María del Mar Sánchez Ramos*

Corpus paralelos y traducción especializada: ejemplificación de diseño, compilación y alineación de un corpus paralelo bilingüe (inglés-español) para la traducción jurídica

https://doi.org/10.1515/les-2019-0015

Abstract: The article reports on the processing steps followed to build a bilingual parallel corpus (English-Spanish) as a resource for legal translation. The corpus, being in its initial stage of development, is made up of 127 and 145 aligned judgments referring to English and Spanish courts respectively. The corpus was aligned by using InterText alignment software, accounting for a total number of 29983 aligned sentence pairs. The paper describes the different design stages and the technical issues related to the compilation process.

Palabras clave: lingüística de corpus, corpus paralelo, bitextos, traducción jurídica

Keywords: corpus linguistics, parallel corpus, bitexts, legal translation

1 Introducción

La lingüística de corpus ocupa ya un lugar destacado dentro de los estudios de traducción (Baker 1993; Blum-Kulka 2004). Considerados como recursos indiscutibles de documentación en la traducción especializada (Sánchez Ramos 2017) o como recursos en la enseñanza de la traducción en general (Fantinuoli y Zanettin 2015), los distintos tipos de corpus son fuente de inspiración dentro de nuestra disciplina. En lo referente a la especialización de la investigación, como es la traducción jurídica, las propias características de la misma pueden plantear la necesidad de la compilación y diseño de un corpus que posibilite el acceso a un material textual ad hoc que, igualmente, aporte luz sobre cuestiones traductológicas concretas. El avance ha sido lento, pero firme, y son ya varios los trabajos que han contribuido a asentar la metodología de corpus dentro del campo de la traducción jurídica. Prueba de ello, y como queda recogido en Vigier Moreno y

***Kontaktperson: María del Mar Sánchez Ramos,** University of Alcalá (Madrid), Modern Philology, C/ Trinidad 3, 28801 Alcalá de Henares, Spain, E-Mail: mar.sanchezr@uah.es

Sánchez Ramos (2017), la metodología de corpus se emplea para la investigación de la fraseología jurídica (Pontrandolfo 2015), el uso de binomios en el discurso jurídico (Andrades Moreno 2016), o para fines pedagógicos (Monzó Nebot 2008), por nombrar algunos ejemplos.

El artículo que aquí se presenta ofrece una descripción del diseño de un corpus paralelo bilingüe (inglés-español) formado por sentencias dictadas por el Tribunal de Justicia de la Unión Europea (TJEU) correspondientes a diversos órganos jurisdiccionales. La principal finalidad es proporcionar a la comunidad académica una metodología de investigación basada en corpus en los contextos especializados y que pueda ser provechosa y aplicable en diversos ámbitos de investigación, como pueden ser los estudios de fraseología jurídica, de técnicas de traducción empleadas, el uso de patrones léxicos y colocacionales, así como para el entrenamiento de sistemas de traducción automática estadísticos, uso de memorias de traducción en herramientas de traducción asistida o como uso para consulta y extracción de terminología (Oliver 2017).

Tal y como nos advierten diversos estudiosos (Prieto 2011; Soriano Barbino 2018), la formación del traductor jurídico es clave para garantizar el éxito de la traducción de los textos de naturaleza jurídica. La dificultad de estos textos no solo radica en su componente lingüístico, caracterizado por una terminología propia, la confluencia de tendencias sintácticas y estilísticas y una tipología de géneros textuales variada, sino también en cuestiones de tipo pragmático o cultural. En esta línea, la labor del traductor jurídico engloba la familiarización con el ordenamiento jurídico y el dominio del lenguaje de especialización del derecho, la identificación del género jurídico y el dominio de las técnicas de documentación (Borja Albi 2000). Dentro de las técnicas de documentación, y tomando como punto de partida el trabajo reciente de Soriano Barbino (2018: 225), hay que destacar la necesidad que todo traductor jurídico debe desarrollar, como es la competencia profesional, interpersonal e instrumental, que incluye "el uso de fuentes documentales especializadas, búsquedas terminológicas, gestión de la información, uso de herramientas informáticas [...]". No hay duda de que los corpus textuales y los programas de gestión de corpus quedan incluidos en las palabras de Soriano Barbino (2018), pues son herramientas informáticas de gestión terminológica o fuentes de documentación especializada.

El uso de los corpus textuales resulta interesante para el estudio (o enseñanza) de aquellos aspectos que puedan ser fuente de dificultades, como son los conceptos culturalmente marcados. Traducir textos jurídicos supone, además del dominio lingüístico, temático y textual, asimilar el funcionamiento de un sistema estructurado e integrado por los llamados *system-bound terms* (Jopek-Bosiacka 2013), además de conocer aquellos vinculados a un ordenamiento jurídico determinado. Como bien señala Prieto (2011: 2016), es la falta de correspondencia

conceptual y la ausencia de equivalentes plenos lo que pone obstáculos al traductor jurídico y limita "el margen de maniobra para la aplicación de procedimientos que serían adecuados en otros contextos de traducción documental". Con todo, la asimetría jurídica derivada de los distintos ordenamientos jurídicos en los que puede darse la traducción jurídica (Soriano Barbino 2018) y el anisomorfismo cultural es lo que lleva a una situación donde la terminología ocupa un lugar primordial a la hora de hacer frente a los textos de carácter jurídico (Biel 2014a). Con todo, son cada vez más los trabajos que ahondan en dicha problemática (Sánchez Ramos y Vigier Moreno 2016). Para hacer frente a ello, desde este trabajo se propone una metodología basada en corpus, que sirvió para compilar un corpus paralelo (inglés-español) formado por sentencias dictadas por el TJEU, garante de la interpretación de la legislación de la Unión Europea (UE) en los distintos estados miembros, y que ponían fin a procesos iniciados en distintos ordenamientos jurídicos dentro del ámbito inglés y español.

En aras de obtener un corpus especializado de calidad que sea adecuado para la investigación, se exige un proceso de compilación o protocolo, tarea cuanto menos ardua y tediosa, ya que el investigador puede verse en la tesitura de adquirir una serie de destrezas técnicas y metodológicas que le aseguren la creación de dicho corpus. En nuestro caso, dicho protocolo quedó sustentado en tres fases: documentación, compilación y alineación.

El trabajo queda articulado de la siguiente forma: tras la introducción (Sección 1), la Sección 2 ofrece una contextualización teórica breve sobre corpus y traducción, y sus aplicaciones; la Sección 3 describe el proceso de documentación del corpus paralelo bilingüe (inglés-español); el proceso de compilación, que incluye la codificación, organización y alineación del corpus; la Sección 4 esboza las principales conclusiones del trabajo, que, a su vez, propone líneas de investigación futuras.

2 Metodología de corpus y traducción

Baker (1993) supuso un punto de partida en la aplicación de la lingüística de corpus en los estudios de traducción, donde la estudiosa habla de "a turning point in the history of the discipline" (Baker 1993: 235). A partir de esa fecha, muchos han sido los trabajos que han apostado por el uso de una metodología de corpus en el campo de la traducción y la interpretación (Monzó Nebot 2008; Sánchez Ramos 2017, Seghiri 2017; Vigier Moreno 2016). Aunque, sin duda, los estudios de traducción basados en corpus cuentan ya con un bagaje que fundamenta su consolidación, las investigaciones son más escasas en el campo de la traducción jurídica, probablemente debido a la propia naturaleza de la gran mayoría de los

documentos legales. No obstante, destacan investigaciones como las recogidas en Goźdź-Roszkowski y Pontrandolfo (2017), los trabajos sobre fraseología jurídica (Pontrandolfo 2015), estudios sobre la traducción en contextos institucionales, como puede ser la Unión Europea (Biel y Engberg 2013) o el uso de corpus en el aula de traducción jurídica (Sánchez Ramos y Vigier Moreno 2016).

En lo referido a la tipología de corpus, existen distintas clasificaciones (Hu 2016; Laviosa 2002; Zanettin 2012). Laviosa (2002) propone una clasificación que distingue entre corpus monolingües, bilingües y multilingües. Los primeros, los corpus monolingües, los desglosa en corpus monolingües simples (recopilación de textos en una única lengua) y comparables (recopilación de textos originales en una lengua A y textos traducidos en esa misma lengua A); los bilingües quedan clasificados en paralelos (textos originales en lengua A y sus traducciones en lengua B) y comparables (textos originales en lengua A y textos originales en lengua B); y los multilingües se agrupan en paralelos (textos originales en lenguas diversas con sus respectivas traducciones) y comparables ("bi/multilingual corpus made up to two or more sets from the same subject domain(s)" (Laviosa 2002: 36). Una clasificación más reciente es la ofrecida por Zanettin (2012), que, de una forma muy ilustrativa, desgrana los distintos corpus y sus posibles combinaciones (Figura 1).

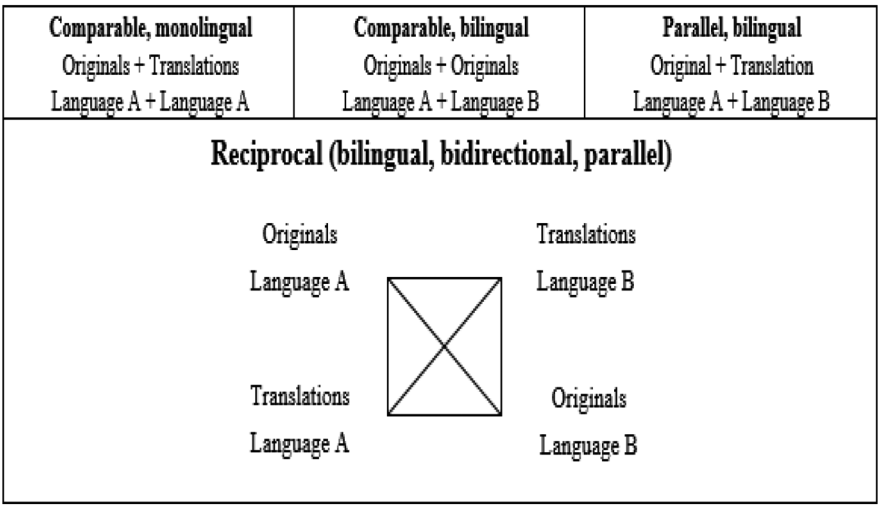

Figura 1: Tipología de corpus (Zanettin 2012: 11).

Debido a la naturaleza de nuestro trabajo, nos detendremos en los corpus paralelos. McEnery y Wilson (1996: 348) definen un corpus paralelo como "a corpus which is composed of source texts and their translations in one or more different

languages; sometimes referred to as translation corpus". Si bien el término "corpus paralelo" puede tener distintas acepciones dependiendo del ámbito de estudio, nosotros nos acercamos a este tipo de corpus desde el punto de vista de la traducción, es decir, corpus formados por originales en una lengua y sus correspondientes traducciones (Olohan 2004). Este tipo de corpus serán de gran utilidad para estudiar, por ejemplo, las distintas estrategias de traducción, la traducción de la fraseología empleada, o la naturaleza de la lengua traducida, entre otros. Como ejemplo de corpus paralelo podemos resaltar el llamado *Hansard Canadian English-French Parallel Corpus*, formado por textos extraídos del Parlamento canadiense y publicados en inglés y francés. Otros ejemplos lo constituyen el llamado *Europarl*, compilado por Koehn (2005), un corpus paralelo inicialmente formado por textos alineados en 11 lenguas. Igualmente destaca *OPUS – An Open Source Parallel Corpus*, quizás uno de los corpus paralelos multilingües más extensos, y que cubre diversos campos especializados, como puede ser el ámbito legislativo y administrativo de la UE. En términos de estructura y tipología, los corpus paralelos suelen ser unidireccionales o bidireccionales o una combinación de ambos. Los corpus paralelos unidireccionales están formados por textos en la lengua de partida y sus traducciones en la lengua de llegada. En palabras de Frankenberg-García (2009: 57): "a bidirectional corpus contains source texts in two different languages (L1 and L2) aligned with their reciprocal translations into L2 and L1". Como puede deducirse, las posibilidades de investigación son numerosas, ya que se pueden analizar las traducciones de una L1 a la L2 y viceversa.

Debido a la especificidad del lenguaje jurídico, como se ha señalado anteriormente, la equivalencia dentro de este ámbito quedará marcada por una dependencia del contexto y una determinada dependencia cultural de las lenguas implicadas en el proceso de traducción. Con todo, en nuestra opinión, la compilación de un corpus paralelo parece de gran utilidad teniendo en cuenta la escasez de los mismos dentro del ámbito jurídico.

De los beneficios que su uso puede suponer, destacamos las implicaciones en la enseñanza de terminología y patrones fraseológicos. En este sentido, rescatamos las palabras de Vargas Sierra (2002: 525):

> Aquellos profesionales que se dedican a la traducción especializada no son ajenos a los problemas que se plantean en casi, por no decir en todos, los encargos de traducción. Nos referimos a aspectos como la traducción de neologismos, la búsqueda de equivalencias y, más concretamente, las colocaciones y los términos dependientes del contexto.

La terminología, disciplina definida como "materia lingüística de carácter interdisciplinario, cuyo objetivo es el estudio y definición de los términos pertenecientes a las lenguas -o lenguajes [...] de especialidad" (Guerrero Ramos 1999: 880), es uno de los aspectos clave a la hora de enseñar traducción especializada. Y es que

no hay duda de que la traducción especializada supone una comprensión de conceptos en el texto origen si se quiere obtener una traducción eficaz, precisa y sin ambigüedades. Distintos trabajos han tratado este tema y coinciden en afirmar que el uso de corpus puede resultar beneficioso para la adquisición de terminología especializada en el aula de traducción (Faber y Jiménez Hurtado 2002; Sanchez Ramos y Vigier Moreno 2016).

Junto con la terminología, la fraseología ocupa un lugar primordial en los contextos de traducción especializada. Caracterizada por su disparidad a la hora de establecer una definición, debido en gran parte a la polisemia del término (Pontrandolfo 2011), no existe un consenso en cuanto a su denominación o taxonomía, como consecuencia de la confluencia de diversas disciplinas en su estudio a lo largo de la historia (Montoro 2017). Como bien señala Toledo Báez y Martínez Lorente (2018), esta dificultad de caracterización de la disciplina se extiende a la fraseología especializada, que es la que nos ocupa, entendida como el conjunto de unidades fraseológicas (UF), términos ampliamente estudiados en la lingüística de corpus (Hoey 2005; Sinclair 2004; Stubbs 2002), de contenido especializado de una determinada lengua. Ante ello, cabe definir dichas unidades fraseológicas especializadas (UFE) como "unidades de conocimiento especializado, que se corresponden con estructuras sintagmáticas u oracionales, no lexicalizadas, pero que presentan una cierta tendencia al estereotipo o un cierto grado de fijación, y que contienen como mínimo un término (Lorente Casafont 2002: 178). Una aportación más reciente sobre UFE la encontramos en Aguado de Cea (2007), donde se expone que las UFE se consideran estructuras sintagmáticas de al menos un término, que pueden incluir distintos elementos (verbo, sustantivo, adjetivo), que mantienen cierto grado de fijación, y, por último y de gran interés para el campo de la traducción especializada, ostentan un significado específico en un campo de especialidad en el que suelen aparecer con cierta frecuencia.

En lo referente al campo del lenguaje jurídico, los estudios han sido escasos (Ruusila y Lindroos 2016) y preferentemente centrados en cuestiones terminológicas, si bien, como indica Biel (2014b), con excepciones como los trabajos de Kjaer (2007) sobre fraseología y lenguaje jurídico, a los que podemos añadir los de Tabares Plasencia (2014).

El lenguaje jurídico se caracteriza por su carácter formulaico y expresiones fijas, entre otras, que hacen que el discurso jurídico presente ciertas peculiaridades. Existen trabajos centrados en la fraseología especializada, como son Gläser (2007) y Kjaer (2007), en el caso de la lengua alemana y que el reciente trabajo de Hourani Martín (2017) recoge de una forma detallada y minuciosa, o los de Biel (2014b) y Pontrandolfo (2011, 2015). Nos detendremos brevemente en los de Kjaer (2007) y Biel (2014b) por estar relacionados con la fraseología jurídica. En el trabajo de Kjaer (2007), la autora propone una clasificación en donde las unida-

des fraseológicas adquieren significado dentro de un contexto comunicativo legal concreto y enfatiza la intertextualidad de los textos jurídicos: "coherence of the sytem is based on the intertextual relations between legal texts, obtained by reproduction (implicit quotation) and recontextualization of words and phrases (Kjaer 2007: 513). Así, distingue entre lo que denomina "multi-word terms, collocations with a term and formulaic expressions and standard phrases" (Kjaer 2007: 509–510) y que, como afirma Hourani Martín (2017), son de importancia para el estudio de la fraseología alemana. En cuanto a la lengua inglesa, creemos oportuno mencionar la clasificación que propone Biel (2014b) por dos razones: 1) toma como punto de partida la propuesta de Kjaer (2007), trabajo esencial en fraseología jurídica y 2) se trata de una propuesta ligada a la legislación europea, temática relacionada con la compilación del corpus que se describe en este trabajo. Biel (2014b: 178–181) establece:

- Text-organizing patterns: repetitive global textual patterns which are often prescribed in drafting guidelines […];
- Grammatical patterns: genre-specific recurrent grammatical patterns […]
- Term-forming patterns (multi-word terms): collocates of a generic term which form more specific multi-word terms of varying degrees of terminologicality […]
- Term-embedding collocations: collocates of terms which embed terms in cognitive scrips and the text, evidencing combinatory property of terms […]
- Lexical collocations: routine formulae at the microstructural level which are not built around terms […]

No podemos obviar que la fraseología especializada es parte indiscutible del trabajo del traductor, y del correcto trasvase de las unidades fraseológicas especializadas dependerá la calidad de la traducción, tal y como afirma Gouadec (2007:23) cuando habla de lo que él denomina "phraseological conformity", y de la que los traductores tienen que ser conscientes, entendida como el conocimiento y correcto trasvase de la fraseología del texto origen, la organización sintáctica y textual característica de un género textual, y a la que también encontramos referencia en el trabajo de Tabares Plasencia (2012) cuando habla de competencia terminofraseológica del traductor. Por ello, el corpus, y más concretamente la metodología de corpus, puede considerarse de ayuda para solventar los problemas que puedan darse en los casos de la traducción especializada en cuanto a cuestiones fraseológicas (Goźdź-Roszkowski y Pontrandolfo 2017), o bien como herramienta para estudios terminológicos y de comportamiento lingüístico de cierto términos, como el trabajo reciente de Tabares Plasencia y Hourani Martín (2018), en el que se analizan los términos *organización criminal* y *grupo criminal organizado* dentro del corpus CRIMO.

El corpus paralelo que aquí se describe es fruto de la puesta en práctica de la metodología de corpus como recurso para hacer frente a los problemas característicos de la traducción jurídica, a la vez que servir como recurso documental de traducción. Está formado por originales y traducciones de sentencias emitidas por el TJEU y que ponen fin a procedimientos iniciados en los ordenamientos jurídicos de Inglaterra y Gales y España, y que incluyen su traducción tanto al español como al inglés respectivamente. El corpus puede ser de utilidad para el estudio del mismo desde un punto linguístico, terminológico y fraseológico, así como modelo de diseño y compilación de un corpus paralelo. Y no podemos olvidar la utilidad que ofrece nuestro corpus para las tareas de entrenamiento de sistemas de traducción automática estadística, por ejemplo (Oliver 2017).

Para la obtención de un corpus de calidad ha de seguirse un proceso de compilación claro, preciso y bien delimitado. Es por ello que la siguientes líneas están dedicadas a la descripción de las distintas fases de compilación del corpus, así como las dificultadas encontradas y soluciones propuestas que, esperemos, puedan ayudar en futuras investigaciones de traducción especializada y creación de corpus.

3 Proceso de creación del corpus paralelo

En cuanto a la creación de un corpus paralelo, deben distinguirse distintas fases bien delimitadas, como puede ser la *fase de documentación* del corpus, que incluye los criterios de diseño y búsqueda de textos originales; la *fase de compilación*, que engloba la descarga de los textos originales, la gestión de los mismos en carpetas y codificación, la conversión de formatos, y la *fase de alineación* para creación de archivos bitextos. Todas estas fases dejarán el corpus preparado para una fase final de análisis.

El proceso o *fase de documentación* es una etapa decisiva en la creación de un corpus, puesto que es en esta en donde se establecen los criterios que se van a seguir para su elaboración (Bowker y Pearson 2002). Atendiendo a la temática del corpus, las fuentes de documentación pueden ser muy variadas, desde revistas especializadas a portales especializados, por ejemplo. No hay que olvidar que hoy en día es Internet la mayor fuente de documentación, por lo que el repertorio documental de un corpus puede ser incalculable. Delimitar, pues, las fuentes documentales y utilizar una herramientas adecuadas nos ayudarán a, en último fin, obtener un corpus de calidad.

En la elaboración de nuestro corpus, se siguieron los criterios expuestos por Bowker y Pearson (2002) (el propósito de la compilación, investigación traducción jurídica; el tamaño, en nuestro caso, y siguiendo las sugerencias sobre el

tamaño de corpus jurídicos (Scott 2012), no era necesario un corpus de grandes dimensiones; el medio, escrito; la temática, jurídica; el tipo textual; la autoría, sentencias emitidas por el TJUE; la fecha de publicación, 1990–2014; y, por último, las lenguas, inglés y español). Una vez delimitado el diseño, se procedió a la búsqueda de sentencias pronunciadas por el TJUE y sus correspondientes traducciones. La fuente para la extracción de dichas sentencias fue el repositorio http://curia.europa.eu/jcms/. Sí hay que señalar que la fase de documentación y de búsqueda de material para nuestro corpus quedó restringida a este portal, puesto que era una fuente fiable para nuestros fines. En otros casos concretos, puede rastrearse la web con programas específicos, como HTTrack. Al ser sentencias públicas, no tuvimos problema en cuestiones de licencias de reproducción, un tema que puede poner trabas en la compilación de cualquier corpus (Koehn 2005). Los formatos de los archivos descargados fueron .html y .pdf, que serían tratados en la siguiente fase de compilación.

Tras el proceso de documentación, en el que, como ya se ha visto, se establecen una serie de criterios para la creación de dicho corpus paralelo de acuerdo a unos objetivos y necesidades, se obtiene el material textual que conforma el mismo y se procede a la *fase de compilación*, que incluye la normalización y almacenamiento de los archivos así como el tratamiento del formato del material. Esta etapa consta de tres estadios esenciales: la descarga de los archivos, la gestión del corpus y el establecimiento de la nomenclatura que se va a seguir, el diseño de la estructura y de la organización del corpus.Una vez ubicados los textos en el repositorio, se procedió a la descarga de los textos. Los textos origen se encontraban en su mayoría en archivos .html y .pdf. Al encontrarnos con estos formatos, el tratamiento electrónico de los archivos no fue excesivamente complicado, aunque sí han tenido que emplearse diversos programas para pasarlos a texto plano (.txt), formato con el que normalmente trabajan los gestores de corpus o programas de concordancias y los programas de alineación de textos. Para los archivos .html se utilizó el programa HtmlAsText, que permite la conversión de archivos en lote, y el programa pdf2text para el caso de los archivos .pdf. Por último, es de máxima importancia que el texto origen y el meta sean prácticamente idénticos, lo que facilitará y agilizará el proceso de alineación. En nuestro caso, se llevó a cabo una comprobación manual de todos los textos (originales y traducidos) para solventar errores menores de edición.

Tras la descarga y tratamiento de los documentos, el corpus queda organizado en dos subcorpus principales delimitados diatópicamente, coincidiendo con las lenguas de nuestro estudio (Inglaterra y Gales; España), con cada subcorpus organizado en carpetas en función de la institución a la que pertenecen los archivos, y cada una de dichas carpetas subdivididas a su vez en dos subcarpetas:

una para los archivos en la lengua origen y otra para sus traducciones. El corpus paralelo quedó configurado de la siguiente forma:

- Un subcorpus paralelo inglés-español de la jurisdicción Inglaterra y Gales, formado por 127 sentencias emitidas por el TJUE.
- Un subcorpus paralelo español-inglés de la jurisdicción España, formado por 145 sentencias emitidas por el TJUE.

El siguiente paso en esta fase es la codificación de los archivos. Esta tiene que ser unívoca para una gestión adecuada y un posterior análisis efectivo. Así, y de forma simultánea al almacenamiento de los textos, se asignó una codificación específica. Igualmente, y con el fin de crear un registro de los textos, se elaboró una tabla en formato .xls, que recogía de forma resumida los datos del corpus paralelo. En cuanto a la estructura y la organización del corpus de jurisprudencia paralelo, el primer subcorpus inglés-español queda formado por un total de 7 carpetas, correspondientes a los distintos órganos jurisdiccionales (COURT_APPEAL, CROWN_COURT, EMPLOYMENT_APPEAL_TRIBUNAL, HIGH_COURT_JUSTICE, HOUSE_LORDS, MAGISTRATES_COURT y SUPREME_COURT), que contienen un total de 127 bitextos. El segundo subcorpus, español-inglés, queda formado por 9 carpetas, correspondientes a los distintos órganos jurisdiccionales españoles (AUDIENCIA_NACIONAL, AUDIENCIA_PROVINCIAL, JUZGADO_CENTRAL_PENAL, JUZGADO_DE_LO_PENAL, JUZGADO_PRIMERA_INSTANCIA, JUZGADO_PRIMERA_INSTANCIA_INSTRUCCIÓN, TRIBUNAL_CONSTITUCIONAL, TRIBUNAL_SUPERIOR_JUSTICIA and TRIBUNAL_SUPREMO), que contienen un total de 145 bitextos.

Siguiendo nuestro protocolo de diseño del corpus paralelo, el último proceso es el de la *alineación* de los textos para su preparación hacia su explotación y análisis. La alineación, definida por Tiedemann (2011: 123) como "a process of making symmetric correspondences explicit in order to enable further processing of parallel resources" y realizada sobre dos archivos, el original y su traducción, dará como resultado un conjunto de textos paralelos denominados *bitextos*. La alineación de un corpus paralelo es una de las fases decisivas y claves en el proceso, pues de ella dependerá una correcta disposición del mismo, lo que facilitará su utilidad y el tiempo empleado en la consulta. El proceso de alineación se establece en niveles de segmentación (párrafo, oración y palabra). De forma generalizada, la segmentación oracional es la más empleada, por lo que será la utilizada en nuestro caso. La fase de alineación conlleva dos operaciones, como son la selección del programa informático más adecuado para el proceso y el proceso de alineación en sí, que requerirá una revisión manual.

Existen diversos programas de alineación, bien integrados en las herramientas de traducción asistida (Wingalign, de SDL TRADOS o LiveDocs, de MemoQ); o

sistemas de alineación como AntPConc, Bitext2tmx, Intertext, Bifid, LFAligner, YouAlign o ABBY Aligner 2.0. En la compilación de nuestro corpus paralelo, la alineación se realizó con el programa informático Intertext (Figura 2), un programa gratuito de código abierto e interfaz sencilla que ha sido desarrollado por Vondřička (2014) para el proyecto InterCorp y que está disponible para Windows, MacOS y Linux. Intertext acepta archivos bilingües en .xml y en .txt. Finalmente, en cuanto a los archivos que se exportan, cabe decir que dichos archivos pueden exportarse como un solo archivo bilingüe con extensión .tmx o como dos separados con extensión .txt. . Además, ofrece la posibilidad de codificar los archivos . txt exportados en UTF-8.

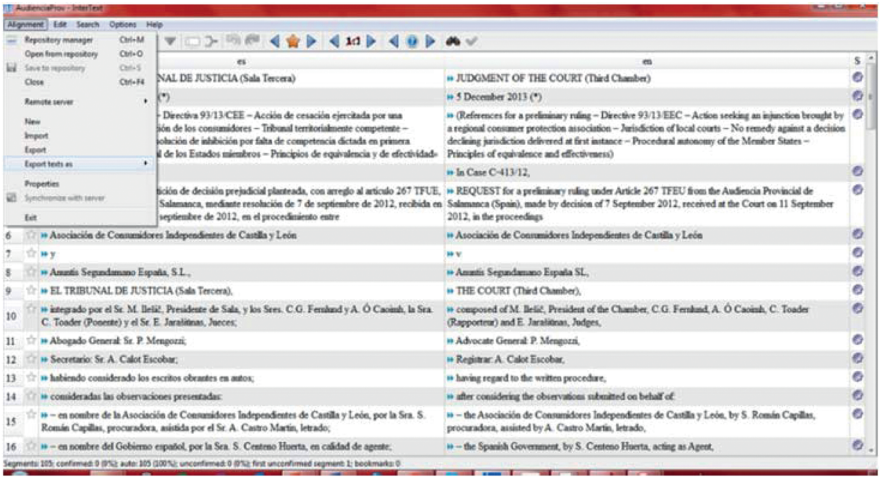

Figura 2: Interfaz de Intertext

Tras el processo de alineación automática, se procedió a la revisión manual de la misma a fin de evitar posibles fallos como la división de oraciones del original en dos o más segmentos, la fusión de oraciones del original o la omisión de información del original (Molés Casés 2016; Frankenberg-García y Santos 2003). Las alineaciones resultantes no han presentado demasiados fallos. A excepción del caso de dos pares de documentos paralelos donde se han observado problemas de alineación por el tipo de codificación de cada uno de ellos, la mayoría de los fallos encontrados se han debido a líneas en blanco existentes en alguno de los dos documentos paralelos que no aparecían en el otro documento. En la mayoría de los casos ha bastado con recurrir a las opciones de edición que InterText ofrece y combinar unos con otros para que la segmentación fuera coincidente. Las alineaciones se van guardando en el repositorio del programa. Para poder usar

posteriormente dichas alineaciones obtenidas con InterText con otros programas, estas fueron exportadas como memorias de traducción con extensión .tmx (Figura 3).

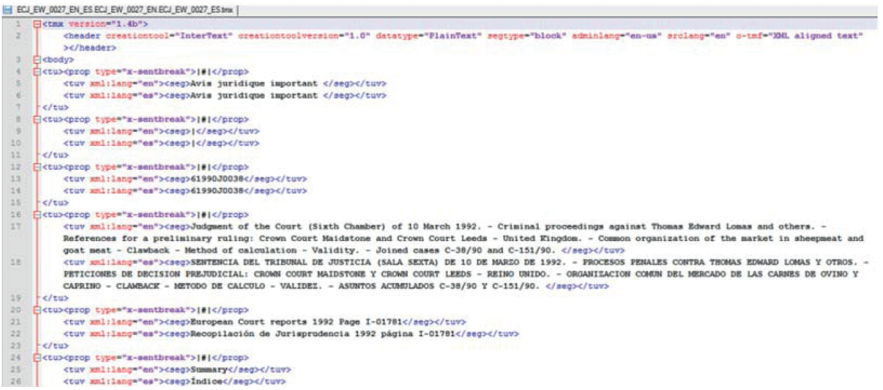

Figura 3: ejemplo de un archivo TMX

Tras el proceso de alineación, se obtuvieron un total de 127 sentencias alineadas para el subcorpus paralelo de jurisprudencia en inglés (Tabla 1), donde se obtuvieron 16012 segmentos alineados, y 145 sentencias alineadas, donde se obtuvieron 13971 segmentos alineados, para el subcorpus paralelo de jurisprudencia español (Tabla2).

Tabla 1: Corpus paralelo de Jurisprudencia inglés-español

	Archivos	Tokens	Types
EN	127	844898	13121
ES	127	938276	20262

Tabla 2: Corpus paralelo de Jurisprudencia español-inglés

	Archivos	Tokens	Types
EN	145	767661	12259
ES	145	832945	18108

Sin duda la configuración del la creación del corpus fue ardua y compleja, debido principalmente a lo oneroso de la descarga y gestión de los archivos. No obstante,

el corpus quedó preparado para su consulta en distintos programas de gestión de corpus.

De un lado, puede emplearse como fuente de documentación fraseológica y terminológica con herramientas de análisis de corpus monolingües como WordSmith, SketchEngine o AntConc, o bien con herramientas para analizar corpus paralelos como ParaConc. En este sentido, el corpus ha sido empleado como fuente de consulta para el estudio de traducción de juriculturemas (Vigier Moreno y Sánchez Ramos 2017). Igualmente, el corpus puede utilizarse como recurso terminológico para la creación de glosarios que puedan servir de ayuda en las tareas de traducción especializada, con herramientas de extracción terminológica como LexTerm, Terminology Extraction Suite (TES), TermStar o SynchroTerm, lo que lo constituye como un recurso de relevancia en las tareas de enseñanza de traducción especializada (Figura 4).

Source Entry	Target Entry	Source Occurrence	Target Occurrence		
Directive	Directiva	463	445		
COURT	TRIBUNAL	444	277		
law	Derecho	443	597		
time	tiempo	376	338		
Article	artículo	328	412		
Member	miembro	315	160		
national	nacional	269	160		
paragraph	apartado	254	389		
part	parcial	215	178		
agreement	Acuerdo	209	196		
proceedings	procedimiento	197	152		
period	período	191	111		
States	Estados	187	157		
European	Europeo	181	58		
United	Unido	159	117		
work	trabajo	158	268		
Kingdom	Reino	155	117		
paid	retribuidas	155	59		
Framework	marco	150	181		
workers	trabajadores	143	170		

Figura 4: Ejemplo de extracción terminológica generada con SynchroTerm

4 Conclusión

Este trabajo ha descrito la investigación realizada en el seno de la traducción especializada jurídica. Concretamente, se ha detallado el proceso de creación de diseño de un corpus paralelo de jurisprudencia europea bilingüe (inglés-español) a través de un protocolo de compilación articulado en tres fases (documentación, compilación y alineación).

En cuanto a la etapa de documentación, se ha utilizado el repositorio http://curia.europa.eu/jcms/ como fuente principal para la descarga de los textos originales y sus traducciones. Asimismo, en relación con la etapa de compilación, se ha seguido una nomenclatura determinada previamente definida y se ha compilado a partir de la creación de dos subcorpus: uno para los documentos legales y sus traducciones propios de órganos de jurisdicción de Inglaterra y Gales y otro para los textos legales y sus traducciones propios de órganos jurisdiccionales de España. Como parte de dicho protocolo, el proceso llamado alineación es uno de los más importantes en el diseño de un corpus paralelo. Tras la comparación de diversos programas, se eligió el programa Intertext. Tras una limpieza de archivos y una comprobación manual de los mismos, se creó un corpus paralelo bilingüe, formado por dos subcorpus atendiendo a los órganos jurisdiccionales de Inglaterra y Gales y España, con un total de 16012 y 13971 segmentos alineados respectivamente.

Tal y como indicábamos en las primeras líneas de este trabajo, el corpus paralelo bilingüe que aquí se presenta se ha diseñado como recurso para la investigación en traducción especializada, concretamente en el ámbito de la traducción jurídica. El corpus paralelo bilingüe, que se encuentra en un periodo inicial de diseño, se plantea en un futuro incorporar otras lenguas, convirtiéndose así en un corpus paralelo multilingüe, que pueda emplearse con otros fines investigadores (i.e. un estudio contrastivo de la fraseología de sentencias pronunciadas por el Tribunal Superior de Justicia de la Unión Europea, extracción de terminología especializada, recurso pedagógico en aula de traducción jurídica, o incluso entrenamiento de sistemas de traducción automática) así como hacerlo disponible a la comunidad académica. Desde un punto de vista más técnico, y con el fin de enriquecer el corpus paralelo, sería interesante utilizar programas específicos para etiquetar el corpus con lenguajes de marcado (XML, Extensive Markup Language) para añadir información metatextual (año de la sentencia, órgano jurisdiccional, lengua, etc.) y aumentar, de este modo, su potencial analítico.

Referencias

Aguado de Cea, Guadalupe (2007): "La fraseologia en las lenguas de especialidade." Alcaraz Varó, Enrique / Mateo Martínez, José / Yus Ramos, Franciso (Eds.) (2007): *Las lenguas profesionales y académicas*. Barcelona: Ariel, 53–65.

Andrades Moreno, Arsenio (2016): "Propuesta de equivalencias de binómios en la traducción jurídica inglés-español." *Estudios de Traducción* 6, 129–145.

Baker, Mona (1993): "Corpus linguistics and translation studies — Implications and applications." Baker, Mona / Gill, Francis / Tognini-Bonelli, Elena (1993) (Eds): *Text and technology*. Ámsterdam: John Benjamins, 233–50.

Biel, Lucja (2014a): "The textual fit of translated EU law: a corpus-based study of deontic modality". *The Translator* 20/3, 332–355.

Biel, Lucja (2014b) "Phraseology in Legal Translation: a Corpus-based Analysis of Textual Mapping in EU Law." Le Cheng, King / King, Sin / Wagner, Anne (Eds.) (2014): *The Ashgate Handbook of Legal Translation*. eds. Le Cheng, King Kui Sin and Anne Wagner. Londres/ Nueva York: Routledge, 177–192.

Biel, Lucja / Engberg, Jan (2013): "Research models and methods in legal translation." *Linguistica antverpiensia* 12, 1–11.

Blum-Kulka, Shoshana (2004): "Shifts of cohesion and coherence in translation." Venuti, Lawrence (2004) (Ed.): *The translation studies reader*. Londres / Nueva York: Routledge, 209–313.

Borja Albi, Anabel (2000): *El texto jurídico inglés y su traducción al español*. Barcelona: Ariel Lenguas Modernas.

Bowker, Lynee / Pearson, Jennifer (2002): *Working with specialized language*. Londres /Nueva York: Routledge.

Faber, Pamela y Catalina Jiménez Hurtado (Eds) (2002): *Investigar en Terminología*. Granada: Comares.

Fantinuoli, Claudio / Zanettin, Federico (2015) (Eds.): New directions in corpus-based translation studies. Berlín: Language Science Press.

Frankenberg-Garcia, Ana (2009). "Compiling and using a parallel corpus for research in translation." *Babel* 21/ 1, 57–71.

Frankenberg-Garcia, Ana y Diana Santos (2003). "Introducing COMPARA, the Portuguese-English parallel corpus." Zanettin, Federico / Bernardini, Silvia / Stewart, Dan (Eds.) (2003): *Corpora in translator education*. Manchester: St. Jerome, 71–87.

Gouadec, Daniel (2007): *Translation as a profession*. Ámsterdan/Philadelphia: John Benjamins Publishing.

Gläser, Rosemarie (1994): "Relations between phraseology terminology with special reference to English." *Alpha, Actes de langue francaise et de linguistique*, 7/8, 41–60.

Goźdź-Roszkowski, Stanislaw / Pontrandolfo, Gianluca (Eds.) (2017): *Phraseology in legal institutional settings: A corpus-based interdisciplinary perspective*. Londres / Nueva York: Routledge.

Guerrero Ramos, Gloria (1999): "Tecnolectos, linguajes (lenguas) específicos, especialees, especializados o de especialidade?" Fernández González, J. et al. (Eds.) (1999): *Lingüística para el siglo XXI*. Salamanca: Universidad de Salamanca, 879–887.

Hoey, Michael (2005): *Lexical priming. A new theory of words and language*. Londres/Nueva York: Routledge.

Hourani Martín, Dunia (2017): *Unidades frasaeológicas especializadas en un corpus de derecho ambiental sobre la protección frente al cambio climático (alemán-español)*. Tesis doctoral. Universidad de Granada.

Hu, Kaibao (2016): Introducing corpus-based translation studies. Berlín: Springer.

Jopek-Bosiacka, Anna (2013). "Comparative law and equivalence assessment of system-bound terms in EU legal translation." *Linguistica antverpiensia* 12, 110–146.

Kjaer, Anne Lise (2007): "Phrasemes in legal texts." Burger, Harald / Dmitri Dobrovol'skij, Dimitri / Kühn, Peter / Norrick, Neal (Eds): Phraseologie. Ein internationales Handbuch zeitgenössischer Forschung/Phraseology. An international

Handbook of contemporary research. Berlín: de Gruyter, 506–516.

Koehn, Philipp (2005): "Europarl: A parallel corpus for statistical machine translation». *Conference Proceedings: the10th Machine Translation Summit*: 79–86.

Laviosa, Sara (2002): *Corpus-based translation studies: Theory, findings, applications*. Ámsterdam: Rodopi.

McEnery, Tony / Wilson, Andrew (1996): *Corpus linguistics*. Edimburgo: Universidad de Edimburgo.

Lorente Casafont, Mercé (2002): "Terminología y fraseologia especializada: del léxico a la sintaxis." Guerrero Ramos, Gloria / Pérez Ramos, Fernando (Eds.) (2002): *Panorama actual de la terminología*. Granada: Comares, 159–180.

Molés-Cases, Teresa (2016) "Compilación y análisis de um corpus paralelo para la investigación en traducción. Proyecto com Déjà Vu, Treetagger e IMS Open Corpus Workbench." *RLA. Revista de Lingüística Aplicada* 54/1, 149–174.

Montoro del Arco, Esteban Tomás (2017): "La intersección entre composición y fraseologia: apuntes historiográficos." Echeniqie Elizondo, Mª Teresa / Pla Colomer, Francisco (Eds.) (2017): *La fraseologia a través de la historia de la lengua española y su historiografia. Valencia: Tirant lo Blach*, 213–245.

Monzó Nebot, Esther (2008). "Corpus-based Activities in Legal Translator Training." *The Interpreter and the Translator Trainer* 2/2, 221–252.

Oliver, Antoni (2017): "El corpus parallel del Diari Oficial de la Generalitat de Catalunya; compilació, anàlisi i exemples d'ús." *Zeitschrift für Katalanistik* 30, 269–291.

Olohan, Maeve (2004): *Introducing corpora in translation studies*. Londres/Nueva York: Routlegde.

Pontrandolfo, Gianluca (2015): "Investigating Judicial Phraseology with COSPE. A Contrastive Corpus-Based Study". Fantinuoli / Zanettin (2015): 137–160.

Pontrandolfo, Gianluca (2011): "Phraseology in criminal judgments: A corpus study of original vs translated italian." *Sendebar* 22, 209–234.

Prieto, Fernando (2011): "Developing legal translation competence: An integrative process-oriented approach." *Comparative legilinguistics* 5, 7–21.

Ruusila, Anna / Lindroos, Emilia (2016): " Conditio sine qua non: On phraseology in legal language and its translation." *Language and law / Linguagem e Direito* 3 (1), 120–140.

Sánchez Ramos, María del Mar (2017): "Metodología de corpus y formación en la traducción especializada (inglés-español): una propuesta para la mejora de la adquisición de vocabulario especializado." *Revista de lingüística y lenguas aplicadas* 12, 137–150.

Sánchez Ramos, María del Mar / Vigier Moreno, Francisco Javier (2016): "Using monolingual virtual corpora in public cervice legal translator training»." Elena Martín-Monje, Elena / Elorza, Izaskun / García Riaza, Blanca (2016) (Eds.). *Technological advances in specialized linguistic domains: practical applications and mobility*. Londres/Nueva York: Routledge, 228–239.

Scott, Juliette (2012): "Can genre-specific DIY corpora compiled by legal translators themselves assist them in 'Learning the Lingo' of Legal Subgenres?" *Comparative legilinguistics* 12, 87–100.

Seghiri, Miriam (2017): "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." *Babel* 63/1 43–64.

Sinclair, John (1994): *Trust the text: language, corpus and discourse*. Londres/Nueva York: Routledge.

Soriano Barbino, Guadalupe (2018): "La formación del traductor jurídico: análisis de la competencia traductora en traducción jurídica y propuesta de programa formativo." *Quaderns. Revista de traducció* 25, 217–229.

Stubbs, Michael (2002): *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell Publishers.

Tabares Plasencia, Encarnación (ed.) (2014): *Fraseología jurídica contrastiva español-alemán / Kontrastive Fachphraseologie der spanischen und deutschen Rechtssprache*. Berlín: Frank & Timme.

Tabares Plasencia, Encarnación (2012): "La competencia terminofraseológica del traductor jurídico." *Revista electrónica de didáctica de la traducción y la interpretación* 8 (1), 13–28.

Tabares Plasencia, Encarnación / Hourani Martín, Dunia (2018): "La creación terminológica en el subdomínio jurídico de la criminalidad organizada en español." *Revista de llengua i dret* 70, 133–151.

Tiedemann, Klaus (2011): *Bitext alignment*. Morgan & Claypool/Toronto.

Vargas Sierra, Chelo (2002): "Utilización de los programas de concordancias en la traducción especializada". El español, lengua de traducción. *Actas del I Congreso Internacional, Almagro*,https://rua.ua.es/dspace/bitstream/10045/13587/1/CE_VargasSierra_2002.pdf.

Toledo Báez, Cristina / Martínez Lorente, Raquel (2018): "Colocaciones, locuciones y compuestos sintagmáticos bilingues (español-francés) sobre diabetes en el corpus comparable Cordiabicom." Panace@ XIX/47, 106–114.

Vigier Moreno, Francisco Javier (2016): "Teaching the use of ad hoc corpora in the translation of legal texts into the second language». *Language and Law / Linguagem e Direito* 3/ 1, 100–113.

Vigier Moreno, Franciso Javier / Sánchez Ramos, María del Mar (2017): "Using parallel corpora to study the translation of legal-system bound terms: the case of names of English and Spanish Courts." En Ruslan Mitkov (2017) (ed.). Computational and corpus-based phraseology. Berna: Springer, 260–273.

Vondřička, Pavel (2014): "Aligning parallel texts with InterText." *Proceedings of the Ninth International Conference on Language Resources and Evaluation ,LREC'14*, 1875–1879.

Zanettin, Federico (2012): *Translation-driven corpora. Corpus resources for descriptive and applied translation studies*. Manchester: St. Jerome.