Herramientas de extracción terminológica

Buscar y encontrar la terminología más adecuada para lograr un trabajo de calidad es un objetivo permanente del traductor profesional. La autora de este artículo propone modos para extraer terminología de sitios web o archivos de diferentes formatos para elaborar glosarios o bases terminológicas especializadas.

| Por la **Traductora Pública Valeria Esterzon**, Secretaria de la Comisión de Recursos Tecnológicos

Como todo traductor profesional sabe, cualquiera que sea su área de especialización, no alcanza con tener conocimientos de un tema en particular. El traductor que ejerce como perito no es la excepción y debe buscar todos los medios que estén a su alcance para encontrar la terminología más adecuada y lograr el resultado de mejor calidad. Para ello, se necesita trabajar con fuentes confiables y saber aprovechar al máximo lo que estas nos brindan. A continuación, aprenderemos a extraer terminología de sitios web o archivos de diferentes formatos. Esto nos permitirá elaborar glosarios o bases terminológicas especializadas.

Los extractores terminológicos trabajan mediante técnicas lingüísticas o estadísticas. Las técnicas lingüísticas filtran el texto para reconocer términos. Suelen ser más precisas, pero presentan problemas a la hora de incorporar términos nuevos y, por lo tanto, resulta más difícil mantener actualizadas las listas de términos. Las técnicas estadísticas utilizan algoritmos para medir la frecuencia de aparición de un término, por eso son más precisas para organizar listas de términos específicos. Sin embargo, al utilizar estas técnicas, las listas suelen ser bastante incompletas. Por estas razones, en la mayoría de los casos, las herramientas de extracción automática utilizan técnicas que combinan ambos procesos para obtener mejores resultados. El usuario puede utilizar técnicas que afinen la selección de la herramienta. Puede usar la *stoplist*, que sirve para que el programa se concentre solo en los términos importantes y deje de lado las posibles interferencias, como artículos, preposiciones, etcétera. También puede usar los *word clusters*, que ayudan a resaltar elementos léxicos que aparecen con más frecuencia. Será el usuario quien haga la verificación final para determinar qué términos será útil agregar a una base terminológica y cuáles no.

Las propias herramientas de traducción asistida pueden usarse para extraer terminología, ya sea que el cliente o un colega nos haya enviado una base terminológica o que tengamos glosarios en diferentes formatos de archivo. Este proceso se puede utilizar para actualizar una base preexistente o para generar una nueva.

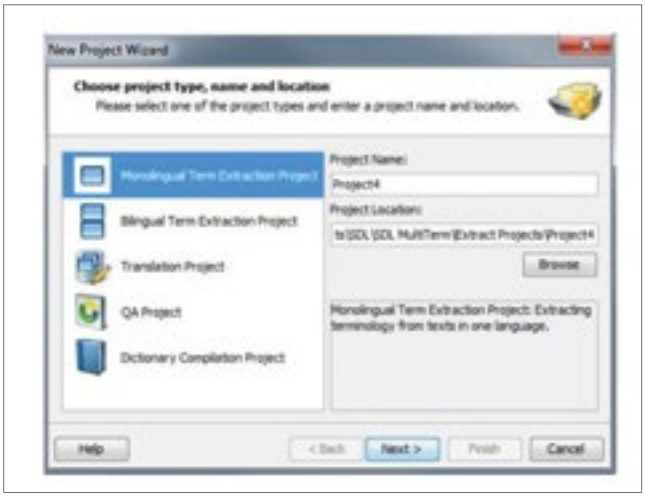
En el caso de memoQ, en la pestaña «Preparación», hay un botón denominado «Extraer términos». Al presionarlo, se accede a una ventana que permite personalizar la extracción. Se puede elegir, por ejemplo, la

Herramientas de extracción terminológica

longitud de caracteres, la frecuencia con la que aparece el término, en qué memoria se quiere incorporar, etcétera. Una vez que se termina la selección, se abre otra ventana con una lista de los términos que se extrajeron. Allí se puede realizar una revisión y decidir si se incorporan todos o si se quiere borrar alguno. Eso permite que el usuario personalice la extracción automática. Para terminar, se exportan todos los términos a la base que se haya elegido o creado.

En el caso de Wordfast, se debe descargar un complemento llamado +Tools, que tiene el mismo propósito. Se puede descargar en el siguiente enlace: https://www.wordfast.net/index.php?whichpage=downloadpage. Es posible optar por la versión demo para probar la herramienta, o bien ingresar con el mismo usuario que se creó para la herramienta TAC.

Por último, si se trabaja con SDL Trados Studio, se puede descargar un *software* pago que trabaja en paralelo. Se llama SDL MultiTerm Extract y es un complemento de SDL MultiTerm, pero se descarga en forma individual en el siguiente enlace: https://www.sdl.com/es/software-and-services/translation-software/terminology-management/sdl-multiterm/extract.html. La extracción se basa en la frecuencia de aparición del término en contexto y tiene la ventaja de poder utilizarse con un gran número de formatos de archivo, como por ejemplo, TXT, PDF, DOCX, entre otros. Cuando comenzamos a trabajar, el programa abre una ventana que nos permite elegir si el proyecto será monolingüe o bilingüe, o, incluso, si queremos hacer la extracción desde diccionarios o glosarios que ya tenemos guardados.

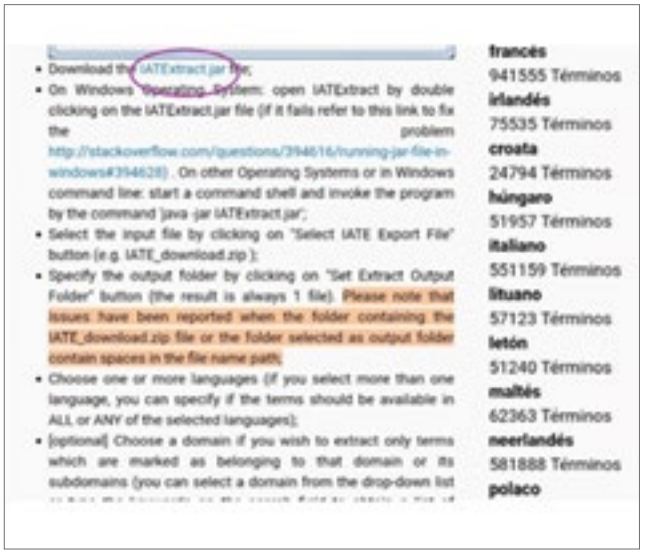


Luego, seleccionamos uno o más idiomas con los que vamos a trabajar y el nombre de la base terminológica. En este paso, podemos elegir combinar una base armada para poder actualizarla o crear una desde cero. Debemos proporcionarle al sistema los patrones que deseamos que utilice, la longitud de caracteres, la cantidad máxima de términos por extraer, etcétera. En esta etapa, también podemos ayudar a la herramienta utilizando las técnicas mencionadas anteriormente, la *stoplist* y los *word clusters*. A continuación, y como ya se mencionó, se debe hacer una selección de los términos que arrojó el programa y determinar cuáles nos interesan para ese caso en específico y cuáles se pueden descartar.

Una fuente de información muy útil para armar bases terminológicas son los corpus. Un corpus es «un conjunto formado por miles de textos (novelas, obras de teatro, guiones de cine, noticias de prensa, ensayos, transcripciones de noticiarios radiofónicos o televisivos, transcripciones de conversaciones, discursos, etc.) y cientos de millones de formas. Son empleados habitualmente para conocer el significado y características de palabras, expresiones y construcciones a partir de los usos reales registrados» (definición de la Real Academia Española). Su ventaja, a diferencia de un diccionario bilingüe, por ejemplo, es que la información está actualizada en tiempo y contexto, lo que nos permite validar el término que buscamos, para el texto que traducimos. A partir de ellos, se pueden armar bases terminológicas monolingües o bilingües.

Uno de los corpus más completos es el de la Unión Europea (IATE), que cuenta con más de siete millones de términos en más de veintiséis idiomas. Además, tiene la ventaja de estar disponible para trabajar en línea o de descargarse de forma gratuita en el siguiente enlace: https://iate.europa.eu/download-iate. El archivo que nos proporciona la IATE está comprimido y se puede descargar para utilizar directamente en una herramienta de traducción asistida.

Si se prefiere, se puede descargar un complemento que permite trabajar con ese archivo para extraer los términos deseados y crear bases terminológicas temáticas. En el mismo enlace anterior encontraremos el complemento, IATExtract.jar.



Al hacer clic, comienza la descarga directa. Una vez que termina, lo ejecutamos y se abre una ventana que nos permite elegir uno o más idiomas y los filtros que queremos aplicar a la selección de términos. Podemos construir cuantas bases terminológicas necesitemos o usar la información para actualizar otra base.

Existen varias herramientas de *software* libre que podemos descargar fácilmente y comenzar a utilizar, sin mayores inconvenientes. Sin embargo, muchas de ellas requieren otros complementos para poder ejecutarlas. Por ejemplo, TBXTools tiene un enlace de descarga directa en la página de su creador: https://sourceforge.net/projects/tbxtools/, pero se debe tener en cuenta que está escrito en lenguaje de programación Phyton y utiliza NLTK (conjunto de programas que permiten el procesamiento del lenguaje simbólico y estadístico de Phyton). Por esta razón, es probable que no podamos leerlo en cualquier computadora, si no corroboramos antes las especificaciones básicas necesarias.

TES (Terminology Extraction Suite) es una herramienta de *software* libre, pero que presentaba varios inconvenientes por este mismo tema. Era necesario instalar algunas herramientas adicionales para poder utilizarlo, y el proceso llevaba bastante tiempo. Ahora, si trabajamos con Windows, se puede descargar directamente del siguiente enlace para comenzar a trabajar: http://lpg.uoc.edu/TES/TES-09.03-win.zip. Para Linux y Mac, debemos tener el intérprete de lenguaje Perl y descargar algunos complementos. Podemos iniciar el proceso de descarga en el siguiente enlace: http://lpg.uoc.edu/TES/TES-09.03.zip.

Los procesos de extracción terminológica llevan un tiempo considerable por las variables que se deben tener en cuenta y porque cada traductor elegirá su propio proceso de trabajo, según el uso que le dé a cada base. En todos los casos, considero que es un tiempo digno de invertir porque cada traductor podrá armar su propia biblioteca terminológica, derivada de fuentes confiables y organizadas según sus criterios.□