METODOLOGÍA PROTOCOLIZADA DE COMPILACIÓN DE UN CORPUS DE SEGUROS DE VIAJES: ASPECTOS DE DISEÑO Y REPRESENTATIVIDAD*

PROTOCOLIZED METHODOLOGY FOR COMPILATION
OF A TRAVEL INSURANCE CORPUS: DESIGN AND
REPRESENTATIVENESS

MIRIAM SEGHIRI

Universidad de Málaga. Málaga, España

seghiri@uma.es

RESUMEN

A pesar de que el concepto de “representatividad” es clave en la definición de corpus, en ocasiones puede resultar ambiguo pues, hasta la fecha, parece no existir una metodología para asegurar la calidad y el tamaño de la muestra compilada (Giouli y Piperidis, 2002 y CORIS/CODIS, 2011). Para ello, el presente artículo entra de lleno en la lingüística del corpus aplicada a la traducción especializada y en las siguientes páginas se presentará una metodología protocolizada –en cuatro fases– para la compilación de corpus virtuales creados exclusivamente a partir de recursos disponibles a través de la red Internet, que servirá para asegurar la representatividad cualitativa de la muestra. Esta metodología se ejemplificará a través del diseño y compilación de un corpus de seguros de viaje cuya re-

* El presente artículo de investigación se enmarca en el seno del proyecto I+D nacional “Espacio Único de Sistemas de Información Ontológica y Tesauro sobre el Medio Ambiente (ECOSISTEMA)” (Ref. FFI2008-06080-C03-01/FILO. Entidad financiadora: Ministerio de Ciencia e Innovación de España). Asimismo, también se enmarca en el seno de la investigación desarrollada en el proyecto europeo “LEARNINGARABIC: Learning Arabic language for approaching Arab countries to increase business and mutual understanding” (Ref. 143422-LLP-1-2008-1-ES-KA2-KA2MP. Lifelong Learning Programme, Key Activity 2: Multilateral Project). Por último, la metodología aquí descrita para la determinación de la representatividad de los corpus lingüísticos a través del algoritmo N-Cor ha recibido el *Premio de Tecnologías de la Traducción de España* (2007) por parte del Observatorio de Tecnologías de la Traducción de la Universidad Europea de Madrid (Red Temática TIC2002-11705-E, financiada por el Ministerio español de Educación y Ciencia en el área de la Traducción y las Tecnologías de la Información).

presentatividad desde el punto de vista cuantitativo se determinará a través de la aplicación informática ReCor, diseñada para tal fin.

Palabras clave: Lingüística de corpus, compilación de corpus, representatividad, traducción especializada.

ABSTRACT

Although the term “representativeness” is a key concept when defining a corpus, it is one of the most controversial aspects among researchers as regards the ambiguity inherent in the qualitative and quantitative connotations (Giouli & Piperidis, 2002 and CORIS/CODIS, 2011). For this reason, in the following pages, we shall present a systematic methodology—divided in four steps—for corpus compilation based on electronic resources available on the Internet, an essential part of corpus linguistics applied to specialized Translation, and it will guarantee the quality of the corpus. The methodology will be illustrated through the example of the creation of a virtual corpus of travel insurance, whose representativeness from a quantitative point of view will subsequently be determined by using a computer programme specifically designed for this purpose, called *ReCor*.

Keywords: Corpus linguistics, corpus compilation, representativeness, specialised corpora.

Recibido: 03.03.2011. Aceptado: 29.08.2011

1. INTRODUCCIÓN

Hoy en día nadie duda de la importancia y necesidad de la utilización de los corpus virtuales pues se perfila como uno de los más útiles dados los cortos plazos de entrega y la rapidez que se exige en el desempeño de esta profesión. Existe una infinidad de trabajos que realizan estudios a partir de corpus, y en los que se insiste en la calidad y representatividad del conjunto como base fundamental para poder extraer resultados válidos de éste. De hecho, como expuso Biber: “The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research” (Biber, Conrad y Reppen, 1998: 246); sin embargo, y a pesar de su importancia, el concepto de *representatividad*, desde un punto de vista tanto cualitativo como cuantitativo, sigue siendo vago y no parece haber un consenso al respecto: “The definition of representativeness is a crucial point in the creation of a corpus, but it is one of the most controversial aspects among specialists, especially as regards the ambiguity inherent in its use due to the intermingling of quantitative and qualitative connotations” (CORIS/CODIS, 2011).

De este modo, en las siguientes páginas nos proponemos presentar una meto-

dología sistemática para la compilación de corpus virtuales representativos desde un punto de vista tanto cualitativo (a través de un protocolo dividido en cuatro fases) como cuantitativo (con la aplicación informática *ReCor*¹, ideada para tal fin). Esta metodología se ilustrará través del diseño de un corpus virtual de seguros de viajes creado exclusivamente a partir de documentos en formato electrónico disponibles en la red Internet. La elección para la ejemplificación se justifica por la enorme demanda de este tipo de traducciones, tanto directa como inversa, a nivel internacional (cfr. ACT, 2005), pues la industria turística supone uno de los pilares fundamentales de la economía global². A ello se suma el hecho de que el consumidor europeo está en posición de exigir la traducción de este tipo de documentos al amparo de las Directivas europeas en materia asegurativa –*Directiva 92/49/CEE*³ y *Directiva 92/96/CEE*⁴–, y de sus correspondientes transposiciones nacionales, en el caso concreto de España la *Ley 18/1997, de 13 de mayo, de modificaciones del artículo 8 de la Ley de Contrato de Seguro, para garantizar la plena utilización de todas las lenguas oficiales en la redacción de los contratos* (BOE, de 14 mayo de 1997)⁵, la cual reconoce al tomador del seguro el derecho a que el contrato⁶ se redacte, además de en una lengua oficial del territorio del Estado miembro donde se formalice, en otra lengua que él elija. Esta medida, tomada con el objetivo de proteger los derechos del consumidor, ha terminado favoreciendo el diálogo turístico en un contexto bilingüe y, por ende, la traducción de este tipo de documentos.

2. LA LINGÜÍSTICA DE CORPUS EN LA PRÁCTICA TRADUCTORA

Las ventajas de aplicar la lingüística del corpus a la Traducción, y más concreta-

¹ El programa ReCor, diseñado por G. Corpas y M. Seghiri, que permite determinar por primera vez, *a posteriori*, la representatividad de corpus, ha recibido el premio en Tecnología de la Traducción de España (2007) otorgado por el Observatorio de Tecnologías de la Traducción de la Universidad Europea de Madrid.

² El turismo mueve un volumen de negocio astronómico en la economía internacional, donde Europa ocupa una posición privilegiada que la sitúa a la cabeza a escala mundial. Así, autores como Alcaraz, Hughes, Campos, Pina y Alesón (2006) señalan que “el turismo de masas es uno de los fenómenos más novedosos desde la segunda mitad del siglo xx. En España es la industria número uno, la fuente principal generadora de riqueza y de puestos de trabajo”.

³ <http://eur-lex.europa.eu/smartapi/cgi/sga_doc?smartapi!celexplus!prod!DocNumber&lg=es&type_doc= Directive&an_doc=92&nu_doc=49>.

⁴ <<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1992L0096:19950807:ES:PDF>>.

⁵ <http://www.boe.es/aeboe/consultas/bases_datos/doc.php?id=BOE-A-1997-10401>.

⁶ La póliza es el documento en que se plasma el contrato de seguro, así como en el que se establecen las obligaciones y derechos tanto de la aseguradora como del asegurado, donde se describen las personas u objetos que se aseguran y se establecen las garantías e indemnizaciones en caso de siniestro, y que supone, además, la formalización y culminación de todo el proceso de contratación, de ahí que en muchos casos la póliza del seguro sea denominada *contrato* (cfr. Ley 18/1997).

mente del uso y gestión de corpus *virtuales*⁷, han sido señaladas por un gran número de investigadores (Bernardini y Zanettin, 2000; Corpas Pastor, 2001; Corpas Pastor, 2004; Pearson, 1998; Sánchez-Gijón, 2003a; Sánchez-Gijón, 2003b; Seghiri, 2006; Seghiri, 2008, Seghiri, en prensa; Zanettin, 2002a, y Zanettin, 2002b). Muestra de ello es la gran cantidad de proyectos que se han desarrollado hasta la fecha basados en corpus, como por ejemplo aquéllos relativos a programas marco de la Unión Europea, donde se han desarrollado trabajos orientados a la generación textual, la extracción de conocimiento y ontologías basadas en corpus virtuales compilados a partir de sitios web y que combinan específicamente la investigación en el ámbito del turismo con la lingüística de corpus aplicada a la Traducción y con las Tecnologías de la Información y las Comunicaciones (TIC) como *ALMA* (*Arabic linguistic multilingual applications in the context of data exchange on sustainable environment and tourism*), *Development of Italian and German language learning multimedia tools in the tourism sector*, *eMATE* (*Electronic Multilingual Tourism Environment*), *MINTOUR* (*Multimedia Information Network For Tourism*) o *MIS* (*A Multilingual Information Service for Travel Agents*)⁸.

De esta forma, un corpus virtual puede definirse como aquel que se crea exclusivamente a partir de recursos electrónicos disponibles en la red Internet para la realización puntual de un determinado encargo de traducción o documentación de un bloque textual, y cuyo principal objetivo es satisfacer las principales necesidades documentales del traductor de forma económica (sólo requiere de la

⁷ Este tipo de corpus, que hemos denominado *virtual* (cfr. Seghiri, 2006), también recibe otras denominaciones como *ad hoc* (Corpas Pastor, 2001: 164; Sánchez-Gijón, 2003a: 3), *disposable* (Zanettin, 2002a), *do-it-yourself/DIY* (Zanettin, 2002a), *domain-specific* (Corpas Pastor, 2004: 226), por citar sólo algunas de las más frecuentes.

⁸ Otros proyectos europeos, aunque de temática diferente a la turística, basados también en la compilación y gestión de corpus virtuales son, por mencionar sólo algunos de los más relevantes, *AGILE* (*Automatic Generation of Instructions in Languages of Eastern Europe*), *APOLLO* (*An Open Workbench for Multinational Document Creation and Maintenance*), *CRISTAL* (*Conceptual Retrieval of Information using Semantic Dictionary in Three Languages*), *DELIS* (*Descriptive Lexical Specifications and Tools for Corpus-based Lexicon-building*), *DOT.KOM* (*Designing Adaptive Information Extraction from Text for Knowledge Management*), *EUROWORDNET* (*Building a Multilingual Wordnet Database with Semantic Relations between Words*), *GIST* (*Generating Instructional Text*), *INTERLEX* (*Developing General and Terminological Multilingual Databases to be exploited in the Internet from Translation Dictionaries in Electronic Format*), *INTERSECT* (*International Sample of English Contrastive Texts*), *MABLE* (*Multilingual Authoring of Business Letters*), *Access to Language Corpora on Computers*, *MANDES* (*Integrated and Efficient Multilingual Document Management System with Translation and Layout/Editing Capabilities*), *METIS* (*Statistical Machine Translation Using Monolingual Corpora*), *MUSI* (*Multilingual Summarisation Tool for the Internet*), *MULTEXT* (*Multilingual Text Tools and Corpora*), *MULTEXT-EAST* (*Multilingual Text Tools and Corpora for Central and Eastern European Languages*), *METEO* (*Development and provision of multilingual information service*), *NL-TRANSLEX* (*Machine Translation for Dutch and English/French/German*), *QUARTET* (*Quality Assurance Techniques for Enhancing multi-lingual Translation*), *TRANSACCOUNT* (*Translation of Annual Account and Financial Reporting Documents between IAS and FR Accounting System*) o *TT2* (*TransType2 - Computer-Assisted Translation*).

conexión a la red Internet), rápida (los textos pueden descargarse en cuestión de segundos con un *click* de ratón) y fiable (pues permite observar, entre otros, la macro- y microestructura textual así como la terminología propia del campo de especialidad *in vivo*). El corpus podrá ser reutilizado para futuros encargos similares y podrá tener múltiples aplicaciones no sólo para la traducción, sino para la validación, corrección y justificación de encargos así como para establecer teorías de la lengua o crear glosarios y diccionarios, por citar sólo algunos de sus usos más frecuentes.

A pesar de las ventajas del uso de los corpus virtuales en traducción, el principal problema con el que se encuentra el traductor a la hora de hacer frente a un texto es que no existen corpus de tal especialización ya compilados disponibles en red para ser consultados, o en el caso de que existieran, difícilmente cubrirían todas sus necesidades documentales. Ante esta situación, al traductor no le queda otra alternativa que compilar él mismo su propio corpus virtual.

Hay que tener en cuenta, en este sentido, que cualquier colección de textos no constituye por sí misma un corpus. Para que dicha colección textual pueda ser considerada un corpus en el sentido recto del término, debe responder a unos criterios claros de *diseño* y ceñirse a un *protocolo* de compilación específico para que la colección resultante sea *representativa* del campo de especialidad o del tipo de documento concreto que se pretende traducir.

3. PARÁMETROS DE DISEÑO Y FASES DE COMPILACIÓN

En el presente apartado se ilustra un protocolo de compilación de corpus virtuales, consistente en cuatro fases bien diferenciadas, a saber: 1) búsqueda y acceso a la información, 2) descarga, 3) normalización y 4) almacenamiento. Antes de ello, será preciso indicar los principales parámetros de diseño que regirán la compilación del corpus virtual. Esta metodología, válida para la implementación de cualquier corpus de especialidad y en cualquier lengua, se ilustrará a través de la creación, como indicamos con anterioridad, de un corpus de seguros de viaje en español.

3.1. Criterios de diseño

Antes de iniciar el proceso de compilación de un corpus de textos es imprescindible que el traductor establezca unos criterios claros de diseño. En este caso pretendemos crear un corpus de seguros de viaje en español, con vistas a documentar este bloque temático, alimentado exclusivamente a partir de los recursos electrónicos jurídico-turísticos disponibles en la red por lo que, por consiguiente, será un cor-

pues *virtual*. Este corpus *monolingüe*⁹ (español) debe estar delimitado diatópicamente, habida cuenta del dilatado número de países en los que el español se habla y es lengua oficial. Con el objeto de ilustrar la metodología propuesta, el corpus quedará restringido a la legislación vigente (comunitaria y nacional) así como a los elementos formales del contrato (solicitudes de seguro, propuestas, cartas de garantía y pólizas¹⁰), que hayan sido redactados en España. Asimismo, será *comparable*, esto es, integrado por textos originales, no traducidos, para la contratación turística, a la par que *textual*, dado que incluirá documentos completos. A su vez, estaremos frente a un corpus muy *homogéneo*, pues pretende documentar un bloque textual específico, aquel de los seguros de viaje, aunque *desequilibrado* en cuanto a número de documentos pues primará la calidad frente a la cantidad, ya que esta última vendrá determinada, en un principio, por la disponibilidad de los textos en red. No obstante, aunque *desequilibrado*, será un corpus *representativo* a nivel *cualitativo* –pues la calidad viene garantizada por los parámetros de diseño y el protocolo de compilación– así como a nivel *cuantitativo*, ya que será posible determinar si se ha cubierto la terminología básica de este campo de especialidad a través de la aplicación del programa diseñado para tal fin, *ReCor*.

3.2. Protocolo de compilación

Una vez establecidos los parámetros de diseño preliminares, el traductor deberá seguir una metodología protocolizada de compilación del corpus, si quiere asegurar la representatividad cualitativa de la muestra. De esta forma, el protocolo comprenderá cuatro fases, que pasamos a describir a continuación.

3.2.1. Búsqueda y acceso de la documentación

Una vez diseñado el tipo de corpus virtual que se pretende compilar, la primera fase del protocolo consiste en localizar y acceder a la información disponible en Internet. Para ello el traductor necesita desarrollar su competencia heurística en materia de recursos electrónicos. Así, son principalmente dos los tipos de búsqueda que serán de mayor ayuda para el traductor, a saber, la *búsqueda institucional*, que es aquella que se realiza en sitios web específicos de instituciones, asociaciones u organizaciones nacionales e internacionales y la *búsqueda por palabra clave*, a

⁹ Para una mayor información en torno a las distintas clasificaciones de corpus véase Seghiri (2006).

¹⁰ También existen, por un lado, el *duplicado de la póliza*, tramitado a petición del tomador, del asegurado o del beneficiario, siempre por escrito, a la aseguradora, quien tiene la obligación de realizar un duplicado o copia de la póliza en caso de extravío, que será idéntico y tendrá la misma validez que el documento original; por otro, el *boletín de adhesión* –un documento acreditativo del aseguramiento–, que no se ha incluido puesto que atañe exclusivamente a los seguros de vida.

través de buscadores.

De este modo iniciamos nuestra incursión en la red a partir de una *búsqueda institucional*¹¹, que ha resultado una de las más fructíferas para la alimentación del corpus, no sólo por la gran cantidad de documentos que albergan hoy en día este tipo de organismos en la red, sino por la alta calidad y el grado de fiabilidad que se les presupone de antemano, ya que los redactores son especialistas en la materia. Esta búsqueda institucional se realizará principalmente, aunque no de forma exclusiva, en fuentes de búsqueda institucionales, normativas y legislativas. Para la localización de legislación, se pueden emplear sitios y páginas web como los que se indican a continuación:

Por lo que se refiere a organismos e instituciones de carácter oficial, se puede extraer información legislativa de la sede de *Mesa del Turismo*¹², que incluye leyes de ordenación y leyes de disciplina turística sectorial. Ocupa un lugar destacado, asimismo, el de la *OMT (Organización Mundial del Turismo)*¹³, que alberga uno de los principales recursos documentales en materia legislativa, *Lextour*¹⁴, la base de datos de legislación turística, que enlaza a sitios web, a otras bases de datos y a servidores externos de legislación turística, todos ellos emitidos por parlamentos, organismos gubernamentales, universidades o asociaciones profesionales, entre otros. Para la obtención de legislación comunitaria también nos hemos valido de *Eur-Lex*¹⁵, el portal del derecho de la Unión Europea, que constituye, hoy por hoy, la mayor base de datos documental del Derecho de la Unión Europea.

Los documentos que conforman la contratación en materia de seguros de viaje así como su legislación se ubican, prácticamente en su totalidad, en las páginas de las grandes aseguradoras como las internacionales *Mondial Assistance*¹⁶ o *Europ Assistance*¹⁷ o las españolas *Allianz*¹⁸, *MAPFRE*¹⁹ u *Ocaso*²⁰, por citar sólo algunas de las más representativas. A su vez, aunque con menos frecuencia, es posible encontrar estos documentos en la gran cantidad de agencias de viajes que ofrecen sus

¹¹ En gran número de ocasiones, cuando se agotan el nombre de entidades conocidas en esta materia, la búsqueda institucional requiere de otra por palabra clave, por lo general en motores de búsqueda. Normalmente, la introducción de descriptores en *Google* <<http://google.com>> utilizando las técnicas booleanas propias de este buscador como, por ejemplo, *organismo OR turismo, organismo AND turismo* u “*organismo turístico*”, en el caso de que quisiéramos incrementar el número de nombres que integran este tipo de instituciones turísticas, con vistas a visitar sus sedes web para extraer información subjetiva de ser incluida en el corpus de seguros turísticos.

¹² <<http://www.mesadelturismo.com>>.

¹³ <<http://www.world-tourism.org>>.

¹⁴ <<http://www.world-tourism.org/doc/S/lextour.htm>>.

¹⁵ <<http://eur-lex.europa.eu>>.

¹⁶ <<http://www.mondial-assistance.com/en/aboutus/homepage.htm>>.

¹⁷ <<http://www.europassistance.com>>.

¹⁸ <<http://www.allianz.es>>.

¹⁹ <<http://www.mapfre.com/pmapfre/es/index.html>>.

²⁰ <<http://www.ocado.es>>.

servicios en red y publican los documentos contractuales que han contratado con estas aseguradoras, poniéndolos a disposición de sus clientes.

En líneas generales, la búsqueda institucional ha resultado más fructífera para la extracción de legislación, pero no de documentos contractuales. De este modo, se necesita dar un paso más, y llevar a cabo una *búsqueda por palabra clave*. Para ello, es posible utilizar un buscador genérico, como por ejemplo *Google*, que según un gran número de analistas es el mejor por calidad en el resultado de las búsquedas (Radev, Fan, Qi, Wu y Grewal, 2005: 580).

De este modo, la localización de documentos *por palabras claves* constituye la forma más sencilla y rápida para rescatar los documentos que integran la contratación de seguros de viaje. En este sentido, los motores de búsqueda dan mejores resultados si se saben utilizar los recursos que ofrecen —operadores booleanos, truncamientos, búsquedas por frase, entre otros— así como cuanto más definida y precisa sea nuestra búsqueda. Por consiguiente, el establecimiento de descriptores es esencial. A continuación ilustramos (Tabla I²¹), a modo de ejemplo, cómo se realizaron las búsquedas para la localización de los textos que conforman el corpus. Para ello se toman como descriptores los tipos textuales y el ámbito de los seguros en el que se desea localizar la información —en este caso, *seguro turístico* o *asistencia en viaje*²², términos con los que se denomina jurídicamente al seguro de viajes (cf. Auriolés Martín, 2005)—, y se aplican técnicas de búsqueda booleanas (OR, AND, NOT) a partir de sencillas ecuaciones como las que se ejemplifican a continuación:

²¹ En esta tabla hemos presentado los descriptores que nos han permitido extraer mayor número de documentos para el tipo textual que se rastreaba en español, aunque, en este sentido, conviene señalar que existen y se han manejado infinitos criterios de búsqueda y que aquí sólo pretendemos recoger una muestra a modo ilustrativo.

²² El término *seguro turístico* cuenta con una larga tradición en España desde la aprobación, en 1964, por parte de la Presidencia de Gobierno, del *Decreto 3404/1964*, de 22 de octubre, por el que se establece el “seguro turístico”. Sin embargo, la *Orden Ministerial de 27 de enero de 1988 por la que se califica la cobertura de las prestaciones de asistencia en viaje como operación de seguro privado* transpuso a nuestro ordenamiento el texto de la *Directiva 84/641/CEE* del Consejo, de 10 de diciembre de 1984, por la que se modifica, en lo que se refiere en particular a la asistencia turística, la *Directiva 73/239*. La orden ministerial transpuesta empleaba el término “travel assistance”, que fue traducido al español por el calco neológico oficialmente aceptado de “asistencia en viaje”. No obstante, preferimos utilizar *seguro turístico* (o *seguros de viaje*, tal y como reza en el título de este trabajo), pues coincidimos con la postura adoptada por Auriolés Martín (2005) y Auriolés Martín, Benavides y González (2004) al entender que es más acertado que el calco español del original inglés *asistencia en viaje*, dado que la asistencia en viaje es tan sólo una de las posibles coberturas que ofrece el seguro turístico o seguro de viajes, como pueden ser otras del tipo anulación del viaje o asistencia médica, por citar algunas de las más frecuentes en esta contratación (cfr. Seghiri 2008 [2007]). Para una visión más amplia al respecto, véase la clasificación trilingüe (español-inglés-italiano) de pólizas de seguro turístico en función de su cobertura propuesta por Seghiri (2006: 279-281).

Tabla I: Descriptores para la localización de elementos formales del contrato de seguro de viajes (español).

Tipo textual	Descriptores	Ecuación de búsqueda
Póliza	Póliza, seguro turístico, asistencia en viaje	"seguro turístico" OR "asistencia en viaje" AND póliza
Solicitud	Solicitud de póliza, seguro turístico, asistencia en viaje	solicitud AND póliza AND "seguro turístico" Solicitud AND póliza AND "asistencia en viaje"
Propuesta	Propuesta, proposición, seguro turístico, asistencia en viaje	póliza AND propuesta OR proposición "seguro turístico" póliza AND propuesta OR proposición "asistencia en viajes"
Carta de Garantía	Carta de garantía, seguro turístico, asistencia en viaje	"carta de garantía" AND "asistencia en viaje" "carta de garantía" AND "seguro turístico"

El principal inconveniente de las palabras clave es la elección de los descriptores más precisos para la búsqueda que se pretende realizar, puesto que, de lo contrario, la información no pertinente tendrá una presencia elevada. Corresponde al traductor, en última instancia, filtrar todo ese “ruido documental” en cada una de las páginas susceptibles de ser incluidas en el corpus.

3.2.2. Descarga de datos

A la fase de búsqueda y acceso de los documentos le sigue una *descarga de datos*. Normalmente, esta fase se realiza de forma manual aunque es posible automatizar la tarea cuando se trata de un grupo de páginas a través de programas que permiten la descarga “en lotes”, como *GNU Wget*²³ o *GetBot*²⁴. Esta segunda fase de descarga también se puede ver dificultada por la estructuración propia de la red ya que nos encontramos ante un lenguaje de marcas o HTML, es decir, la información se organiza en nodos hipertextuales a los que, en muchas ocasiones, no es fácil acceder, normalmente debido a una denominación poco apropiada para el contenido que alberga así como a una ubicación visualmente dificultosa en la página. Por otro lado, mencionaremos la infinidad de formatos en los que puede aparecer la información, que da paso a la tercera fase de compilación denominada de *normalización*.

²³ Este *software* libre, listo para descarga y acompañado de su manual de instrucciones se encuentra en la siguiente dirección URL: <<http://www.gnu.org/software/wget/>>.

²⁴ GetBot es un programa gratuito que se encuentra disponible para su descarga en la siguiente dirección URL: <<http://descargar.mp3.es/lv/group/view/kl42345/GetBot.htm>>.

3.2.3. Normalización

Es posible observar que en Internet, tanto para la legislación como para los documentos contractuales relativos a los seguros de viaje, existe una predilección por los formatos HTML (.html) y PDF (.pdf). El primero no supone grandes problemas de conversión, puesto que puede copiarse la información y ser pegada en un documento de texto sin mayores dificultades. Por lo que se refiere a los documentos en .pdf, el buscador *Google* permite, en la mayoría de las ocasiones, visualizarlos como .html a través de la opción “Versión en HTML”, por lo que se puede proceder del mismo modo. Cuando esto no es posible, los programas de conversión como *PDFtoWord*²⁵ o *ABBYY FineReader* pueden resultar de gran ayuda. Se cierra así esta tercera fase que sigue a la descarga y, ahora, todos los documentos presentarán un formato ASCII o de texto plano (.txt), es decir, sin marcas HTML o código de otro tipo, siguiendo, en este punto, la *clean-text policy* expuesta por Sinclair (1991: 21): “The safest policy is to keep the text as it is, unprocessed and clean of other code”.

3.2.4. Almacenamiento

La última fase consiste en el *almacenamiento* de datos. De este modo, los documentos descargados se guardan, correctamente identificados y relacionados. Una forma posible de hacerlo es mediante una carpeta principal que indique la temática del corpus y sucesivas divisiones en subcarpetas en función de la lengua y del formato (original y ASCII), de forma que permita una perfecta identificación para futuros usos así como posibles ampliaciones a otras lenguas o tipos textuales, por ejemplo (Figura 1).

Figura 1: Almacenamiento del corpus de seguros de viaje.

²⁵ Puede descargarse la versión limitada de PDFtoWord en la dirección URL: <<http://pdf-to-word.softonic.com>>.

Una vez que se han llevado a cabo los cuatro pasos –búsqueda y acceso de la documentación, descarga de datos, normalización y almacenamiento– el resultado ha sido la creación de un corpus virtual monolingüe (español) integrado por 159 textos²⁶ (1.837.869 palabras). Ahora que ya se ha creado una muestra representativa desde el punto de vista cualitativo, sólo queda verificar que se ha alcanzado el tamaño mínimo de palabras y documentos para que la colección sea considerada también representativa desde el punto de vista cuantitativo en este campo de especialidad, los seguros de viaje.

4. ESTABLECIMIENTO DEL UMBRAL DE REPRESENTATIVIDAD

Hasta la fecha, mucho se ha escrito e investigado en torno a la cantidad como criterio representativo y sobre las posibles fórmulas capaces de estimar un mínimo de palabras y documentos a partir del cual un corpus especializado puede considerarse representativo sin llegar a resultados concluyentes. De hecho, tal y como apuntan Giouli y Piperidis (2002): “There is no general agreement as to what the size of a corpus should ideally be. In practice, however, the size of a corpus tends to reflect the ease or difficulty of acquiring the material”.

De este modo, los intentos de fijar un tamaño, al menos mínimo, para los corpus especializados han sido varios. Algunos de los más significativos son los expuestos por Heaps (1978), Sánchez Pérez y Cantos Gómez (1997) o Young-Mi (1995). Según Yang, Cantos y Song (2000: 21), tales propuestas presentan importantes deficiencias porque “the critical problem is, however, how to determine the value of tolerance error for positive predictions”. La determinación del tamaño mínimo de un corpus sigue siendo uno de los aspectos más controvertidos en la actualidad (Seghiri, 2006). Las cifras tan dispares que se han manejado hasta la fecha, así como la poca fiabilidad que dan las propuestas para su cálculo, nos llevaron a reflexionar sobre una posible solución, que se ha visto materializada en la aplicación informática denominada *ReCor*²⁷, que pasamos a describir a continuación. Con este método se pretende plantear una solución rápida y eficaz

²⁶ Por lo que se refiere a los documentos legislativos que integran el corpus (16 textos en español), hemos de decir que los seguros turísticos no cuentan con legislación sustantiva, pero se ven afectados por las regulaciones de los seguros distintos del de vida mediante diversas directivas comunitarias como: 73/239/CEE, 73/240/CEE, 76/580/CEE, 78/473/CEE, 84/641/CEE, 87/343/CEE, 87/344/CEE, 88/357/CEE, 90/618/CEE, 92/49/CEE, 95/26/CEE, 2000/26/CE, 2000/64/CE y 2002/13/CE. Además, en España, los contratos de seguros turísticos se rigen en la actualidad por la *Ley 50/1980, de 8 de octubre, de Contrato de Seguro*, así como por la *Ley 30/1995, de 8 de noviembre, de ordenación y supervisión de los Seguros Privados*. Por lo que se refiere a las pólizas, documento en torno al cual gira este tipo de contratación, se han localizado 101 documentos (1.000.067 palabras). El resto del corpus lo integran el resto de elementos formales del contrato.

²⁷ Para una descripción más detallada de *ReCor*, véase Seghiri (2006).

para determinar, por primera vez, *a posteriori*, el tamaño mínimo de un corpus independientemente de la lengua o tipo textual de dicha colección, estableciendo, por lo tanto, el umbral mínimo de representatividad mediante el análisis de la densidad léxica en función del aumento incremental del corpus.

4.1. Especificaciones del programa

ReCor es una aplicación informática creada con objeto de poder estimar la representatividad de los corpus en función de su tamaño y que se caracteriza, ante todo, por la sencillez de su interfaz de usuario (Figura 2), frente a la carga eminentemente matemática y de formulación que abundan en este tipo de trabajos.

Figura 2: Interfaz de ReCor (versión 2.1).

Hasta el momento se han implementado tres versiones del programa *ReCor*: 1.0, 2.0 y 2.1. El funcionamiento es básicamente similar y corresponde a la descripción genérica que ofrecemos a continuación. Ahora bien, la versión 2.0 difiere de la versión 1.0 en que permite a) seleccionar automáticamente un directorio completo de documentos (en vez de tener que pulsar la tecla *Shift* como en la versión anterior) y b) permite seleccionar un número de n-gramas para el cálculo, donde $n \geq 1$ y $n \leq 10$. Ambas versiones (1.0 y 2.0) generan archivos estadísticos en texto plano (.txt). La versión 2.1. se diferencia de su predecesora en que presenta los archivos estadísticos simultáneamente en formato .txt y en forma de tablas en Excel.

4.2. Funcionamiento del programa

En este apartado mostraremos el programa *ReCor* en funcionamiento (versión 2.1.). Para la ilustración del manejo del programa utilizaremos el corpus de seguros de viaje en español compilado. Gracias a una sencilla interfaz, *ReCor* resulta de fácil utilización y así, en primer lugar, procedemos a la selección de los archivos que conforman el corpus mediante el botón “Selección de los ficheros del corpus”. Una vez seleccionados los archivos, podremos incorporar, si se desea, un “filtro de palabras”. En nuestro caso hemos incluido un filtro que contiene numeración romana. Además, el programa genera tres ficheros de salida (*Análisis estadístico*, *Palabras ord. alf.* y *Palabras ord. frec.*) que se crearán por defecto en la ubicación que determine la aplicación. Si se desea otra localización de los archivos de salida generados, puede indicarse una nueva ruta. El primero, “Análisis estadístico”, recoge los resultados de dos análisis distintos: de un lado, los ficheros ordenados alfabéticamente por nombre; de otro, para los ficheros ordenados en orden aleatorio. El documento aparecerá estructurado en cinco columnas, a saber, muestra de *types*, *tokens*, cociente entre palabras distintas y totales (*types/tokens*), número de palabras con una aparición (V1) y número de palabras con dos apariciones (V2). El segundo, “Palabras ord. alfa.”, generará dos columnas en las que aparecerán las palabras ordenadas por orden alfabético, de una parte, y sus correspondientes ocurrencias, de otra. En tercer lugar, “Palabras ord. frec.”, presenta la misma información que el fichero de salida anterior, pero esta vez las palabras se ordenan en función de su frecuencia, es decir, por rango.

Por último, procederemos a especificar “Grupo de palabras”, esto es, los *n*-gramas ($n \geq 1$ y $n \leq 10$). Escogemos, para una primera ejemplificación, “1” (Figura 2). Asimismo, indicaremos “sí” en la opción “Filtrar números”.

4.2.1. Representaciones gráficas

Una vez que se han seguido los pasos descritos más arriba, la aplicación está lista para realizar el análisis, cuyo resultado se expresa en forma de representaciones gráficas, de una parte, y ficheros de salida en .txt con datos estadísticos exportables a tablas en Excel, ya descritos (cfr. 4.2), de otra. Para generar las representaciones gráficas A y B pulsamos “Aceptar”. Éstas nos permitirán determinar si, efectivamente, nuestra colección es representativa desde el punto de vista cuantitativo, es decir, si se ha cubierto la terminología básica existente en este campo de especialidad (Figura 3). De esta forma, *ReCor* permite establecer *a posteriori*, esto es, una vez creado el corpus o durante el proceso de compilación, sin tener que establecer valores prefijados, el umbral de representatividad de un corpus bien construido, es decir, compilado conforme a criterios de diseño cualitativos. Concretamente, se parte de la idea de que el cociente entre las palabras reales de un texto y las

totales (*types/tokens*), que da cuenta de la densidad o riqueza léxica de un texto, no aumenta proporcionalmente a partir de un número de textos determinado (representación gráfica A). Lo mismo ocurre cuando la representatividad se calcula en función de la densidad léxica a partir de secuencias de palabras o n-gramas (representación gráfica B). Sobre esta base teórica, las representaciones gráficas A y B permiten ilustrar gráficamente el punto a partir del cual un corpus que ha sido compilado según criterios cualitativos comienza a ser representativo en términos cuantitativos. Las representaciones gráficas, a partir de dos líneas que reflejan, de un lado, los documentos incluidos alfabéticamente y, de otro, aleatoriamente, que se unen y se estabilizan a medida que se aproximan al valor cero, muestran el tamaño mínimo de la colección para ser considerada representativa.

El tiempo que invierta el programa en generar las representaciones gráficas y los archivos de análisis dependerá del número de n-gramas seleccionados para el cálculo, del tamaño del corpus analizado y de la versión utilizada.

Figura 3: Representatividad del corpus de seguros turísticos (1-grama).

A partir de los datos arrojados por *ReCor* podemos concluir que el corpus español de seguros de viaje (Figura 3) es representativo a partir de 120 documentos y de 740.000 palabras.

Si deseamos ver los resultados para dos o más gramas, esto es, verificar que el corpus también es representativo desde el punto de vista colocacional, locucional y fraseológico, repetiremos los pasos anteriormente expuestos y especificaremos la cifra en “Grupo de palabras”. A continuación mostramos los resultados arrojados por *ReCor* para 2-gramas (Figura 4).

Figura 4: Representatividad del corpus de seguros turísticos (2-gramas).

De este modo, a partir de los datos que nos ofrece el programa para 2-gramas, se desprende que el corpus español de seguros de viaje (Figura 4) es representativo a partir de 150 documentos y un millón de palabras.

5. CONCLUSIONES

En el presente artículo nos hemos ocupado del corpus virtual como recurso documental por excelencia para la práctica de la traducción; sin embargo, la metodología para su compilación a veces no resulta demasiado clara. Así, para que dicha colección pueda ser denominada *corpus* tendrá que cumplir con unos parámetros de diseño claros, seguidos de un protocolo de compilación específico dividido en cuatro fases bien diferenciadas –1) búsqueda y acceso a la información, 2) descarga, 3) normalización y 4) almacenamiento– para poder asegurar la representatividad cualitativa de la muestra. Por lo que se refiere a la representatividad desde el punto de vista cuantitativo, una de las características principales de los corpus virtuales hasta la fecha consistía en que solían ser eminentemente desequilibrados, puesto que su tamaño y composición finales venían determinados, normalmente, por la disponibilidad. En este sentido, es imprescindible contar con herramientas que nos aseguren su representatividad de la muestra desde el punto de vista terminológico, colocacional, locucional y fraseológico. Sin embargo, las propuestas realizadas hasta la fecha para el cálculo de la representatividad no resultaban fiables, pues intentaban establecer la representatividad cuantitativa *a priori*. Por este mo-

tivo se ha implementado la aplicación informática de sencillo manejo *ReCor*, que permite determinar *a posteriori*, por primera vez, la representatividad cuantitativa de los corpus. Hasta el momento esta metodología se ha probado con éxito para un macrocorpus de turismo en alemán, español, francés, inglés e italiano (Seghiri, 2006), corpus especializados de seguros turísticos y contratos de viaje combinado en alemán, español, inglés e italiano (Corpas Pastor y Seghiri Domínguez, 2007; Seghiri, 2006; Seghiri, 2008 y Seghiri, en prensa), en corpus de turismo de salud y belleza en español, inglés, italiano y francés (Castillo Rodríguez, 2010) así como en corpus de resúmenes de artículos científicos en inglés y español (Toledo Báez, 2009).

Un corpus comparable, virtual y representativo desde el punto de vista cualitativo y cuantitativo resulta de enorme utilidad a la hora de estudiar el discurso propio del campo de especialidad en cuestión, en este caso el de seguros de viaje, sus manifestaciones correspondientes, así como sus realizaciones y formas textuales, todo ello desde una perspectiva monolingüe y monocultural, así como desde el punto de vista de la traducción, la comparación y mediación interlingüística e intercultural. Asimismo, el corpus virtual compilado se revela tremendamente útil también para la docencia de la traducción especializada, pues ayuda a los procesos autónomos de enseñanza-aprendizaje, estableciendo los mecanismos adecuados para la especialización y diversificación del traductor, además de favorecer el estudio de textos traducidos por alumnos con objeto de corregir y validar encargos de traducción, además de otros posibles usos aún por descubrir.

REFERENCIAS

- ACT. 2005. *Primer estudio de mercado de los servicios de traducción profesional en España de la Asociación de Empresas de Traducción (ACT)*. Madrid: ACT.
- Alcaraz, Enrique, Brian Hughes, Miguel Ángel Campos Pardillo, Víctor Manuel Pina y María Amparo Alesón Carbonell. 2006. *Diccionario de términos de turismo y ocio. Inglés-Español /Spanish-English*. Editorial Ariel: Barcelona.
- Aurioles Martín, Adolfo. 2005. *Introducción al Derecho Turístico (Derecho Privado del Turismo)*, 2ª ed. Madrid Tecnos.
- Aurioles Martín, Adolfo, Patricia Guillermina Benavides Velasco y María Belén González Fernández. 2004. *Contratación Turística*. Documento Técnico. [BFF2003-04616 MCYT/TI-DT-2004-1].
- Bernardini, Silvia y Federico Zanettin (eds.). 2000. *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate*. Bolonia: CLUEB.
- Biber, Douglas; Susan Conrad y Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Castillo Rodríguez, Cristina M. 2010. *El trabajo terminográfico en traducción: ex-*

- plotación de un corpus multilingüe de turismo de salud y belleza (español-francés/inglés/italiano)*. Tesis doctoral. Málaga: Universidad de Málaga.
- CORIS/CODIS. 2011. "Design and implementation of a CORpus di Italiano Scritto" CORIS/CODIS. Bolonia: CILTA. [En línea]. Disponible en http://dslo.unibo.it/coris_engDesign.html. [Consulta: 28/08/2011].
- Corpas Pastor, Gloria. 2001. "Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada". *TRANS. Revista de Traductología* 5, pp. 155-184.
- Corpas Pastor, Gloria. 2004. "Localización de recursos y compilación de corpus vía Internet: Aplicaciones para la didáctica de la traducción médica especializada". En Consuelo Gonzalo García y Valentín García Yebra (eds.). *Manual de documentación y terminología para la traducción especializada*. Madrid: Arco/Libros, pp. 223-257.
- Corpas Pastor, Gloria y Miriam Seghiri Domínguez. 2007. "Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness". *Translation Journal* 11 (3). [En línea]. Disponible en <http://accurapid.com/journal/41corpus.htm>. [Consulta: 28/08/2011].
- Giouli, Voula y Stelios Piperidis. 2002. *Corpora and HLT. Current trends in corpus processing and annotation*. Bulgaria: Institute for Language and Speech Processing. [En línea]. Disponible en http://www.larflast.bas.bg/balric/eng_files/corpora1.php. [Consulta: 28/08/2011].
- Heaps, H. S. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Nueva York: Academic Press.
- Pearson, Jennifer. 1998. *Terms in Context, Studies in Corpus Linguistics*. Amsterdam y Filadelfia: John Benjamins Publishing.
- Radev, Dragomir, Weiguo Fan, Hong Qi, Harris Wu y Amardeep Grewal. 2005. Probabilistic question answering on the web. *Journal of the American Society for Information Science and Technology (JASIST)* 56 (6), pp. 571-583.
- Sánchez-Gijón, Pilar. 2003a. "Es la web pública la nova biblioteca del traductor?". *Tradumàtica: Traducció i tecnologies de la informació i la comunicació* 2, [En línea]. Disponible en <http://www.bib.uab.es/pub/tradumatica/15787559n2a7.pdf>. [Consulta: 28/08/2011].
- Sánchez-Gijón, Pilar. 2003b. *Els documents digitals especialitzats: utilització de la lingüística de corpus com a front de recursos per a la traducció*. Tesis doctoral. Barcelona: Universidad Autónoma de Barcelona.
- Sánchez Pérez, Aquilino y Pascual Cantos Gómez. 1997. "Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish". *International Journal of Corpus Linguistics* 2 (2), pp. 259-280.
- Seghiri, Miriam. 2006. *Compilación de un corpus trilingüe de seguros turísticos (español-inglés-italiano): aspectos de evaluación, catalogación, diseño y represen-*

- tatividad*. Málaga, España: Servicio de Publicaciones de la Universidad de Málaga. [En línea]. Disponible en <http://www.biblioteca.uma.es/bbldoc/tesisuma/16754888.pdf>. [Consulta: 29/01/2011].
- Seghiri, Miriam. 2008 [2007]. “La traducción al español del término *travel insurance*”. *PUNTOYCOMA, Boletín de los traductores españoles de las instituciones de la Unión Europea*. Bruselas: Comisión Europea. [En línea]. Disponible en http://ec.europa.eu/translation/bulletins/puntoycoma/107/pyc1071_es.htm. [Consulta: 28/08/2011].
- Seghiri, Miriam. 2008. “Creating virtual corpora step by step”. *Researching and Teaching specialized languages: New contexts, new challenges. VII Annual Conference of the European Association of Languages for Specific Purposes (AELFE)*. Murcia: Universidad de Murcia, pp. 435-449.
- Seghiri, Miriam. En prensa. “El corpus comparable para la didáctica de la traducción jurídica inversa (español-inglés)”. *IV Congreso de la Asociación Ibérica de Estudios de Traducción e Interpretación: “Traducir en la frontera”*. Vigo: Universidad de Vigo.
- Sinclair, John M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Toledo Báez, Ma Cristina. 2009. *El resumen automático como recurso documental para la traducción de artículos de investigación del ámbito jurídico-tecnológico (español-inglés-francés)*. Tesis doctoral. Málaga, España: Universidad de Málaga.
- Yang, Dan-Hee; Pascual Cantos Gómez y Mansuk Song. 2000. “An Algorithm for Predicting the Relationship between Lemmas and Corpus Size”. *ETRI Journal* 22 (2), pp. 20-31.
- Young-Mi Jeong. 1995. “Statistical Characteristics of Korean Vocabulary and its Application”. *Lexicographic Study* 5 (6), pp. 134-163.
- Zanettin, Federico. 2002a. “DIY Corpora: The WWW and the Translator”. En Belinda Maia; Jonathan Haller y Margherita Urlrych (eds.). *Training the Language Services Provider for the New Millennium*. Oporto: Faculdade de Letras, Universidade do Porto.
- Zanettin, Federico. 2002b. “CEXI. Designing an English Italian Translational Corpus”. En Bernhard Ketteman y Georg Marko (eds). *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, pp. 329-343.