KB
Eurocode7.en	Archivo EN	732 KB
Eurocode7.es	Archivo ES	853 KB
glosario-mod-eng	Documento de texto	785 KB
澷 learn-patterns	Python File	1 KB
learnt-patterns	Documento de texto	1 KB
澷 linguistic-extract	Python File	1 KB
patterns	Documento de texto	1 KB
projecte_linguistic.sqlite	Archivo SQLITE	14.575 KB
TBXTools	Python File	66 KB

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-modeng.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 (candidateslinguistic-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).

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



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).

Nombre	Тіро	Tamaño
pycache	Carpeta de archivos	
🗹 📄 candidates-eng-spa	Documento de texto	50 KB
candidates-linguistic-mod-eng	Documento de texto	36 KB
Eurocode7.en	Archivo EN	732 KB
Eurocode7.es	Archivo ES	853 KB
projecte_estadistic.sqlite	Archivo SQLITE	3.020 KB
stop-eng	Documento de texto	3 KB
stop-spa	Documento de texto	664 KB
📄 TBXTools	Python File	68 KB
澷 tond1	Python File	1 KB

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).

CATBXTools\BTE\candidates-eng-spa.tt - Notepad++ - □ × Archivo Editar Buscar Vista Codificación Lenguaje Configuración Herramientas Marco Ejecutar Plugins Ventana ? × Competentia a contante resistencia ^ condidates-eng-spa.tt 1 shear strength resistencia a cortante resistencia ^ civil engineering civil engineering engineerin 3 pore pressure presión intersticial ayuas subt 4 earth pressure unitarios de tierras unitario d 5 shearing resistance rozamiento interno ángulo de 6 ground conditions condiciones del terreno 1) P: 7 undrained shear strength corte sin drenar re 8 rock mass macizo rocoso macizos rocosos:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a tracción P Los: 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales presió 18 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales presió 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v		
Archivo Editar Buscar Vista Codificación Lenguaje Configuración Herramientas Marco Ejecutar Plugins Ventana ? X Configuration Plugins Plugins ? X Configuration Plugins Ventana ? X Configuration Plugins Plugins ? X Configuration Plugins Plugins ? X Configuration Plugins ? X Configurati	📔 C:\T	BXTools\BTE\candidates-eng-spa.txt - Notepad++ - C X
Macro Ejecutar Plugins Ventana ? X Construction of the service servic	Archivo	Editar Buscar Vista Codificación Lenguaje Configuración Herramientas
I shear strength resistencia a cortante resistenci civil engineering civil engineering engineerin pore pressure presión intersticial aguas subt a earth pressure unitarios de tierras unitario d shearing resistance rozamiento interno ángulo de ground conditions condiciones del terreno 1) P: undrained shear strength corte sin drenar re rock mass macizo roccoso macizos roccoss:masas test specimen muestra de ensayo muestras de en load test ensayos de carga carga de pilotes: compressive resistance resistencia a tracción P Los: field tests ensayos de campo ensayos de laborat pile foundation cimentación por pilotes estados lí field tests pressure presiones intersticiales pre- ru utimate limit state 1) P valores de cálculo contact pressure presión de contacto presiones serviceability limit state límite de servicio es corck material material roccoso materiales roccosos v corck material material roccoso materiales roccosos v corce sume corce contacto presiones roccoso v corck material material roccoso materiales roccosos v corck material material roccoso materiales roccoso v corce sume corce	Macro	Ejecutar Plugins Ventana ? X
<pre>candidateseng-spaint I shear strength resistencia a cortante resistenci civil engineering civil engineering engineerin pore pressure presión intersticial aguas subt 4 earth pressure unitarios de tierras unitario d 5 shearing resistance rozamiento interno ángulo de 6 ground conditions condiciones del terreno 1) F: 7 undrained shear strength corte sin drenar re 8 rock mass macizo roccoso macizos roccoso:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 20 rock material material roccoso materiales roccosos v </pre>	🕞 🚽 I	Ê 🖻 🗟 🐻 📥 🕹 🛍 🌔 🗩 😋 📾 🍇 🔍 🔍 🖾 🖘 🛚 👋
shear strength resistencia a cortante resistenci civil engineering civil engineering engineerin pore pressure presión intersticial aguas subt earth pressure unitarios de tierras unitario d shearing resistance rozamiento interno ángulo de ground conditions condiciones del terreno 1) F: undrained shear strength corte sin drenar re rock mass macizo roccoso macizos roccoss:masas test specimen muestra de ensayo muestras de en load test ensayos de carga carga de pilotes:e compressive resistance resistencia a compresión water pressure presiones intersticiales presió tensile resistance resistencia a tracción P Los: field tests ensayos de campo ensayos de laborat pile foundation cimentación por pilotes estados lí pore-water pressure presiones intersticiales pr ultimate limit state 1) P valores de cálculo contact pressure presión de contacto presiones serviceability limit state límite de servicio es corck material material roccoso materiales roccosos v corch material material roccoso materiales roccosos v corch material material roccoso materiales roccoso v corch material roccoso materiales roccoso v corch material material roccoso	📄 candid	ates eng-spa.txt 🗵
<pre>2 civil engineering civil engineering engineerin 3 pore pressure presión intersticial aguas subt 4 earth pressure unitarios de tierras unitario d 5 shearing resistance rozamiento interno ángulo de 6 ground conditions condiciones del terreno 1) P: 7 undrained shear strength corte sin drenar re 8 rock mass macizo rocoso macizos rocosos:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v</pre>	1	shear strength resistencia a cortante resistenci ٨
3 pore pressure presión intersticial aguas subt 4 earth pressure unitarios de tierras unitario d 5 shearing resistance rozamiento interno ángulo de 6 ground conditions condiciones del terreno 1) P: 7 undrained shear strength corte sin drenar re 8 rock mass macizo rocoso macizos rocosos:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes: 11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v 14 contacto presiones intersticiales pr 15 contacto pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v	2	civil engineering civil engineering engineerin
<pre>4 earth pressure unitarios de tierras unitario d shearing resistance rozamiento interno ángulo de 6 ground conditions condiciones del terreno 1) P: 1 undrained shear strength corte sin drenar re 8 rock mass macizo rocoso macizos rocosos:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v 1 tota 6 tento</pre>	3	pore pressure presión intersticial aguas subt
5 shearing resistance rozamiento interno ángulo de ground conditions condiciones del terreno 1) P: 7 undrained shear strength corte sin drenar re 8 rock mass macizo rocoso macizos rocosos:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v x to ta cheta de una producto presiones	4	earth pressure unitarios de tierras unitario d
6 ground conditions condiciones del terreno 1) P: undrained shear strength corte sin drenar re rock mass macizo rocoso macizos rocosos:masas test specimen muestra de ensayo muestras de en load test ensayos de carga carga de pilotes:e compressive resistance resistencia a compresión water pressure presiones intersticiales presió is tensile resistance resistencia a tracción P Los: tifield tests ensayos de campo ensayos de laborat pile foundation cimentación por pilotes estados lí fo pore-water pressure presiones intersticiales pr rultimate limit state 1) P valores de cálculo contact pressure presión de contacto pressiones serviceability limit state límite de servicio es corck material material rocoso materiales rocosos v corche de contacto pressones variantes de cálculo de servicio es corck material material variantes de termina de servicio es corck material material variantes de servicio es corck materiantes de servicio es corck materis de servicio es cor	5	shearing resistance rozamiento interno ángulo de
7 undrained shear strength corte sin drenar re rock mass macizo roccoso macizos roccoss:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material roccoso materiales roccosos v	6	ground conditions condiciones del terreno 1) P:
8 rock mass macizo rocoso macizos rocosos:masas 9 test specimen muestra de ensayo muestras de en 10 load test ensayos de carga carga de pilotes:e 11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v 1 contacto presure presión de contacto presiones	7	undrained shear strength corte sin drenar re
9 test specimen muestra de ensayo muestras de en load test ensayos de carga carga de pilotes:e compressive resistance resistencia a compresión water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v x	8	rock mass macizo rocoso macizos rocosos:masas
10 load test ensayos de carga carga de pilotes:e compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material roccoso materiales roccosos v	9	test specimen muestra de ensayo muestras de en
<pre>11 compressive resistance resistencia a compresión 12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v </pre>	10	load test ensayos de carga carga de pilotes:e
12 water pressure presiones intersticiales presió 13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C	11	compressive resistance resistencia a compresión
13 tensile resistance resistencia a tracción P Los: 14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C	12	water pressure presiones intersticiales presió
14 field tests ensayos de campo ensayos de laborat 15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C > >	13	tensile resistance resistencia a tracción P Los:
15 pile foundation cimentación por pilotes estados lí 16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C > >	14	field tests ensayos de campo 🛛 ensayos de laborat
16 pore-water pressure presiones intersticiales pr 17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C > >	15	pile foundation cimentación por pilotes estados lí
17 ultimate limit state 1) P valores de cálculo 18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C > 1 clta cltato	16	pore-water pressure presiones intersticiales pr
18 contact pressure presión de contacto presiones 19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v C v v v v v v v v v v v v v v v v v v v	17	ultimate limit state 1) P valores de cálculo
19 serviceability limit state límite de servicio es 20 rock material material rocoso materiales rocosos v <	18	contact pressure presión de contacto presiones
20 rock material material rocoso materiales rocoso v <	19	serviceability limit state límite de servicio es
< <u>></u>	20	rock material 🛛 material rocoso materiales rocosos 🗸
	<	>
.n : I Col: I Sel: U U U U IX (LF) U IF-8 INS	Ln:1 Co	ol : 1 Sel : 0 0 Unix (LF) UTF-8 INS

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-modeng.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.

Exportar a un archivo de texto	×
Opciones de campo	Aceptar
Conjunto de <u>c</u> aracteri Unicode (UTF-8)	
Delimitador de <u>c</u> ampı ,	Cancelar
Delimitador de <u>t</u> exto	Ay <u>u</u> da
Entrecomillar todas las celdas de texto	
Guardar el <u>c</u> ontenido de la celda como se ve	
Ancho de columna fijo	

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.

Convert to GlossML					
SV File C:\U	C:\Users\Icaba\TFM\Base_de_datos\terms-A Browse				
GlossML File C:\U	sers\lcaba\TFN	1\Base_de_d	atos\terms-A Brows	se	
CSV Options					
Character Set	UTF-8			\sim	
Column Separator				~	
Text Delimiter	"			~	
Columns					
1					
2			Set Column Typ	be	
			Save Settings		
		Load Settings	Load Settings		
Preview					
1		2		^	
acceptance test		ensayo de aceptación			
angle of incidence		ángulo de	ángulo de incidencia		
angle of shearing resistance		ángulo de rozamiento interno			
artesian water		agua artesiana			
atmospheric pressure		presión atmosférica >			
251 rows 2 columns					
	Pefrech Draview Convertite GlassMI Format				

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).

🛷 Colum	п Туре	×		
 Comment Column Term Column Term Definition Column Ignore Column 				
Language	en English	~		
		Save		

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).

≫.	Anchovy - C:\Users\Icaba\TI	FM\Base_de_d	atos\terms-A-Z-eng-spa.csv.gls	-		×
File	Edit Options Tasks H	lelp				
B	New File				1	2
~	Onen File					•
	opennie					^
	Save File					
	Save File As					
	Export as HTML					
	Export as TMX					
	Export as TBX					
	Export as CSV					
	0 - C:\Users\Icaba\TFM\Ba	se_de_datos\te	erms-A-Z-eng-spa.csv.gls			
	Exit			Alt + F4		
11	bearing stratum		estrato portante		-	
12	block sampling		muestreo por bloque			
13	bond length		longitud de sellado			
14	borehole dilation		dilatación del sondeo			
15	boundary condition		condición de contorno			
16	building site		obra			
17	bulk density		densidad aparente			
18	calibration cylinder		cilindro de calibración			
10	civil engineering		ingeniería civil			
20	classification testing		ensayo de clasificación			
21	clay content		contenido de arcilla			
22	coarse gravel		grava gruesa			
23	coarse soil		suelo grueso			
24	cohesive soil		suelo cohesivo			
25	compaction test		ensayo de compactación			
26	compressive resistance		resistencia a compresión			
27	cone resistance		resistencia en el cono			
28	consistency limit		límite de consistencia			
29	construction sequence		secuencia constructiva			
30	construction site		emplazamiento de la obra			
31	construction work		obra de construcción			
32	contact pressure		presión de contacto			
33	contact surface		superficie de contacto			
34	continuous flight auger pil	e	Pilote con hélice continua			
35	control and calibration uni	t	unidad de control y calibración			~

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).



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).



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