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

Miriam Seghiri
Universidad de Málaga

The sources of information that translators may use are extremely varied, ranging from oral consultation with an expert to a search using specialised dictionaries and glossaries. Nowadays, however, one the most relevant documentation activities in the field of Translation involves the use of Internet resources and, closely related to this, the compilation and management of virtual corpora. For this reason, in the present paper we present a systematic methodology for extracting bilingual and bidirectional glossaries (English-Spanish/Spanish-English) based on parallel corpora to translate TV User Manuals. In fact, according to art. 5 of the Council Resolution of 17 December 1998 on operating instructions for technical consumer goods (98/C 411/01) it is essential to control the quality when writing and translating these manuals. In order to illustrate this methodology we focus on corpus design (according to the skopo) and on the compilation protocol (in four steps: searching, downloading, text formatting and saving data) in order to ensure quality. As for the quantity, we check the quantitative representativeness with the ReCor software (cfr. Seghiri 2006: 387). Once the corpus is representative from the qualitative and the quantitative points of view, it can be managed with a concordance program. So, we illustrate how to extract the terms semiautomatically in order to build a bilingual and bidirectional glossary with a parallel concordance named *ParaConc*. Thus, in the present paper we combine the main resource for researchers (cfr. Bowker 1998; Varantola 2000; Seghiri 2011) within the Translation field: corpora, in order to ensure quality; and the main documentation resource for prospective translators (cfr. Corpas et al. 2001): bilingual glossaries.

Keywords: corpus linguistics, Paraconc, parallel corpora, glossary, representativeness

Babel 63:1 (2017), **43–64**.
DOI 10.1075/babel.63.1.04seg ISSN 0521–9744 E-ISSN 1569–9668

1. Introducción

Las herramientas de las que puede valerse el traductor del siglo XXI son muy variadas, pues van desde la consulta a un profesional, hasta la creación y gestión de corpus virtuales a través de la red Internet; precisamente, son números los autores (cfr. Bowker 1998; Zanettin 1998; Varantola 2000 o Seghiri 2011, entre otros) que apuntan que la utilización de corpus virtuales en Traducción se perfila como uno de los principales recursos documentales a la hora de abordar una traducción especializada pues, con este único recurso, se pueden alcanzar todas las competencias que un proveedor de servicios de traducción debe tener según la norma europea de calidad UNE EN-15038:2006,[1] aprobada por el Comité Europeo de Normalización.

No obstante, tal y como reveló el estudio realizado por Corpas, Leiva y Varela (2001) en torno a los hábitos de los estudiantes de la Licenciatura de Traducción e Interpretación de la Universidad de Málaga, tras indicar su contundente preferencia del diccionario o glosario frente a cualquier otro tipo de recurso, señalaron, además, aquél de corte bilingüe como su herramienta de trabajo por excelencia frente al monolingüe. Los resultados de este trabajo llevado en la Universidad de Málaga coinciden con los presentados en estudios análogos por Atkins y Knowles, en la Universidad de Támpere (Finlandia) o el de Meyer y Roberts, en la Universidad de Ottawa.

Así las cosas, el recurso ideal podría ser aquel que aunara el formato preferido por los futuros traductores, como es el glosario, pero que se basara en el recurso ideal para los investigadores, que permite asegurar la calidad, como es el corpus virtual. De esta forma, en el presente trabajo nos proponemos presentar una metodología propia para la extracción de glosarios basada en corpus paralelos. Esta metodología se ilustrará a través de la creación de un glosario bilingüe y bidireccional (inglés-español) para la traducción de manuales de instrucciones generales de televisores. La elección de este género y de esta temática responde a la gran demanda actual en el mercado profesional de la traducción en España, debido a los avances tecnológicos de dispositivos electrónicos (cf. ACT 2005). Por lo que se refiere al género –manuales de instrucciones generales–, esta demanda viene motivada, además, jurídicamente, puesto que el consumidor europeo tiene derecho a que los manuales de productos técnicos, como son aquellos de los televisores, se encuentren redactados en la lengua oficial de su país en la Unión Europea, tal y como establece la *Resolución del Consejo de 17 de diciembre de 1998*

1. Para más información en torno a la Norma EN 15038:2006, consúltese http://www.aenor.es/aenor/normas/normas/fichanorma.asp?tipo=N&codigo=N0037193#.UtkDvrS6kQM.

sobre las instrucciones de uso de los bienes de consumo técnicos (98/C 411/01)[2] en su artículo quinto. Asimismo, este artículo quinto, no sólo se hace referencia a la *lengua* sino que, además, se subraya la importancia de la *calidad* en la redacción de estos manuales:

> 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.

Por lo que se refiere a la temática elegida, se suma el hecho de que los televisores se sitúan entre los dispositivos electrónicos más utilizados en el hogar, según el Ministerio de Industria, Turismo y Comercio[3] de España, por lo que generan un alto volumen de ventas.

2. Los corpus virtuales

Son numerosos los investigadores –como Bowker (1998 y 1999), Zanettin (1998 y 2002), o Barlow (2002), entre otros– que defienden en sus respectivos trabajos que los corpus virtuales son herramientas que promueven el desarrollo de la competencia traductora, sirven de base para implementar mejoras en la labor docente y ayudan tanto a alumnos como a traductores profesionales a enfrentarse a los desafíos que presenta el mercado laboral. Los corpus virtuales[4] han sido definidos en múltiples ocasiones por diferentes lingüistas que tienen distintas opiniones al respecto, aunque todos ellos coinciden en que están compuestos por textos compilados a través de la red Internet. Así, Aston (1999) considera que un corpus es "compiled

---

2. El texto completo de 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) se encuentra disponible en la siguiente dirección: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31998Y1231(02):ES:HTML.

3. Puede consultarse esta información del Ministerio de Industria, Turismo y Comercio en la siguiente dirección URL:<http://www.lamoncloa.gob.es/ServiciosdePrensa/NotasPrensa/MIN/_2009/ntpr20090225_definicion.htm>

4. En este sentido, utilizamos el término "virtual" siguiendo a Seghiri (2006 y 2011) para referirnos a los corpus electrónicos o también conocidos como corpus ad hoc (Aston, 1999).

'on the fly' by the translator in order to investigate a specific problem encountered during a particular translation", mientras que Corpas Pastor (2002: 201) va más allá y defiende que un corpus virtual no se crea sólo para solucionar un problema concreto de traducción en un encargo específico, sino que el objetivo de éstos es "reunir toda la documentación disponible sobre un tema en muy poco tiempo, ya se trate de documentar un único texto o bien de preparar todo un bloque textual". Un corpus virtual, a su vez, puede ser *comparable*, es decir, aquel corpus que se compone únicamente por textos originales, no traducidos; o bien, *paralelo*,[5] que es aquel que se integra por textos originales y sus respectivas traducciones, o simplemente traducciones. Una de las definiciones más completas de corpus paralelo que podemos encontrar es la aportada por Danielsson y Ridings (1996: 1):

> In its simplest form this refers to a text in one language that has been translated into one or more other languages. All of these taken together are called "parallel texts", since they, ideally, contain the same information in parallel with each other.

En esta línea, Olohan (2004) añade que los corpus paralelos pueden ser, a su vez, *monodireccionales* o *bidireccionales*. Los corpus *monodireccionales* están formados por textos originales en lengua A con sus traducciones correspondientes en lengua B. En cambio, los corpus *bidireccionales* están compuestos por textos originales en lengua A con sus respectivas traducciones en lengua B, al igual que los monodireccionales, pero también incluyen textos originales en lengua B con sus respectivas traducciones en lengua A.

La principal ventaja del uso de corpus virtuales para la traducción, ya sean paralelos o comparables, según Sánchez Trigo (2005: 138), es que:

> constituyen una herramienta interesante para solucionar problemas de diferente naturaleza (temáticos, terminológicos, textuales, estilísticos, etc.) […] ya que permiten compilar una documentación fiable y específica de manera económica (en tiempo y coste) y muy eficaz.

A ello se suma el hecho de que facilitan el trabajo del traductor, permitiéndole alcanzar dos objetivos: de una parte, identificar de manera estadística unidades terminológicas, fraseologías o expresiones habituales en el discurso especializado de un dominio determinado; de otra, puede localizar en los textos segmentos que contienen información conceptual sobre aspectos que puedan resultar relevantes. De este modo, del corpus virtual se puede extraer información de tipo

5. Aunque la denominación de corpus paralelo es la más utilizada, Leech (1991) y otros lingüistas utilizan la denominación de corpus bilingües. Otros investigadores tales como Hallebeek (1999) o Andújar Moreno (2002) aportan una tercera denominación, que es la de corpus de traducción.

tanto lingüístico como cognitivo que permite al traductor especializado alcanzar los conocimientos necesarios para traducir en un dominio determinado.

Sin embargo, como no todo pueden ser ventajas, y los corpus virtuales también presentan algunos inconvenientes para el traductor. Tal y como apunta Seghiri (2006, 230), los corpus que encontramos en la red no son altamente especializados en su mayoría o, en el caso de que encontrásemos corpus especializados, es probable que no cubran todas las necesidades del género textual y temática que deseamos documentar, por lo que al traductor no le queda otra alternativa que compilar su propio corpus virtual. A ello se suma el hecho de que el uso de Internet implica una serie de dificultades pues "[f]inding data on the worldwide is no problem at all. But finding reliable information is rather a difficult task. And finding the information you really need can be very time-consuming and often frustrating" (Austermühl 2001:52). Por consiguiente, a la hora de compilar un corpus será necesario seguir una metodología protocolizada que asegure la calidad tanto desde el punto de vista cualitativo como cuantitativo para que el corpus pueda ser considerado representativo.

3. Compilación de un corpus virtual paralelo

Es necesario contar con una metodología clara de compilación de corpus paralelo para asegurar la representatividad de la muestra. Esta metodología se puede dividir en tres partes bien diferenciadas: en primer lugar, los parámetros de diseño y, en segundo lugar, el protocolo de compilación, en cuatro fases, que permitirán asegurar la calidad del corpus. Finalmente, se procederá a la comprobación de la representatividad cuantitativa de la muestra compilada a través del programa ReCor.

3.1 Parámetros de diseño

En esta sección nos centraremos en el diseño del corpus, que servirá como base para la creación del glosario bilingüe y bidireccional (inglés-español) para la traducción de manuales de instrucciones generales de televisores. Con este objetivo en mente, procederemos a su diseño. De este modo, el corpus será *virtual*, pues lo conformarán textos descargados exclusivamente de la red Internet; *paralelo*, dado que lo integrarán documentos redactados originariamente en inglés y sus correspondientes traducciones al castellano, por lo que también será *monodireccional* y *bilingüe*. Por último, será un corpus *textual*, ya que se compone de manuales de instrucciones generales completos.[6]

---

6. Para más información en torno a las taxonomías de corpus existentes y su clasificación, véase Seghiri (2006).

3.2 Protocolo de compilación

Una vez que se han establecido claramente los parámetros para diseñar el corpus, procedemos a presentar los pasos[7] necesarios para su compilación en cuatro fases, a saber, búsqueda, descarga, formato y almacenamiento.

3.2.1 Búsqueda

La primera fase es aquella que se centra en la localización de los textos que conformarán el corpus. Debemos realizar distintas búsquedas en línea para encontrar los manuales de instrucciones generales tanto en inglés como en español. El tipo de búsqueda más fructífero fue aquella de corte institucional, es decir, en páginas de empresas de reconocido prestigio en la comercialización de televisores como pueden ser Acer,[8] Airis,[9] Panasonic,[10] Philips,[11] Samsung,[12] Sharp[13] o Toshiba,[14] por citar sólo algunas de las más relevantes, así como en bases de datos en red de manuales de instrucciones.[15]

---

7. Para el protocolo de compilación se han seguido las directrices de Seghiri (2006 y 2011) para corpus virtuales comparables y se han adaptado, por primera vez, a los corpus paralelos.

8. Acer es una empresa que se dedica a la venta de productos electrónicos e informáticos (televisores, portátiles, tablets, etc.). Se encuentra disponible en: <http://www.acer.es>.

9. Es una empresa dedicada a la venta de productos informáticos y electrónica de consumo. Se puede consultar su página en: <http://www.airis.es>.

10. Se trata de una empresa que comercializa una amplia gama de productos informáticos y electrónicos (DVDs, teléfonos, electrodomésticos, etc.). Podemos acceder a su página web en la siguiente dirección: <http://www.panasonic.com>.

11. Es una empresa que trabaja con todo tipo de productos electrónicos e informáticos (televisores, afeitadoras, vídeos, etc.). Se puede consultar su página web en: <http://www.philips.es>.

12. Se trata de una empresa que comercializa gran variedad de productos informáticos y electrónicos tales como teléfonos móviles, televisores y ordenadores, entre otros. Su página web se encuentra disponible en la siguiente dirección: <http://www.samsung.com/es>.

13. Esta empresa se dedica a la venta de electrónica de consumo (televisores, portátiles, telefonía móvil, etc.), equipo de oficina e incluso células fotovoltaicas. Podemos consultar su página web en la siguiente dirección: <http://www.sharp.es>.

14. Empresa dedicada a la comercialización de productos electrónicos e informáticos (televisores, portátiles, DVD, Blu-Ray, etc.). Podemos acceder a su página web en la siguiente dirección: <http://www.toshiba.es>.

15. Bases de datos de manuales de instrucciones como la siguiente: <http://www.manual-instrucciones.es/television>.

3.2.2 Descarga

La segunda fase consiste en la descarga de los manuales desde el sitio web que los alberga para guardarlos, posteriormente, en el ordenador. La descarga de los textos puede hacerse de forma manual (descargando uno a uno cada texto con la opción "Ctrl+G"), aunque nosotros la hemos automatizado con la utilización del programa BootCaT.[16] El mencionado programa BootCat permite la descarga en lotes de documentos un sitio web determinado, mediante el empleo de palabras claves (en nuestro caso "manuales de instrucciones" y "televisor").

3.2.3 Formato

Una vez que se han descargado los manuales de instrucciones generales de televisores, se observa una clara predilección por el formato PDF. En este sentido, Sinclair (1991, 21) es muy claro cuando afirma que "the safest policy is to keep the text as it is, unprocessed and clean of any other codes" para que puedan ser procesados por un programa de gestión de corpus. De esta forma hay que proceder al cambio de formato de PDF (.pdf) a texto plano (.txt). Para la conversión, nos valimos del avanzado programa de reconocimiento de OCR llamado Abby Fine Reader.[17]

3.2.4 Almacenamiento

Esta última fase consiste en codificar y archivar convenientemente la totalidad de los documentos que componen el corpus en carpetas y subcarpetas. Para archivarlos, hemos creado una carpeta llamada "Manuales de televisores" que se divide en dos subcarpetas, en función de la lengua. La subcarpeta que contiene los textos en inglés recibe la denominación de "EN", mientras que la subcarpeta dedicada a los textos en español ha sido nombrada como "ES". Cada una de estas carpetas se subdivide, a su vez, según el formato de los documentos: una de ella se llama PDF, que contiene los textos en el formato original con el que se albergan en la red (.pdf), mientras que la otra recibe el nombre de TXT, que contiene los textos ya convertidos a texto plano (.txt).

Asimismo, es esencial que todos los manuales que almacenemos en estas subcarpetas se hayan codificado de forma clara y ordenada para que el programa de gestión de corpus paralelo (*ParaConc*) pueda, seguidamente, alinear cada original

---

16. El programa BootCat puede descargarse en la siguiente dirección URL: <http://bootcat.sslmit.unibo.it>. Sobre las ventajas del uso de BootCat para la compilación de corpus, véase Seghiri, Corpas y Gutiérrez. (2013).

17. Para más información sobre el programa Abby Fine Reader, véase <http://finereaderonline.net>.

con su correspondiente traducción. De esta forma (cfr. Ilustración 1), numeramos cada texto original igual que su traducción (01, 02, 03, etc.); a continuación añadimos, bien TO y EN, para los textos originales (TO) en lengua inglesa (EN) o bien TM y ES para los manuales traducidos (TM, de texto meta) al español (ES). Finalmente añadimos el género, MIG (manuales de instrucciones generales) y la temática, TV (televisor). Es posible automatizar el proceso de codificación de todos los textos que conforman el corpus gracias al programa Lupas Rename.[18] Esta codificación permitirá, asimismo, la futura ampliación del corpus.

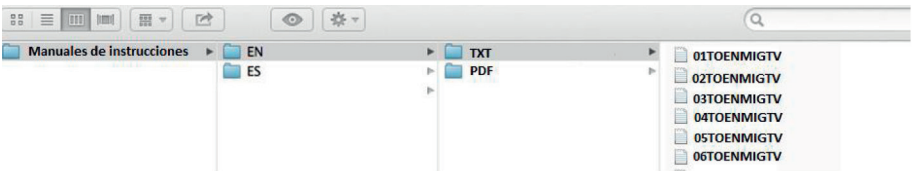

Ilustración 1. Subcorpus en inglés de manuales de instrucciones de televisores en texto plano

El resultado es un corpus paralelo, bilingüe y monodireccional (inglés → español) de manuales generales de televisores integrado por 20 textos en inglés (330.363 tokens o palabras) y de 20 textos en español (y 355.389 tokens o palabras), que es representativo a nivel cualitativo gracias a los parámetros de diseños y protocolo de compilación seguidos. Para concluir el proceso, ahora sólo queda comprobar la representatividad a nivel cuantitativo, cuestión que abordaremos a continuación.

3.3 Determinación de la representatividad cuantitativa

Una vez asegurada la representatividad cualitativa de la muestra a través de los dos pasos anteriores (parámetros de diseño y protocolo de compilación), nos valdremos del programa ReCor[19] para determinar la representatividad cuantitativa de la muestra, es decir, si se ha cubierto la terminología básica empleada en este género (manuales de instrucciones generales) y temática (televisor). Tras analizar cada subcorpus de lengua (inglés y español), el programa genera las gráficas de

---

18. LupasRename puede descargarse desde la siguiente dirección URL: <http://rename.softonic.com/descargar>.

19. El programa ReCor fue diseñado por las doctoras Gloria Corpas y Miriam Seghiri y, actualmente, su algoritmo (N-Cor) se encuentra patentado a través de la Oficina Española de Patentes y Marcas (http://patentados.com/invento/nuevo-metodo-determinar-representatividad-corpus.html). Las inventoras han recibido el Premio de Tecnología de la Traducción de España (2007) en la Universidad Europea de Madrid. Para más información sobre el funcionamiento del programa ReCor, léase Seghiri (2006) y Corpas y Seghiri (2007 a y b).

representatividad (cf. Ilustraciones 2 y 3). De este modo, el subcorpus de manuales de instrucciones generales en español resulta representativo a partir de los 19 documentos y 300.000 tokens o palabras (cf. Ilustración 3).

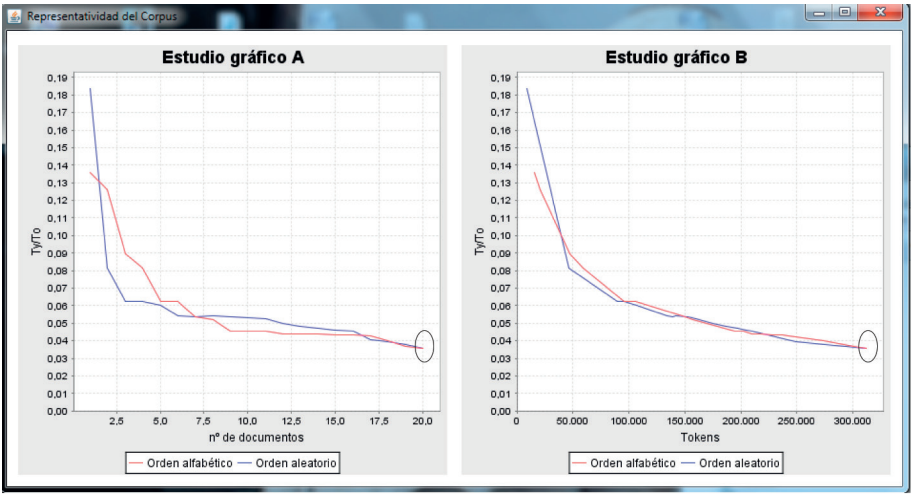

Ilustración 2. Estudios gráficos del subcorpus de manuales de instrucciones generales en español

Por su parte, el subcorpus en inglés alcanza antes la representatividad, a partir de 12 documentos y 200.000 tokens o palabras (cf. Ilustración 3).

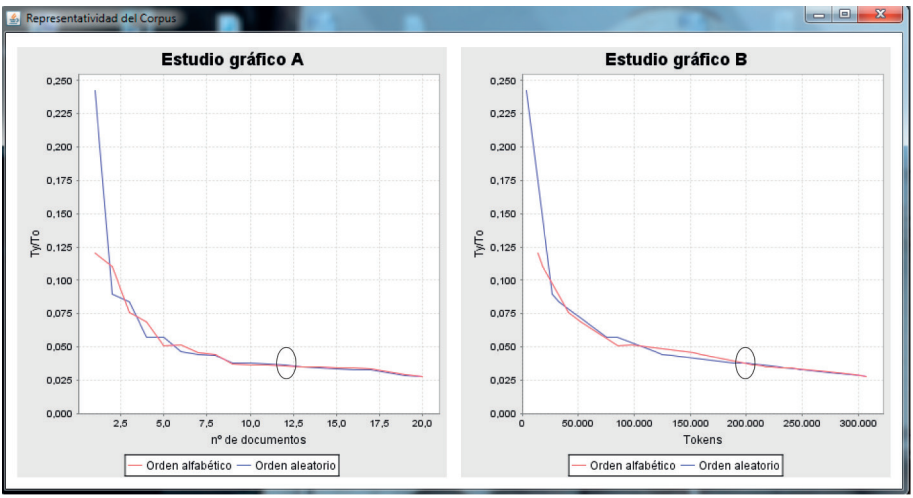

Ilustración 3. Estudios gráficos del subcorpus de manuales de instrucciones generales en inglés

Así, tras asegurar la representatividad cualitativa –a través de los parámetros de diseño y el protocolo de compilación– y cuantitativa –mediante el programa ReCor– de la muestra de manuales de instrucciones generales de televisores, podemos afirmar que nos encontramos ante un corpus representativo, fiable y de calidad, listo para ser utilizado para cualquier tipo de análisis. En nuestro caso, se procederá a la gestión del corpus con el programa de concordancias *ParaConc* para la extracción de un glosario bilingüe y bidireccional (inglés-español/español-inglés) para la traducción de manuales de instrucciones generales de televisores.

4. Elaboración de un glosario bilingüe y bidireccional con ParaConc

ParaConc es uno de los principales programas de concordancias para corpus que existe actualmente en el mercado. Está diseñado especialmente para la gestión y explotación de hasta cuatro corpus paralelos simultáneamente. El programa puede descargarse fácilmente desde la página de Athelstan,[20] la empresa encargada de su distribución. Entre las principales funcionalidades del programa destaca *Text Search*, que permite buscar en el corpus una palabra, o grupo de palabras, bien completa o por truncamiento (a través de *Enter pattern to search for*). En esta opción es muy importante que seleccionemos el idioma (*Language*) de la palabra o grupos de palabras que el usuario desea buscar (véase Ilustración 4):

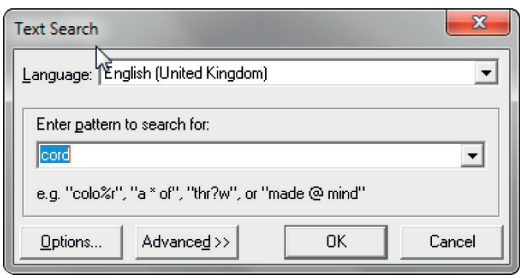

Ilustración 4. Ventana de búsqueda (Text Search) en ParaConc

Al introducir la palabra (en nuestro caso *cord*) o grupo de palabras, *ParaConc* extrae todas las concordancias (cfr. Ilustración 5) que encuentra en la lengua seleccionada (inglés) en ventana superior –el término o términos buscados en su

20. Para consultar el manual completo de ParaConc, se puede acceder a la siguiente dirección: <http://www.athel.com/paraconc.html>.

contexto– y, en la ventana inferior, recoge sus equivalentes. Esta opción recibe el nombre de *Parallel Concordance* o concordancia paralela. En este caso, el programa nos revela todos los resultados correspondientes al término inglés *cord* y su equivalente en español, *cable*. El equivalente puede aparecer marcado en color (azul) para que sea más fácil su localización en contexto gracias a la opción *Search Query* (cfr. Ilustración 5).

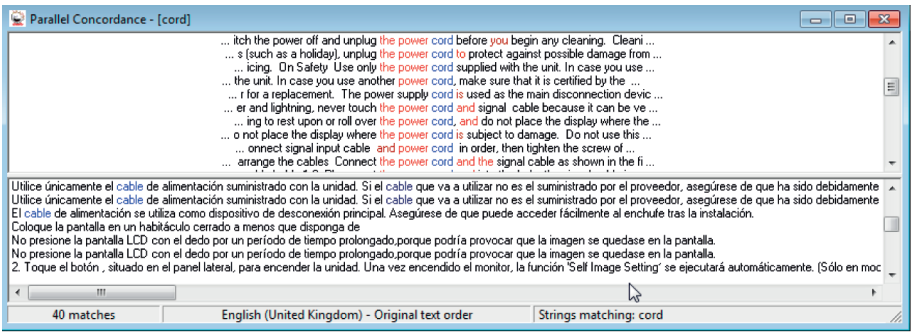

Ilustración 5. Resultados de la búsqueda de "cord"

Otra función interesante que ofrece el programa es la búsqueda avanzada (*Advanced Search*), ya que permite al usuario restringir el número de resultados en la búsqueda (cfr. Ilustración 6).

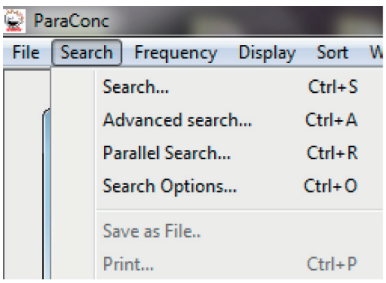

Ilustración 6. Menú de búsqueda avanzada

Destaca, asimismo, la opción *Workspace*, que permite guardar todo el trabajo realizado por el usuario sin necesidad de empezar de nuevo (cfr. Ilustración 7). Cabe destacar que la información sobre los resultados de la búsqueda o las estadísticas de frecuencia no se guardan, aunque sí se almacenan en el historial de búsqueda.

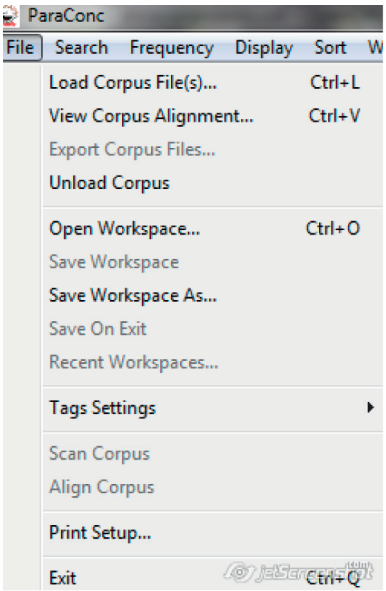

Ilustración 7. Opción Workspace

En último lugar, destacaremos la opción de *Frecuency* (cfr. Ilustración 8) que permite extraer los términos tanto ordenados alfabéticamente (*Alphabetical Order*) como por frecuencia de aparición en el corpus (*Frequency Order*), entre otros.

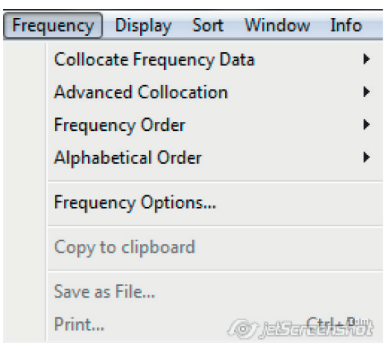

Ilustración 8. Menú de búsqueda por frecuencia

Precisamente, esta última funcionalidad (*Frequency Order*) será la que nos permitirá extraer los términos del corpus que conformarán el glosario, tal y como ilustraremos a continuación.

4.1 Metodología de extracción terminológica para la elaboración del glosario

Para la extracción de los términos que conformarán el glosario bilingüe y bidireccional (inglés-español/español-inglés) de manuales de instrucciones generales

de televisores será necesario gestionar el corpus compilado con el programa *ParaConc*. En primer lugar, debemos seleccionar *File* para desplegar el menú y seleccionar la primera opción: *Load Corpus File(s)* para cargar nuestro corpus en el programa (cfr. Ilustración 9).

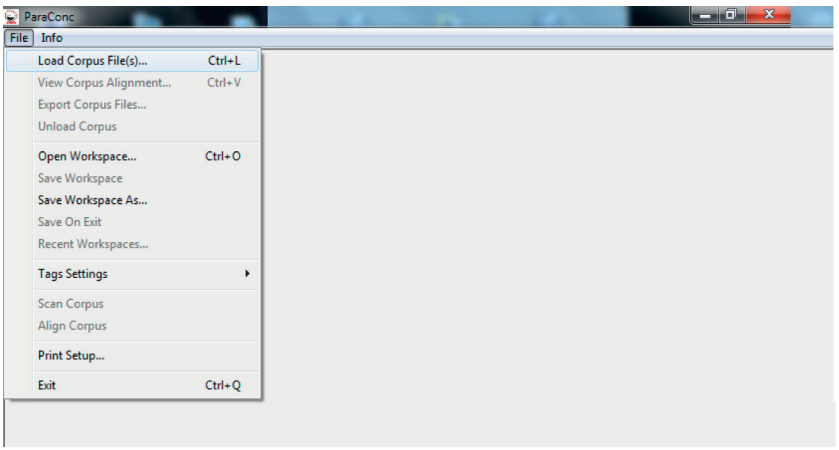

Ilustración 9. Interfaz de ParaConc para cargar los subcorpus

Al seleccionar *Load Corpus File(s)* aparece una nueva ventana donde indicar los idiomas del corpus a saber, inglés y español (cf. Ilustración 10).

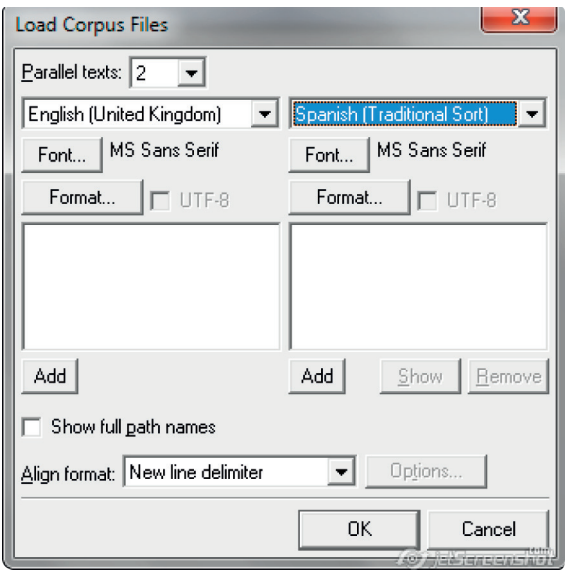

Ilustración 10. Interfaz para subir los subcorpus en los idiomas correspondientes.

Una vez hayamos elegido los idiomas correspondientes, subimos el subcorpus en inglés (en formato .txt) y, para ello, seleccionamos el botón *Add* de la ventana situada a la izquierda (cfr. Ilustración 10). Posteriormente, realizaremos el mismo procedimiento con los textos en lengua española, en la ventana de la derecha (cfr. Ilustración 10). Cuando hayamos subido los textos en lengua inglesa y española, pulsaremos *OK* para que se carguen ambos subcorpus en *ParaConc* (cf. Ilustración 11).

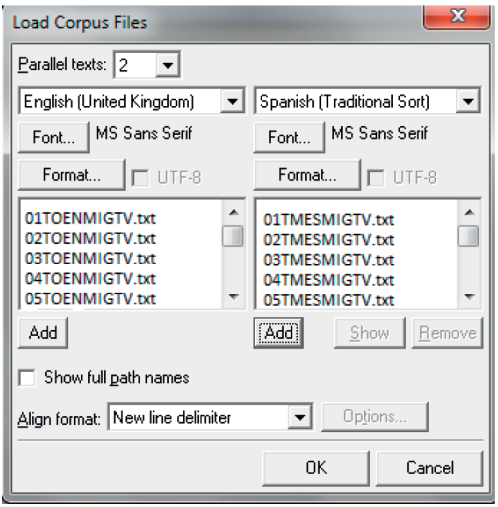

Ilustración 11. Subcorpus (inglés, a la izquierda y en español, a la derecha) subidos a ParaConc

Una vez subidos los dos subcorpus al programa, la opción que nos permitirá crear los glosarios será aquella de frecuencia (*Frequecy*). Extraeremos los términos ordenados por frecuencia –seleccionaremos *Frequency Order*– de ambos subcorpus en paralelo –escogeremos para ello la opción *All*– (cfr. Ilustración 12).

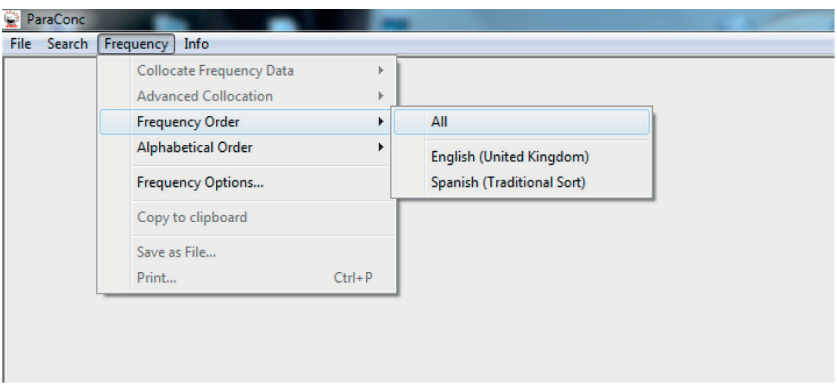

Ilustración 12. Selección de la opción All de Frequency Order en ParaConc para la extracción del glosario

Al tratarse de textos originales (en inglés) y sus traducciones (al español) es muy probable que cada término en inglés y su equivalente en castellano cuente con una frecuencia o número de ocurrencias similar en el corpus, de forma que el programa los muestre prácticamente en paralelo. De este modo, como se puede observar (cfr. Ilustración 13), se genera una lista de frecuencias[21] para el subcorpus inglés (columna de la izquierda) y una lista para el subcorpus español (columna de la derecha).

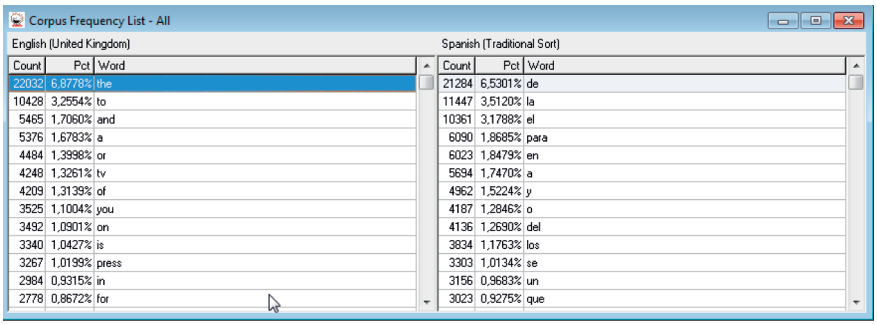

Ilustración 13. Listas de frecuencias del subcorpus inglés (izquierda) y subcorpus español (derecha)

Las listas de frecuencia, normalmente, se encuentran encabezadas por palabras del corpus que están vacías de significado, como por ejemplo, preposiciones o artículos, ya que son las palabras que presentan un mayor número de repeticiones u ocurrencias (cf. Ilustración 13). Si se sigue avanzando en el listado, se puede observar cómo, tras estas palabras vacías de significado, aparece la terminología propia de este género y temática –manuales de instrucciones generales de televisores– (cf. Ilustración 14). Como podemos observar, los términos en inglés y sus equivalentes al castellano cuentan con un número de apariciones similar, mostrándonos prácticamente en paralelo cada término con su equivalente correspondiente (cfr. Ilustración 14). Así, en la columna de la izquierda aparecen términos tales como *screen*, *menu* o *button,* entre otros, y en la columna derecha aparecen sus respectivas equivalencias*: pantalla, menú* o *botón.*

21. Las listas de frecuencias pueden ser de 1 grama (1 término), de 2 o más gramas. En nuestro caso, hemos extraído listas de frecuencias de hasta 10 gramas, con objeto de que nuestro glosario incorpore locuciones, colocaciones y unidades fraseológicas, en su caso.

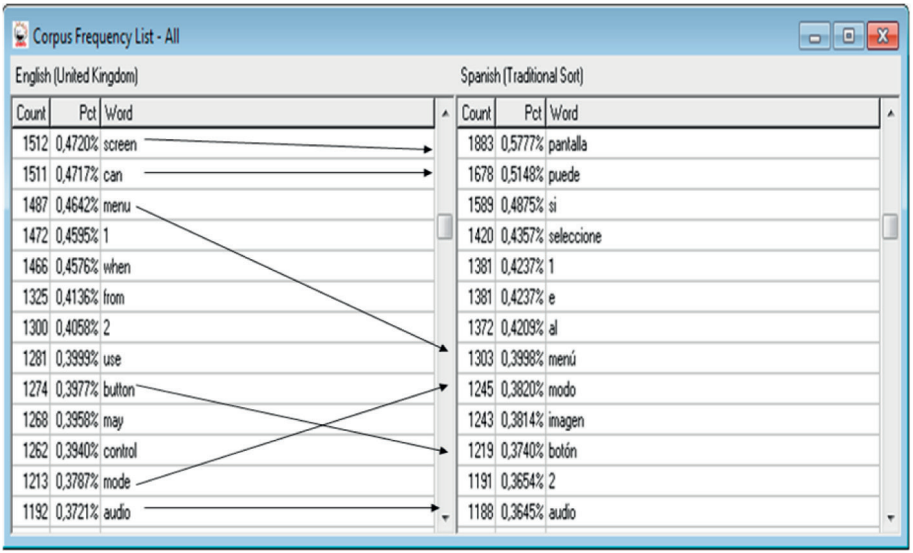

Ilustración 14. Listas de frecuencias del subcorpus inglés y subcorpus español.

Una vez que hemos extraído la lista de frecuencia del subcorpus en inglés y del subcorpus español con *ParaConc*, procederemos a su descarga en formato Excel (a través de la opción *Save as*). No obstante, a pesar de que la terminología aparece casi en paralelo, se siguen intercalando los términos que pueden formar parte del glosario con aquellas palabras vacías de significado (como preposiciones, números, artículos o verbos modales, entre otros). El proceso de eliminación de estas palabras vacías de significado puede automatizarse del siguiente modo:

En primer lugar, descargaremos de la red Internet un listado de palabras vacías de significado (*Stop Words*) en inglés, como *Default English Stopwords List,*[22] *List of English Stop Words,*[23] *Full-Text Stop Words*[24] *o English Stop Words*,[25] y

22. Disponible en la siguiente dirección URL: <http://www.ranks.nl/resources/stopwords.html>.

23. Disponible en la siguiente dirección URL: <http://norm.al/2009/04/14/list-of-english-stop-words>.

24. Disponible en la siguiente dirección URL: <http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html>.

25. Disponible en la siguiente dirección URL: <http://www.textfixer.com/resources/common-english-words.txt>.

en español, como *Spanish StopWords*,[26] *Lista de Stopwords en español*[27] o *Spanish Stopwords in WebpageAnalyse*.[28]

Una vez que tenemos las *Stop Words*, eliminar registros repetidos en Excel es una tarea relativamente sencilla. Para confeccionar el glosario en inglés haremos lo siguiente: abriremos el listado de frecuencia en Excel, con los términos en inglés que hemos descargado de *ParaConc*. Copiaremos las *Stop Words* en el mencionado Excel, en otra zona de la hoja de cálculo, y procederemos a indicarle al documento Excel que queremos eliminar los datos repetidos. Para ello, se deben seguir los siguientes pasos: a) hay que seleccionar en el menú de datos de Excel la opción de "Filtro Avanzado"; b) se abrirá una pequeña ventana con diferentes casillas, en la que debe seleccionarse "Copiar a otro lugar"; c) en la casilla "Rango de la lista" debe introducirse el rango donde están situadas las tablas de términos (también se puede hacer clic sobre las casillas o seleccionar con el ratón); d) en la casilla "Copiar a" hay que seleccionar una celda vacía de la hoja de cálculo, pues a partir de esta celda se copiará el resultado; e) se debe activar la casilla "Sólo registros únicos"; y, f) una vez terminados estos pasos, hay que pinchar el botón "Aceptar" y se observará que los datos se han copiado en la nueva zona indicada pero omitiendo aquellos que están duplicados (gracias a la *Stop List*). Se procederá de igual modo con el listado de términos en español.

Como resultado tendremos dos columnas de Excel (una en inglés y otra en español) en las que se han eliminado las palabras vacías de significado y ha quedado sólo la terminología que formará parte del glosario. Ambos columnas de Excel pueden, ahora, ponerse al lado de la otra en un documento único de Excel (por ejemplo, los términos en inglés en la columna A y los términos en español en la columna B), de forma que se proceda a comprobar que los términos han quedado, efectivamente, ordenados en paralelo o, si fuera necesario, se hagan las modificaciones mínimas necesarias manualmente. Una vez hecho esto, basta seleccionar la columna A (en inglés) y la opción "Ordenar alfabéticamente" de Excel, para que los términos en lengua inglesa aparezcan ordenados de la A a la Z. Una muestra del Glosario para la traducción de manuales de televisores (inglés-español) se recoge a continuación (cfr. Ilustración 15).

---

26. Disponible en la siguiente dirección URL: <http://www.ranks.nl/stopwords/spanish.html>.

27. Disponible en la siguiente dirección URL: <http://es.scribd.com/doc/94217611/Lista-de-stopwords-en-espanol>.

28. Disponible en la siguiente dirección URL: <http://www.webpageanalyse.com/dev/stopwords/es>.

INGLÉS	ESPAÑOL
AC Adapter	Adaptador CA
AC Cord fixing	Fijación del cable de alimentación
Accumulator	Acumulador
Activate	Activar
Adaptation	Adaptación
Add channels	Añadir canales
Adjustments	Ajustes
Advanced playlist creation	Creación de listas de reproducción avanzadas
Antenna	Antena
Alphanumeric	Alfanumérico
Ambilight Technology	Tecnología Ambilight

Ilustración 15. Muestra del glosario para la traducción de manuales de televisores (inglés-español)

ESPAÑOL	INGLÉS
Adaptador CA	AC Adapter
Abrir una aplicación	Launch an application
Accescrio	Attachment
Accesorios incluidos	Supplied accessories
Activar	Activate
Actiialización del software de red	Network software update
Acumulador	Accumulator
Adaptatión	Adaptation
Ajustede inclinatión de imagen	Image tilt correction
Ajustes	Adjustments
Ajustes de imagen	Picture adjustments
Ajustes de zoom	Zoom adjustments
Alfanumérico	Alphanumeric
Almacenamiento de informatión	Data storage
Altavoz de graves	Subwoofer

Ilustración 16. Muestra del glosario para la traducción de manuales de televisores (español-inglés)

Para la extracción del glosario español-inglés, basta con cambiar el orden de las columnas en el archivo Excel (los términos en inglés pasan a la columna B y los términos en español a la columna A). Una vez hecho esto, se ordenarán alfabéticamente los términos de la columna A, en español, para que quede ordenado alfabéticamente de nuevo el glosario (cfr. Ilustración 16).

De este modo ya se encuentra confeccionado el glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus paralelo, listo para ser utilizado para abordar la traducción de manuales de instrucciones generales de televisores.

5. Conclusiones

En el presente trabajo hemos ilustrado una metodología de creación de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basada en corpus paralelo monodireccional (inglés-español) para la traducción de manuales generales de televisores pues, tal y como recoge 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)* en su artículo quinto, es imprescindible asegurar la *calidad* en la redacción de estos manuales. De esta forma, se ha aunado el recurso preferido por los investigadores, los corpus, para asegurar la calidad, con el de los futuros profesionales de la traducción, los glosarios bilingües. Asimismo, la metodología aquí descrita es válida y puede aplicarse para la creación de cualquier glosario, sea cual sea el género, la temática o las lenguas implicadas en el estudio.

Acknowledgments

El presente trabajo se enmarca en el seno del proyecto europeo Marie Curie EXPERT (EXPloiting Empirical appRoaches to Translation, ref. 317471-FP7-PEOPLE-2012-ITN) así como, parcialmente, en los proyectos de I+D INTELITERM (Sistema inteligente de gestión terminológica para traductores, ref. FFI2012-38881), TERMITUR (Diccionario inteligente TERMInológico para el sector TURístico, ref. HUM-2754), VIP (Sistema integrado Voz-texto para IntérPretes, ref. FFI2016-75831), La integración de las nuevas herramientas de traducción automática y posedición en Traducción e Interpretación (ref. PIE 115) y NOVATIC (Integración de nuevas herramientas TIC basadas en corpus en el aula de traducción especializada, ref. PIE15-145).

References

ACT. 2005. *Primer estudio de mercado de los servicios de traducción profesional en España de la Asociación de Empresas de Traducción (ACT)*. Madrid: ACT.

Andújar Moreno, Gemma. 2002. *Construcción de sentido y mecanismos anafóricos: la traducción de las marcas anafóricas TEL y VOILÀ en textos periodísticos*. Tesis doctoral. Barcelona: Universidad Pompeu Fabra.

Aston, Guy. 1999. "Corpus use and learning to translate". *Textus* 12 [en línea]. [http://www.sslmit.unibo.it/~guy/textus.htm].

Austermühl, Frank. 2001. *Electronic Tools for Translators*. Manchester: St. Jerome Publishing.

Bowker, Lynne. 1998. "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study". *Meta* 43 (4): 631–651. doi: 10.7202/002134ar

Bowker, Lynne. 1999. "The Design and Development of a Corpus-based Aid for Assessing Translations". *Teanga* 18: 11–24.

Corpas Pastor, Gloria. 2002. "Traducir con corpus: de la teoría a la práctica". En *Texto, Terminología y Traducción*, ed. por Joaquín García Palacios y M.Teresa Fuentes Morán, 189–226. Salamanca: Almar.

Corpas, Gloria, Jorge J. Leiva y María-José Varela. 2001. "El papel del diccionario en Traducción e Interpretación: análisis de necesidades y encuestas de uso". En *La utilidad de los diccionarios para la enseñanza de las lenguas*, ed. por María C. Ayala Castro, 237–272. Sevilla: Servicio de Publicaciones de la Universidad de Sevilla.

Corpas Pastor, Gloria y Miriam Seghiri. 2007a. "Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness". *Translation Journal* 11 (3) [en línea]. [http://translationjournal.net/journal/41corpus.htm].

Corpas Pastor, Gloria y Miriam Seghiri. 2007b. "Determinación del umbral de representatividad de un corpus mediante el algoritmo N- Cor". *SEPLN: Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural* 39: 165–172.

Danielsson, Pernilla y Daniel Ridings. 1996. *PEDANT: Parallel Texts in Göteborg*. Göteborgs (Suecia): Universidad de Göteborgs.

Hallebeek, Jos. 1999. "El corpus paralelo". *Procesamiento del lenguaje natural* 24: 49–56.

Leech, Geoffrey. 1991. "The state of the art incorpus linguistics". En *English Corpus Linguistics*, ed. por Karin Aijmer y Bengt Altenberg, 8–29. Londres: Longman.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. Nueva York: Routledge.

Sánchez Trigo, Elena. 2005. "Investigación traductológica en la traducción científica y técnica". *TRANS: revista de traductología* 9: 131–150.

Seghiri Domínguez, Miriam. 2006. *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y representatividad* Málaga: Universidad de Málaga.

Seghiri, Miriam. 2011. "Metodología protocolizada de compilación de un corpus de seguros de viaje: aspecto de diseño y representatividad". *Revista de lingüística teórica y aplicada* 49 (2): 13–30. doi: 10.4067/S0718-48832011000200002

Seghiri, Miriam, Gloria Corpas y Rut Gutiérrez. 2013. "Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain". 10th International Conference on Terminology and Artificial Intelligence (TIA 2013). París: Université Paris 13 – Paris Sorbonne Cité [en línea]. [https://lipn.univ-paris13.fr/tia2013/Workshop_Proceedings_files/ workshop.hospitals.tia2013.pdf].

Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Varantola, Krista. 2000. "Translators, dictionaries and text corpora". En *I corpora nella didattica della traduzione*, ed. por Silvia Bernardini y Federico Zanettin, 117–133. Boloña: CLUEB.

Zanettin, Federico. 1998. "Bilingual Corpora and the Training of Translators". *Meta* 43 (4): 616–630. doi: 10.7202/004638ar

Zanettin, Federico. 2002. "DIY Corpora: The WWW and the Translator". En *Training the Language Services Provider for the New Millennium*, ed. por Belinda Maia, Johnattan Haller y Margherita Urlrych, 239–248. Oporto: Facultade de Letras, Universidade do Porto.

Résumé

Les sources d'information pouvant être utilisées par les traducteurs sont extrêmement variées et vont de la consultation orale d'un expert à la recherche à l'aide de dictionnaires et de glossaires spécialisés. Cependant, une des activités documentaires les plus pertinentes de nos jours dans le domaine de la traduction implique l'utilisation des ressources d'Internet et, en étroite relation avec ceci, la compilation et la gestion de corpus virtuels. C'est pourquoi nous présenterons dans cet article une méthodologie systématique visant à extraire des glossaires bilingues et bidirectionnels (anglais-espagnol/espagnol-anglais), basés sur des corpus parallèles, afin de traduire les manuels destinés aux utilisateurs d'un téléviseur. Conformément à l'article 5 de la Résolution du Conseil du 17 décembre 1998 relative au mode d'emploi des biens de consommation techniques (98/C 411/01), il est essentiel de contrôler la qualité lorsque l'on rédige et traduit ces manuels. Pour illustrer cette méthodologie, nous nous concentrerons sur un modèle de corpus (selon le skopos) et sur le protocole de compilation (en quatre étapes : recherche, téléchargement, formatage du texte et sauvegarde des données) pour garantir la qualité. En ce qui concerne la quantité, nous vérifierons la représentativité quantitative avec le logiciel ReCor (cf. Seghiri 2006, 387). Une fois que le corpus est représentatif du point de vue qualitatif et quantitatif, il peut être géré avec un programme de concordance. Par conséquent, nous expliquerons comment extraire les termes de manière semi-automatique pour élaborer un glossaire bilingue et bidirectionnel, avec un programme de concordance parallèle appelé *ParaConc*. Pour garantir la qualité, nous avons donc combiné dans cet article la principale ressource pour les chercheurs (cf. Bowker 1998, Varantola 2000, Seghiri 2011) dans le domaine de la traduction : les corpus ; et la principale ressource documentaire des futurs traducteurs (cf. Corpas et al. 2001) : les glossaires bilingues.

Mots-clés: **linguistique de corpus,** Paraconc, corpus parallèles, glossaire, représentativité

About the author

Miriam Seghiri holds a BA in Translation and Interpreting (Spanish- English, French, Italian) and received her PhD in Translation and Interpreting (with high honours) at the University of Málaga in 2006 (with Ph.D. Best Student Prize). Nowadays she is Senior Lecturer (with tenure) at the Department of Translation and Interpreting at the University of Málaga (Spain). Her research fields range from specialized translation (technical and legal) to corpus linguistics and ICTs (Translation Technologies Research Award in 2007 and María Zambrano Award in 2013). She holds one patent (N-Cor). She has been an invited lecturer in several Masters Programs (University of Cordoba and University of Malaga), and at the internacional (Erasmus Mundus)

Master's Programme on Natural Language Processing (Univerity of Wolverhamptom). She is co-editor in chief of four Journals (REDNMA, REHIPIP, RCDCP and RCHLLPS). The outcome of her research has been made public in national and international academic publications, conferences and invited talks.
Address: Universidad de Málaga, Facultad de Filosofía y Letras, Departamento de Traducción e Interpretación, Campus de Teatinos s/n, 29071 Málaga, Spain
E-mail: seghiri@uma.es