Extracción Terminológica Basada en Corpus Para la Traducción de Fichas Técnicas de Impresoras 3D

Ángela Luque Giráldez and Míriam Seghiri[1]

[1] University of Málaga, 29010 Málaga, Spain

Angelaluquegiraldez@gmail.com
Seghiri@uma.es

Abstract. En el presente trabajo presentaremos una metodología para la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) que será de utilidad para la traducción fichas técnicas de impresoras 3D. La extracción de los términos que integrarán el mencionado glosario se realizará a partir de un corpus bilingüe creado a tal efecto al que denominaremos 3DCOR. De este modo, aunamos el formato preferido por los traductores (durante la fase documental), como es el glosario, pero cuya implementación se basará en el recurso ideal para los investigadores, como es el corpus. En cuanto al campo elegido, como hemos apuntado, será técnico, y más concretamente aquel de las impresoras 3D por su novedad y auge en el mercado (cfr. TMT Deloitte, 2019), y, en lo que respecta al género, se ha seleccionado la ficha técnica, pues es uno de los que más demanda genera (cfr. Resolución del Consejo Europeo 98/C 411/01).

Keywords: Corpus paralelo, Extracción terminológica, Glosario, Traducción técnica, Ficha técnica, de mpresoras 3D

1 Introducción

La tecnología avanza a gran velocidad hasta llegar al punto de que ya se comercializan impresoras 3D que pueden ser adquiridas por cualquier usuario común. De hecho, según estudios de TMT Deloitte [1], la industria de la impresión 3D está creciendo a un ritmo aproximado del 12,5 % al año y se prevé que tan solo en 2019 las ventas superen los 2 700 millones de dólares, llegando a los 3 000 millones en 2020. Esta situación, no cabe duda, contribuirá al aumento de las traducciones técnicas en este campo y, en concreto, de sus manuales de instrucciones y fichas técnicas, ya que la *Resolución del Consejo de 17 de diciembre de 1998 sobre las instrucciones de uso de los bienes de consumo técnicos* (98/C 411/01)[1] establece en su quinto artículo lo siguiente:

[1] La presente *Resolución del Consejo de 17 de diciembre de 1998 sobre las instrucciones de uso de los bienes de consumo (98/C 411/01)* puede consultarse en: https://europa.eu/!Hj67Ug.

EUROPHRAS 2019, pages 99–105,
Malaga, Spain, September 25-27, 2019.
https://doi.org/10.26615/978-2-9701095-6-3_013

> Los consumidores deberán poder acceder fácilmente a las instrucciones de uso al menos en su propio idioma oficial de la Comunidad de manera que el usuario pueda leerlas y comprenderlas con facilidad.
> Por razones de claridad y facilidad de uso, cada versión lingüística deberá estar separada de las demás.
> Las traducciones deberán basarse sólo en el idioma original y tener en cuenta las características culturales distintivas de la zona en la que se usa el idioma correspondiente; esto requiere que las traducciones sean hechas por expertos con la formación adecuada, que utilicen el idioma de los consumidores a los que está destinado el producto, y que, en la medida de lo posible, sean sometidas a una prueba de comprensión de los consumidores.

De esta manera, la necesidad de traducciones de calidad de los manuales y fichas técnicas viene dada, además, por imperativo legal. Sin embargo, en el campo que nos ocupa, el de las impresoras 3D, al tratarse de un producto novedoso, es de prever que el traductor se encontrará con una escasez de recursos para abordar su traducción. Y es así como surge el objetivo principal del este trabajo: la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) que será de utilidad para la traducción fichas técnicas de impresoras 3D. La extracción de los términos que integrarán el mencionado glosario se realizará a partir de un corpus bilingüe creado a tal efecto. De este modo, aunamos el formato preferido por los traductores (en particular los noveles) durante la fase documental, como es el glosario, pero cuya implementación se basará en el recurso ideal para los investigadores, como es el corpus (cfr. Seghiri, 2015 [2]).

2 Definición de *Ficha Técnica*

Autores como Gamero (2001) [3] no recogen la denominación de *ficha técnica* dentro de los géneros del campo de la técnica. No obstante, son cada vez más los investigadores, como Byrne (2012) [4], que señalan la ficha técnica como un género independiente con características propias. Así, podemos definir la ficha técnica como un documento que describe las características principales, la composición y las aplicaciones de un producto y que aporta información detallada sobre una serie de aspectos del producto en cuestión. La información suele aparecer presentada en tablas y difícilmente aparecen oraciones completas. Por último, el foco predominante es el foco expositivo, a diferencia de los manuales, donde predomina el foco exhortativo.

3 Creación de un Corpus de Impresoras 3D

Para proceder a la creación del corpus a partir del cual se implementará el glosario es necesario tener claro su diseño para, seguidamente, aplicar un protocolo de compilación y alineación, todo ello siguiendo los postulados expuestos por Seghiri (2006, 2015 y 2017) [5, 2, 6].

3.1 Diseño

Es frecuente encontrar en la red fichas técnicas escritas originariamente en lengua inglesa junto a sus traducciones a diferentes lenguas. Aprovechando este hecho, nos proponemos compilar un corpus *paralelo* de fichas técnicas de impresoras 3D disponibles en Internet; es decir, el corpus estará integrado por bitextos, ya que los documentos que se incluirán serán fichas técnicas originales (escritas originariamente en inglés) y sus traducciones (al español); y, por consiguiente, también será *bilingüe* y *monodireccional*. A su vez, es un corpus *virtual*, pues está integrado por textos descargados exclusivamente de la red, y *textual*, dado que se incorporarán las fichas al completo (y no fragmentos de estas); por último, será *especializado*, y más concretamente *técnico*, en el género de fichas técnicas y la temática de impresoras 3D.

3.2 Protocolo de Compilación y Alineación

Una vez que se ha diseñado el corpus, se procederá en dos fases (cfr. Seghiri, 2006, 2015 y 2017) [5, 2, 6]: la primera fase se dedicará al protocolo de compilación dividida en cinco pasos, a saber, búsqueda, descarga, formato, almacenamiento y determinación de la representatividad cuantitativa; posteriormente, le seguirá una segunda fase consistente en la alineación de los bitextos.

Fase 1: Compilación. Una vez establecido el diseño del corpus, se compilará el corpus, en cinco pasos, con objeto de asegurar la representatividad cualitativa y cuantitativa de la muestra:

Búsqueda: El primer paso consiste en el acceso a la información y la localización de las fichas técnicas que se incluirán en el corpus. Como se trata de un corpus virtual, se descargarán los textos exclusivamente de Internet. Para ello, nos dirigiremos a páginas de empresas de comercialización de impresoras 3D, como HP, Bq, BCN3D o XYZ Printing, por mencionar solo algunas de las más relevantes.

Descarga: El segundo paso consiste en la descarga de las fichas técnicas de las impresoras 3D, que puede llevarse a cabo de forma manual (recurriendo a las teclas Ctrl+G), o bien se puede ir más allá a través del empleo de programas, como BootCat[2], que permiten la descarga de textos en lotes desde una página web determinada mediante el uso de palabras clave.

Formato: El tercer paso consiste en dar el mismo formato a todos los textos para que el programa de gestión de corpus pueda interrogarlos. En este sentido, todas las fichas técnicas que se han descargado se encuentran en formato HTML (.html) o PDF (.pdf), por lo que es necesario convertirlas a ASCII o texto plano (.txt). Para ello, puede sencillamente copiarse el texto y pegarlo en un archivo .txt. Si la ficha en PDF se encontrara encriptada o bloqueada, este proceso se puede llevar a cabo con la ayuda de un programa de reconocimiento de OCR, como Abby FineReader[3].

[2] El programa BootCat puede descargarse en: https://bootcat.dipintra.it/?section=download.

[3] El programa Abbyy FineReader puede descargarse en: https://www.abbyy.com/es-la/bajar/.

Almacenamiento: El cuarto paso consiste en codificar y archivar todos los textos descargados en carpetas y subcarpetas. Para ello, se creó, en primer lugar, una carpeta llamada «Fichas técnicas» que se divide en dos subcarpetas, una para cada lengua de trabajo: para los textos en español, denominada «ES», y para los textos en inglés, llamada «EN». Dentro de estas carpetas destinadas a cada lengua se han creado paralelamente dos más, una llamada «PDF-HTML», en la que se incluirán los documentos en su formato original, y otra llamada «TXT», en la que se almacenarán los textos ya convertidos a texto plano. Una vez estructuradas las carpetas, se organizaron los textos siguiendo una codificación, que permita su organización y explotación en paralelo (así como futuras ampliaciones del corpus, incluso a otras lenguas). De esta forma, la codificación ideada es la siguiente:

— Número: 01, 02, 03, etc.
— Original o traducción: texto original (TO)/texto meta (TM)
— Lengua: español (ES)/inglés (EN)
— Género: fichas técnicas (FT)

Así, el primer texto descargado en lengua inglesa se denominará 01TOENFT, el segundo 02TOENFT y así sucesivamente, mientras que sus traducciones se codificarán como 01TMESFT, 02TMESFT, etc. respectivamente. La codificación también puede llevarse a cabo de forma automática gracias a programas como Lupas Rename[4].

Tras la aplicación de los cuatro primeros pasos hemos asegurado la representatividad cualitativa de la muestra compilada y el resultado ha sido la creación de un corpus *paralelo monodireccional* (compuesto por fichas técnicas redactadas originariamente en inglés y sus traducciones al español), *virtual* (integrado exclusivamente por documentos electrónicos), *bilingüe* (inglés-español) y *textual* (recoge fichas completas), que se encuentra integrado por 110 fichas técnicas, de las cuales 55 son en inglés y 55 en español.

El último paso sería, una vez asegurada la calidad, determinar si se ha alcanzado la representatividad desde el punto de vista de la cantidad a través del empleo del programa ReCor[5]. Esta herramienta fue diseñada por Corpas Pastor y Seghiri (2007) [9], por la que recibieron el Premio de Tecnología de la Traducción de España en 2008, y sirve para determinar el tamaño mínimo de un corpus dado; así, en palabras de Corpas Pastor y Seghiri (2007) [9], para calcular el tamaño mínimo del corpus se establece:

> […] el umbral mínimo de representatividad a partir de un algoritmo (N-Cor) de análisis de la densidad léxica en función del aumento incremental del corpus […]. Se analizan gradualmente todos los archivos que componen el corpus, extrayendo

[4] El programa Lupas Rename puede descargarse en: https://es.ccm.net/download/descargar-456-lupas-rename.

[5] El programa ReCor se encuentra patentado y su licencia de uso gratuita puede solicitarse a través del siguiente correo electrónico: alinares@uma.es.

información sobre la frecuencia de palabras tipo (types) y las ocurrencias o instancias (tokens) de cada archivo del corpus.

Tras subir los documentos al programa, este devuelve unas gráficas con las que se ha podido determinar que el suborpus español es representativo a partir de 47 documentos, con 2 504 *types* y 12 765 *tokens*. De la misma manera, el subcorpus en lengua inglesa es representativo a partir de 48 documentos, 2 387 *types* y 12 056 *tokens*.

El resultado obtenido tras esta primera fase es un corpus representativo tanto desde el punto de vista cuantitativo (gracias a la aplicación de ReCor) como cualitativo (gracias al protocolo de diseño y compilación seguidos), al que denominaremos 3DCOR.

Fase 2: Alineación. Para la extracción terminológica a partir del corpus de bitextos emplearemos, como veremos más adelante, LexTerm[6]. Este programa requiere de una alineación previa de los textos del corpus (cada original con su respectiva traducción), y que supone la segunda fase del proceso. Para alinear los archivos recurriremos a LF Aligner[7] que presenta una interfaz y un funcionamiento sencillos. Este programa detecta automáticamente los segmentos equivalentes y permite una revisión manual en la que se pueden unir o separar los segmentos a través de los botones «Merge», «Split», «Shift up» y «Shift down».

4 Extracción de Unidades Terminológicas

Alineado el corpus de bitextos 3DCOR, utilizaremos el software Lexterm para llevar a cabo la extracción terminológica.

4.1 Identificación de Candidatos a Término

Una vez alineados los textos del corpus, se puede proceder a la identificación de los candidatos a término con el programa LexTerm. Para ello, subiremos los textos al programa y seleccionamos en la barra de herramientas la opción «n-gramas», con objeto de que se active la búsqueda de términos. El programa creará, así, una lista con todos los candidatos a términos encontrados. En la primera columna aparece un cuadrado para seleccionar (o no) los candidatos a término y exportarlos a una futura lista; en la segunda columna se indica el número de veces que aparece el término en cuestión en el texto; en la tercera columna se recoge el término en cuestión, que puede modificarse de forma manual si se necesita; y, en la cuarta y última columna, se albergan las posibles traducciones localizadas. En este punto cabe indicar que, para que el programa extraiga las traducciones de cada término, se debe hacer clic en el botón «Tond» de la barra de herramientas, donde aparecerá una pequeña pestaña que contiene

[6] El programa Lexterm puede descargarse en: http://aulaint.es/software-libre-para-traductores-e-interpretes/herramientas-linguisticas/.

[7] El programa LF Aligner puede descargarse en: https://lf-aligner.soft112.com/.

esta opción. Si no aparece el equivalente adecuado o se duda sobre su validez, se puede ver el término en su contexto pinchando en «Cerca» y, además, siempre cabe la posibilidad de escribir el equivalente de forma manual.

Cuando se ha finalizado la selección de los candidatos a término, se puede guardar el corpus y también se puede exportar una lista con los términos seleccionados en un archivo TXT. En este caso, se guardarán en un archivo TXT todos los términos pertenecientes a los 55 textos que conforman el corpus con objeto de crear, como veremos a continuación, un glosario.

4.2 Creación de un Glosario Bilingüe y Bidireccional

Una vez analizados y extraídos los listados de términos seleccionados, procederemos a la creación del glosario bilingüe y bidireccional para la traducción de fichas técnicas de impresoras 3D.

En primer lugar, se unirán los 55 archivos para manipularlos de forma más sencilla utilizando la herramienta online Files Merge[8] gratuita y de fácil uso.

A continuación, se abre el archivo TXT que aparecerá con los términos separados por una tabulación y se copiarán a un archivo Excel, en dos columnas, la primera en inglés y la segunda en español. Dado que se han unido todos los archivos en un único documento (ahora en Excel), los términos aparecerán desordenados, por lo que habrá que ordenarlos. Para ello, seleccionaremos en la barra de herramientas de Excel la pestaña «Datos», que contiene la opción «Ordenar» en la que se seleccionará el criterio «A a Z». Así, se ordenará alfabéticamente la columna de términos en lengua inglesa (junto a sus respectivos equivalentes). Procederemos a eliminar los términos repetidos y, con ello, quedará listo el glosario inglés-español.

Para obtener el glosario español-inglés basta simplemente con cambiar el orden de las columnas y elegir de nuevo la opción «Ordenar de A a Z». Así, quedarán ordenados alfabéticamente los términos en español junto a sus equivalentes en lengua inglesa, conformando el glosario en su versión español-inglés.

Por último, además de en Excel[9], el glosario bilingüe y bidireccional creado se ha almacenado en formato PDF[10] y en Word[11], con objeto de que el traductor pueda utilizar tres formatos de salida, .xls, .doc y .pdf, en función de sus preferencias.

5 Conclusiones

El presente trabajo ha tenido como objetivo principal la creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) para la traducción de fichas técnicas de impresoras 3D. Este glosario ha sido creado a partir de un corpus diseñado para tal fin, al que hemos denominado 3DCOR. Así, hemos aunado el recurso documental preferido por los traductores, el glosario, con aquel preferido por múltiples

[8] El programa Files Merge puede utilizarse online en: https://www.filesmerge.com/sp/.
[9] El glosario en Excel puede consultarse en: https://bit.ly/2lufSPV.
[10] El glosario en PDF puede consultarse en: https://bit.ly/2lR6GoH.
[11] El glosario en Word puede consultarse en: https://bit.ly/2lwYav7.

investigadores, como es el corpus. El resultado ha sido la creación de un glosario compuesto por 148 términos, y sus respectivos equivalentes, que esperamos que sea de utilidad para realizar traducciones directas o inversas de fichas técnicas de impresoras 3D, así como de otros géneros análogos (como, por ejemplo, los manuales de instrucciones) y temáticas (como aquellas de lápices o escáneres 3D). Asimismo, el corpus 3DCOR podría abrir múltiples líneas de investigación desde el punto de vista monolingüe y monocultural, como desde el punto de vista de la traducción. Por último, la metodología aquí presentada puede ser aplicada para la creación de cualquier otro glosario bilingüe o multilingüe basado en corpus.

6 Agradecimientos

El presente volumen ha sido realizado en el seno de la red temática TRAJUTEC y de la red docente de excelencia TACTRAD (Ref. 719/2018), ambas de la Universidad de Málaga, así como en el marco de los proyectos VIP (Ref. FFI2016-75831-P), NOVATIC (Ref. PIE15-145, UMA), INTERPRETA 2.0 (Ref. PIE17-015, UMA), El traductor autónomo: fiscalidad, impuestos y empleabilidad (Ref. 433/2019, UMA) y PROFETA (Ref. PIE19-033, UMA).

Referencias

1. Technology, media & telecommunications Deloitte, https://www2.deloitte.com/content/dam/Deloitte/ec/Documents/technology-media-telecommunications/Deloitte-ES-TMT-Trends-2019-Folleto.pdf, last accessed 2019/05/06.
2. Seghiri, M.: Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos/Establishing the quantitative representativiness of an E-Reader User's Guide ad hoc corpus (English-Spanish). En: Sánchez Nieto. M. (ed.). Corpus-based Translation and Interpreting Studies: From description to application. pp. 125-146. Frank & Timme, Berlín (2015).
3. Gamero Pérez, S.: La traducción de textos técnicos. Ariel, Barcelona (2001).
4. Byrne, J.: Scientific and Technical Translation Explained. A Nuts and Bolts Guide for Beginners. St. Jerome Publishing, Manchester (2012).
5. Seghiri, M.: Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad. SPICUM, Málaga (2006).
6. Seghiri, M.: Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. En Babel, 63 (1), pp. 43-64 (2017).
7. EAGLES: Preliminary Recommendations on Corpus Typology. En EAGLES (1996), http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html, last accessed 2019/05/10.
8. Seghiri, M.: Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad. SPICUM, Málaga (2006).
9. Corpas Pastor, G., Seghiri, M.: Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor». Procesamiento del Lenguaje Natural 39, 165-172 (2007).