Moenia 23 (2017): 717-734.
ISSN: 2340-003X.

La construcción de un corpus paralelo bilingüe multifuncional

Irene DOVAL REIXA
Universidade de Santiago de Compostela

RESUMEN: Este artículo describe los pasos y aborda los diferentes aspectos a considerar en la construcción de un corpus bilingüe paralelo a fin de que pueda ser usado para múltiples propósitos, especialmente para la investigación en lingüística contrastiva, la traducción y la enseñanza de lenguas extranjeras. Este proceso es ejemplificado con la creación del corpus PaGeS, un corpus paralelo español/alemán, disponible vía web, que, aunque originalmente creado para la investigación lingüística, pretende cubrir un amplio rango de aplicaciones. Se describe el proceso de elaboración del corpus: compilación, preprocesado, anotación documental y lingüística y alineación de los textos. Finalmente, se presenta la interfaz web y las posibilidades de consulta para los distintos grupos de usuarios.
PALABRAS CLAVE: lingüística de corpus, corpus paralelo, lingüística contrastiva, traducción.

ABSTRACT: This article describes the steps and addresses the different aspects/issues to consider in the construction of a bilingual parallel corpus aimed to be used for multiple purposes, with special focus on the cross-linguistic research, translation and teaching of foreign languages. This process is exemplified by the creation of the corpus PaGeS, a parallel corpus German / Spanish, available for online searches via web interface. This corpus, although originally created for cross-linguistic research, aims to cover a wide range of uses. The paper describes the different phases / processes in the construction of the corpus: compilation, preprocessing, corpus markup, linguistic annotation and alignment of the data. Finally, the web interface and the search possibilities for the different user groups are presented.Keywords: corpus linguistics, parallel corpus, cross-linguistics, translation.
KEYWORDS: corpus linguistics, parallel corpus, cross-linguistics, translation.

1. INTRODUCCIÓN

En este artículo se describe la construcción de un corpus paralelo bilingüe alemán / español de textos contemporáneos que está siendo elaborado por el equipo de investigación SpatiAlEs, de la Universidad de Santiago de Compostela[1]. Este grupo se dedica desde hace años a investigaciones contrastivas alemán / español, ocupando en los últimos años un lugar central el análisis de las expresiones para los eventos de localización y movimiento, así como

[1] Este equipo, dirigido por Irene Doval Reixa, consta de miembros de las universidades de Santiago de Compostela, Salamanca, Complutense de Madrid y Valladolid. Este proyecto de elaboración del corpus está siendo financiado por el Ministerio de Economía, Industria y Competitividad (FFI2013-42571-P y FFI2017-85938-R).

Recibido: 2-5-2017. Aceptado: 29-6-2017.

a las repercusiones didácticas de este análisis para la enseñanza del alemán como lengua extranjera.

Dado que corpus paralelos español / alemán, apropiados para nuestra investigación, o bien no existen o no son accesibles (*vid.* cap. 3), constituía una necesidad la recopilación de un corpus paralelo alineado de tamaño y variedad léxica suficiente que nos suministrara la base empírica para llevar a cabo nuestras investigaciones. El corpus, aunque originalmente creado para satisfacer esta necesidad, pretende ser un recurso multifuncional útil en una amplia gama de aplicaciones. La parte del corpus PaGes (Parallel Corpus German / Spanish) actualmente en línea[2] contiene cerca de veinte millones de tokens.

El artículo está organizado de la siguiente manera: después de presentar una tipología de los corpus y sus aplicaciones se abordan las necesidades específicas de los diferentes usuarios. En el capítulo 3 se presenta una panorámica de los principales corpus paralelos existentes que incluyen al español y exponemos sus limitaciones en relación con las aplicaciones pretendidas. A continuación, se abordan los diferentes aspectos a tener en cuenta a la hora del diseño y la planificación de un corpus paralelo, la compilación de textos y su preprocesado. Los capítulos 5 y 6 tratan respectivamente la alineación oracional y la anotación de los textos. Luego se describe la presentación web del corpus y las posibilidades de búsqueda y, por último, se esbozan las líneas de trabajo futuras y se hace una breve recapitulación de los rasgos distintivos del corpus PaGeS.

2. TIPOLOGÍA DE LOS CORPUS BILINGÜES Y SUS APLICACIONES

Con respecto a la tipología de los corpus bilingües o multilingües, la terminología comúnmente aceptada (McEnery & Hardie 2012: 19-20) establece una distinción entre corpus comparables y corpus paralelos. Los corpus comparables se componen de textos monolingües en diferentes lenguas que comparten materia, género, tipo de texto, registro y, además, tienen un origen y extensión similares, tales como predicciones meteorológicas, ofertas de trabajo, artículos sobre un tema concreto, etc. Los textos de un corpus comparable no son traducciones unos de otros, pero se seleccionan de acuerdo con criterios comunes. Los corpus paralelos, por el contrario, contienen la misma colección de textos en dos o más idiomas, uno de ellos es el idioma original y los otros la traducción. Uno de los primeros corpus paralelos es el Hansard Corpus[3], formado por actas del Parlamento canadiense publicadas en inglés y francés. Los corpus paralelos pueden ser bilingües o multilingües, según consten de textos de dos o más idiomas. Pueden ser unidireccionales, si la dirección de la traducción es unívoca (por ejemplo, textos españoles traducidos al alemán), bidireccionales, cuando la traducción se efectúa en los dos sentidos (por ejemplo, textos españoles traducidos al alemán y viceversa), o multidireccionales, cuando un mismo original es traducido a varios idiomas. Un

[2] *Vid.* <www.corpuspages.eu>.

[3] La versión más actualizada comprende alrededor de 450 millones de palabras y se distribuye a través de TransSearch http://www.tsrali.com/Main.aspx?cc=true) abonando una suscripción. Una versión más antigua (1997-2000) puede descargarse gratuitamente en http://www.isi.edu/natural-language/download/hansard/.

tipo híbrido lo representan los llamados «corpus paralelos multilingües» (Zhekova *et al.* 2016: 47-8), que consisten en un texto en lengua original y múltiples traducciones en una misma lengua. Este tipo de corpus es especialmente relevante para los estudios de traducción, pues permite contrastar fácilmente diferentes estrategias de traducción y su impacto en el producto final. El diagrama siguiente presenta esquematizado la tipología de los corpus multilingües.

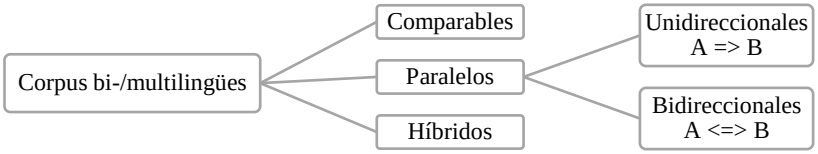

Figura 1. Tipología de los corpus multilingües.

Los corpus paralelos alineados han devenido un recurso indispensable para un amplio rango de aplicaciones multilingües y son utilizados en diferentes campos de investigación. En general se pueden distinguir cinco grandes campos de aplicaciones, cada uno de ellos con sus usuarios específicos: (a) la investigación básica en lingüística contrastiva y translatología, (b) la lexicografía, (c) la traducción, (d) la enseñanza de lenguas extranjeras y de la traducción y (e) el procesamiento del lenguaje natural, especialmente la traducción automática.

En la investigación general en lingüística contrastiva los corpus paralelos proporcionan una base empírica fácilmente accesible, produciendo patrones de correspondencia, permitiendo analizar similitudes y diferencias entre dos sistemas lingüísticos y facilitando datos cuantitativos (Doval 2016: 89-91). En los estudios de traducción resultan muy útiles para descubrir patrones colocacionales y sintácticos entre dos lenguas (Bernardini 2004: 28). Además, en el caso de los corpus bidireccionales, pueden también ser utilizados para comparar monolingualmente la lengua original y la lengua traducida y comprobar la hipótesis de Baker (1996: 175 y ss) sobre los llamados «translation universals», fenómenos típicos exclusivos de las traducciones.

Heid (2008: 137 y ss) ofrece una panorámica de las aplicaciones de los corpus paralelos alineados para la lexicografía bilingüe, subrayando su utilidad a varios niveles, al suministrar datos sintagmáticos sobre el uso de palabras en ambas lenguas. Muy relacionada con este uso se encuentra su aplicación en las labores de traducción (Johansson 2007: 21), proporcionando a los traductores equivalentes de traducción y descubriendo usos específicos de los elementos léxicos.

La utilidad de los corpus paralelos en la enseñanza de lenguas ha sido ampliamente ilustrada en la bibliografía específica (Bernardini 2004: 15 y ss). Por un lado, pueden servir de base para la elaboración de materiales didácticos y gramáticas de referencia. Además, los sistemas de concordancia bilingüe pueden utilizarse complementariamente a los diccionarios bilingües, o incluso en su lugar, ya que ofrecen una valiosa información contextual en multitud de ejemplos de uso traducidos.

Por último, los corpus paralelos se han convertido en un recurso fundamental para una amplia gama de tareas de procesamiento de lenguaje natural, como extracción de información y terminología multilingüe y, especialmente, la traducción automática. Como señalan Wetzel & Bond (2012: 28), los corpus paralelos extensos y de calidad son un recurso indispensable para entrenar sistemas estadísticos de traducción automática, que son actualmente los más utilizados.

Cada una de estas aplicaciones tiene grupos de usuarios diferenciados que demandan requisitos específicos relacionados con el diseño del corpus, el tipo de textos, su grado de anotación y los metadatos a almacenar. Tal como se ha indicado en la introducción, las aplicaciones inmediatamente previstas para el corpus PaGeS se relacionan con la investigación lingüística y con el aprendizaje y enseñanza de lenguas. Teniendo en cuenta que la creación de un corpus es una vasta tarea tanto en tiempo como en esfuerzo, el equipo ha pretendido que el corpus pueda ser explotado para múltiples propósitos y no solo para el que ha motivado su creación. Así, pretendemos que, además de la investigación lingüística y translatológica, pueda ser útil a lexicógrafos y a traductores. Como recurso de enseñanza en la clase de lengua y traducción el corpus PaGeS puede ser usado por aprendices de alemán o español de nivel intermedio a avanzado para obtener un gran número de sugerencias de traducción para un determinado ítem, mostradas directamente en ejemplos de uso.

Por ello hemos planificado la creación del corpus como un recurso multifuncional para satisfacer las diversas necesidades de estos usuarios potenciales. Para ello la interoperabilidad y estandarización de los recursos del corpus es un requisito imprescindible, que solo parcialmente ha sido alcanzado en el estadio actual de elaboración del corpus (marzo de 2017).

A continuación, presentamos una breve panorámica de los corpus bilingües español/alemán, sus características principales así como sus carencias para ciertas aplicaciones.

3. CORPUS PARALELOS DEL ESPAÑOL Y ALEMÁN

La mayoría de los corpus paralelos que incluyen español y alemán, son corpus multilingües que incluyen un buen número de lenguas europeas. Los corpus paralelos más importantes tanto por su tamaño como por su difusión son los que se derivan de textos producidos en diferentes organismos de la Unión Europea[4]. A partir de 2006, el Centro Común de Investigación (JRC)[5] de la Comisión Europea y las organizaciones de la Unión Europea han puesto a disposición una amplísima serie de recursos multilingües. Steinberger *et al.* (2014) ofrecen una visión comparativa de las diferentes colecciones multilingües proporcionadas por estas instituciones.

[4] *Vid.* <https://ec.europa.eu/jrc/en/language-technologies>. Todas las URLs han sido visitadas por última vez en marzo de 2017.

[5] *Vid.* http://ec.europa.eu/dgs/jrc/.

Koehn (2005) publicó el EuroParl, una colección de textos paralelos alineados a nivel de oración que recogen las actas de las sesiones plenarias del Parlamento Europeo desde 1996 en 11 idiomas. Actualmente el Europarl dispone de textos en las 21 lenguas oficiales de la Unión Europea[6]. La versión más reciente (Versión v7) comprende más de 50 millones de palabras para las lenguas más representadas. Los textos alemanes contienen en esta versión 47 236 849 palabras y los españoles 54 806 927.

JRC-Acquis fue el primero de los corpus preprocesados y alineados a nivel de oración distribuidos por la Comisión Europea. En su última versión (versión 3.0), comprendía 22 idiomas. La colección textual está basada en Acquis Communautaire, una colección de textos legislativos de la UE a partir de los años cincuenta.

El Corpus Digital del Parlamento Europeo (DCEP)[7] contiene los documentos publicados en el sitio web oficial del Parlamento Europeo y «constitutes the largest release of documents by a European Union institution» (Hajlaoui *et al.* 2014: 3170). Comprende una buena variedad de documentos, producidos entre 2001 y 2012, desde comunicados de prensa hasta documentos de sesión y legislativos. La versión actual del corpus contiene 162 141 documentos en inglés que funciona como lengua pivote con un total de 100 millones de palabras. El par alemán/español dispone de 103 016 documentos alineados a nivel oracional.

Hay que advertir sin embargo que todos estos corpus multilingües mencionados pueden ser descargados, muchos de ellos en formato XML, pero no ofrecen una interfaz web para ser consultados en línea, con lo cual no son en la práctica accesibles al común de los usuarios, ya que se precisan ciertos conocimientos técnicos para manejarlos. Por ello hay algunas iniciativas que hacen accesible algunos de estos corpus a través de una interfaz web. Uno de los intentos más conseguidos es el corpus Multilingwis[8], que usa los corpus paralelos del Parlamento Europeo. Contiene 220 millones de palabras en cinco lenguas (español, alemán, inglés, francés e italiano), alineados a nivel de palabra. Datos específicos sobre el tamaño del par alemán / español no se ofrecen.

Digno de mención especial es el proyecto OPUS (http://opus.lingfil.uu.se/), probablemente la mayor colección de corpus paralelos multilingües de acceso libre. Es una colección creciente de textos que proporciona a su vez herramientas para procesar datos paralelos, así como alguna interfaz para la explotación de esos datos. Según indicaciones del propio autor (Tiedemann 2012), es el par español-inglés, con aproximadamente 36 millones de oraciones paralelas, el par de lenguas con una mayor representación. Los principales ámbitos cubiertos por el OPUS son textos legislativos y administrativos, principalmente de la Unión Europea y de instituciones asociadas. También hay una buena cantidad de textos periodísticos y otras colecciones más pequeñas de varias fuentes en línea, como datos del Banco Central Europeo, subtítulos y documentación técnica.

Hoy en día, están adquiriendo gran popularidad un número creciente de herramientas en línea que combinan ciertas prestaciones de diccionarios bilingües con ejemplos de uso

6 *Vid.* <http://www.statmt.org/europarl>.
7 *Vid.* <https://ec.europa.eu/jrc/en/language-technologies/dcep>.
8 *Vid.* <https://pub.cl.uzh.ch/projects/sparcling/ multilingwis/>.

alineados paralelamente. En ellas los datos se compilan mediante un algoritmo que rastrea los sitios web bi- o multilingües de Internet y extrae automáticamente los textos correspondientes. El ejemplo más conocido y exitoso de este tipo de recursos es Linguee (<www.linguee.com>), que actualmente cubre unos 25 idiomas, según informaciones propias. Después de identificar sitios web bilingües a través de un motor de búsqueda, los documentos correspondientes se emparejan y alinean usando una programación dinámica. A diferencia de los diccionarios en línea standard[9], Linguee proporciona de este modo acceso a grandes cantidades de textos bilingües traducidos, acercándose también a lo que es una memoria de traducción. Obviamente, los pares mejor representados son los del inglés. Aunque incluye una cierta variedad de textos, la mayoría están vinculados al tipo de texto administrativo o comercial. Hay que tener presente que Linguee es un sitio comercial que ofrece explicaciones más bien parcas, tanto en cuanto a contenido exacto para los diferentes pares de lenguas como en cuanto a las herramientas usadas.

Como se indicó anteriormente, la motivación original para crear el corpus PaGeS fueron nuestros objetivos específicos de investigación contrastiva, esto es, el análisis de las expresiones de eventos de localización y sus propiedades semánticas y sintácticas en alemán y español, así como su aplicación a la didáctica del alemán como lengua extranjera. Para esta finalidad los corpus existentes, anteriormente mencionados, presentan importantes limitaciones. La limitación más importante se debe a que no aportan material suficiente. Ello obedece a que los corpus paralelos bilingües entre el alemán y el español se limitan a textos con dominios muy específicos, principalmente el lenguaje administrativo y comercial, como ya señaló hace casi una década Heid (2008: 137):«only few parallel corpora are available, and many of them are specialized in terms of text types (e. g. parliament debates) and/or domains (e. g. technical documentation)», y Steinberger *et al.* (2014: 4) siguen reconociendo: «Another restriction is linked to the text domain, which for the bulk of the corpus is legal and administrative». Debido a esta limitación, aunque los citados corpus contienen ingentes cantidades de textos, estos son monótonos tanto a nivel sintáctico como léxico y las expresiones espaciales de localización y movimiento no aparecen en ellos en un número suficiente.

Por otro lado, es unánimemente reconocido (Johansson 2007a) que para la investigación lingüística contrastiva los textos base han de ser traducciones directas entre los idiomas investigados, en nuestro caso el alemán y el español, y no dos traducciones de una tercera lengua, al margen de que se puedan incluir, tal como hemos hecho en nuestro corpus (*vid. infra*), algunos de estos textos, como mero elemento referencial. Además, ha de poder determinarse con claridad qué texto es el texto fuente y cuál el texto meta, para asegurar un equilibrio en la bidireccionalidad del corpus.

Para los textos de la Unión Europea se puede suponer que en muchos casos han sido escritos originalmente en inglés y luego traducidos a diferentes idiomas, aunque esta suposición no pasa de la mera especulación, ya que se desconoce la lengua original, tal como reconocen Steinberger *et al.* (2014: 6):

[9] En esta categoría entran, referidos al par español / alemán, entre otros los diccionarios LEO (<www.leo.org/>), dict.cc (<dees.dict.cc/>), Super-Spanisch (<www.super-spanisch.de/woerterbuch>) o PONS (<de.pons.com/%C3%BCbersetzung/deutsch-spanisch>).

> The source language for most documents produced by the EU institutions is no longer known. This information is not part of the explicit meta-information available for the documents. […] It is likely that at least some documents were translated via an intermediate language, i.e. that there are translations of translations.

En el caso de los corpus compilados automáticamente de sitios web bilingües como *Linguee*, la lengua original es todavía más difícil de determinar, y no se ofrece ninguna información al respecto. Aquí los textos podrían haber sido redactados originalmente en el idioma nativo del sitio web y luego traducidos a diferentes idiomas, pero, obviamente, se carece de cualquier evidencia que confirme esa suposición. Lo que sí podemos colegir con cierta seguridad es que en la totalidad de los corpus multilingües existentes, tanto en los textos procedentes de la UE como en los demás, las traducciones directas entre el alemán y español son excepcionales.

Por último, hay que señalar que para cualquier investigación lingüística o recurso didáctico es condición indispensable la calidad de los materiales que constituyen la base empírica de trabajo. Los corpus mencionados, a excepción de los textos de la UE, no pueden asegurar estándares de calidad ni para los textos originales ni para las traducciones, ya que no han sido sometidos a controles de calidad comprobables.

4. DISEÑO DEL CORPUS Y PREPROCESADO DE LOS DATOS

En el apartado anterior se presentaron los corpus existentes y se expusieron las razones por las que no son adecuados para ciertas aplicaciones. Esta carencia de corpus bilingües español/alemán adecuados es lo que ha constituido la motivación inmediata de la creación del corpus PaGeS. A continuación, se describirá el diseño del corpus, la compilación de datos, así como el rango de variedades de lengua y el periodo de tiempo que abarcan sus textos.

Dado que un corpus no es una representación aleatoria de textos, estos deben ser seleccionados de acuerdo a criterios específicos relacionados con el propósito para el que se crea. Por otro lado, tratándose de un corpus bilingüe, la elección de textos se tiene que restringir, obviamente, a aquellos disponibles en ambas lenguas como originales y traducciones. Aquí no se puede ignorar que la obtención de material bilingüe directo alemán / español suficiente para ciertos tipos de texto, como, por ejemplo, el material periodístico, es extremadamente problemático, ya que muy raramente se traducen artículos periodísticos entre el español y el alemán.

Dados los objetivos del corpus, la investigación lingüística y la enseñanza de lenguas, resultaba imprescindible asegurar la calidad del material, tanto con respecto a los textos originales como a su traducción. Para garantizarla, los textos tenían necesariamente que haber pasado algún control de calidad. Para ello la única vía era acudir a materiales publicados por editoriales reconocidas, donde originales y traducciones se someten a un estricto control de calidad.

Así se ha compilado un corpus de textos narrativos escritos después de 1960, aunque con un claro predominio de obras de las dos últimas décadas. En cuanto al género, están

integrados tanto textos de ficción como de no ficción, con claro predominio de los primeros, ya que constituyen la mayor parte de los recursos bilingües disponibles. Planeamos incluir textos de unas 280 obras originales y traducciones, con un tamaño total de alrededor de 20 millones de palabras. En las obras literarias se ha tratado de que cubran un rango amplio de géneros, con cierto predominio de la literatura infantil, y de autores de distintos orígenes. En los textos españoles hay una amplia representación de autores americanos y en los textos alemanes los autores suizos y austríacos están también representados. Entre las obras de no ficción se encuentran ensayos políticos e históricos, así como en general prosa de divulgación científica[10].

Como se ha mencionado anteriormente, tanto para los estudios en lingüística contrastiva, como para los de traducción, donde se estudia el proceso mismo de traducción, el equilibrio de los datos en cuanto a la dirección de la traducción es un criterio importante. Se planificó la siguiente distribución aproximada: el 45% de los textos son originales españoles, el 45% originales alemanes y el 10% restante traducciones de una tercera lengua, como elemento referencial para poder comparar también dos traducciones entre sí. La figura 2 presenta el diseño del corpus PaGeS, según la lengua original, las dobles flechas señalan los textos alineados.

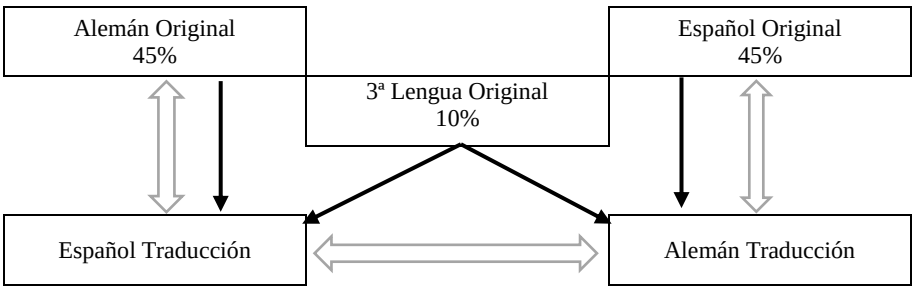

Figura 2. Diseño del corpus PaGeS.

La tabla 1 muestra la composición del corpus actualmente en línea en número de obras y en tokens distribuidas por la lengua original.

	Alemán Original	Alemán Traducción	Español Original	Español Traducción	3.ª Lengua Alemán	3.ª Lengua Español	**TOTAL**
Obras	54	38	38	54	12	12	**208**
Tokens	4 253 900	3 564 688	3 584 908	4 507 832	1 577 794	1 528 715	**19 017 837**

Tabla 1. Composición actual del corpus PaGeS (julio 2017).

Dado que se trata de obras recientes, todos los textos están protegidos por derechos de autor. Este es un tema importante en la construcción de un corpus y un factor que limita su disponibilidad. Por esta razón, muchos investigadores optan por utilizar únicamente textos

[10] En <http://www.corpuspages.eu/> puede consultarse la lista completa de obras incluidas en la versión actual en línea del corpus.

libres de derechos de autor. Pero una investigación sobre el uso contemporáneo de la lengua no es posible hacerla tomando como base textos de una antigüedad superior a los 70 años, cuando expiran los derechos de autor. Por esta razón y para respetar los derechos de autor, hemos tenido que adoptar una serie de medidas y aplicar ciertas restricciones de uso. Por un lado, la mayoría de las obras no están recogidas en su totalidad, sino solo parcialmente, pudiéndose además visualizar únicamente una pequeña cantidad de texto limitada a enmarcar contextualmente la palabra buscada (*vid. infra*). Por otro lado, el uso del corpus está restringido a fines académicos y científicos, excluyéndose cualquier uso de tipo comercial. Por ello, es obligatorio el registro consignándose la institución a la que pertenece el usuario para acceder a la totalidad de los resultados, así como para su eventual descarga.

Después de haber seleccionado y digitalizado los textos, estos se someten a un proceso manual para prepararlos para la alineación. Este consiste básicamente en lograr tanto paralelismo como sea posible entre el texto fuente y el texto meta a fin de obtener los mejores resultados en la alineación automática. Este proceso comprende la eliminación de textos no correspondientes, de caracteres erróneos y de imágenes, así como la revisión de los textos. Se eliminan todos los pasajes que no forman parte del cuerpo narrativo, tales como información bibliográfica, dedicatorias, notas del autor o traductor. De la misma manera, todos los apéndices o cualquier otro añadido que no tenga correspondencia en la otra versión se elimina.

Posteriormente son revisadas y corregidas ambas versiones, el original y su traducción. A menudo el proceso de digitalización causa algunos errores, como la inserción de un espacio dentro de una palabra, la supresión de espacios entre palabras o confusiones ocasionales de caracteres. Solo se corrigen estos errores relacionados con la digitalización, pero no eventuales errores en la edición base. Por la misma razón no adaptamos los textos a las actuales normas ortográficas alemanas o españolas en caso de que el texto usado difiera de ellas. Una vez corregidos, los textos se guardan en texto plano en un esquema de codificación común UTF-8. Estos archivos, que son los que usará el programa de alineación, forman lo que se llama un bitexto, esto es, un texto fuente y un texto meta que, en el caso de los bitextos bilingües, se corresponden por la equivalencia traductológica (Tiedemann 2011: 7).

5. ALINEACIÓN ORACIONAL

Un paso crucial en la construcción y explotación de un corpus paralelo es la alineación. Tiedemann (2011: 123) la define «as a process of making symmetric correspondences explicit in order to enable further processing of parallel resources». Este conjunto de correspondencias de dos textos, forma el bitexto alineado. Las unidades de alineación dependen de los niveles de granularidad de segmentación que se consideren: párrafos, frases o palabras. Actualmente la alineación a nivel oracional es la estándar para los corpus paralelos, por lo que nos hemos centrado en la oración como la unidad de alineación básica.

En el proceso de alineación se combinan dos tareas, una previa de segmentación de los textos y la posterior vinculación de esos segmentos con sus correspondientes en la otra

lengua para formar bisegmentos, pares alineados de segmentos (Tiedemann 2011: 7). La segmentación se lleva a cabo para cada lengua independientemente y la alineación se realiza sobre los segmentos resultantes. Tiedemann (2011: 9) insiste en la importancia de la segmentación para la precisión de la alineación posterior: «The importance of segmentation is often ignored in the literature on text alignment. However, it plays a crucial role in the success of the algorithm». Efectivamente, buen número de los fallos en la alineación se deben a fallos en la segmentación previa. Esto ocurre especialmente en los casos de puntuación no coincidente, tal como ilustra el ejemplo siguiente, en el que en español hay un punto y coma y no se han segmentado, mientras que en alemán hay un punto y se ha introducido, por tanto, una segmentación:

	Das bin ich ihm schuldig.
Se lo debo; si conservo la cordura es gracias a él.	Wenn ich bei Verstand bleibe, dann seinetwegen.

Tabla 2. Segmentación (y alineación) erróneas (Hill 2011: cap. 41).

Para la alineación oracional se han propuesto numerosos algoritmos en la literatura, los cuales se pueden clasificar en tres grupos según su metodología: enfoques basados en la longitud de los segmentos, enfoques de correspondencia léxica y enfoques híbridos. Por un lado, los enfoques basados en la longitud explotan la longitud de la oración en términos de caracteres (Gale & Church 1993) o palabras (Brown *et al.* 1993) para evaluar la probabilidad de alinear una oración de la lengua fuente conteniendo un cierto número de caracteres o palabras con otra en la lengua meta con un número similar de palabras o caracteres. Los enfoques de correspondencia léxica (Kay & Röscheisen 1993), en cambio, alinean las oraciones utilizando un método basado en el léxico, al identificar puntos de anclaje seguros para la alineación usando diccionarios bilingües o similitudes formales entre las palabras. Por último, los métodos híbridos combinan ambos enfoques y son hoy en día los más utilizados (Braune & Fraser 2010).

En el corpus PaGeS nos hemos decantado por LF-Aligner (<http://sourceforge.net/projects/aligner/>), ya que en varias pruebas esta herramienta de alineación alcanzó la mayor precisión[11]. Se basa en Hunalign (Varga *et al.* 2007), una opción común entre los creadores de corpus paralelos multilingües. Utiliza tanto la longitud de la oración como las correspondencias léxicas para derivar la alineación final. Dado que las correspondencias léxicas se derivan automáticamente, no requiere un diccionario externo. LF-Aligner realiza la alineación en tres etapas: (1) Los datos de entrada consisten en textos segmentados en dos idiomas; (2) a continuación se realiza una primera alineación, (3) finalmente, se construye un diccionario automático basado en esta alineación y luego se realinea el texto en una segunda pasada, usando ese diccionario automático. LF-Aligner ofrece varios formatos de salida: TMX, texto delimitado por tabuladores o Excel. Hemos elegido el formato de tabla de Excel, que se importa posteriormente a las hojas de cálculo de Google que son las que editamos y sobre las que hacemos la revisión, por ser más manejables y fácilmente compartibles.

[11] Hemos probado además los siguientes alineadores: CWB-align (<http://cwb.sourceforge.net/>) Vanilla (<http://nl.ijs.si/telri/Vanilla/>) y Abbyy Aligner (<https://www.abbyy.com/en-eu/aligner/>).

La alineación sería una labor trivial si a una oración en la lengua fuente le correspondiera siempre una oración en la lengua meta. Obviamente, esto no es así, ya que durante el proceso de traducción las oraciones pueden ser divididas, fusionadas, suprimidas, insertadas o reordenadas por el traductor para crear un texto natural en la lengua meta. A veces los párrafos originales se resumen, en lugar de traducirlos. Todas estas cuestiones son desafíos importantes para una alineación automática.

La precisión de la autoalineación depende enteramente de la calidad del material de origen y de la forma de la traducción. Así, LF-Aligner alcanza en algunas obras un porcentaje de precisión del 98%, mientras que en otras puede descender a niveles inferiores al 90%. En este último caso, hemos descartado las obras, ya que no disponemos de recursos para el enorme trabajo que supondría la revisión manual. Estas diferencias se deben a que el grado de correspondencia entre los textos fuente y meta varía significativamente dependiendo de los textos mismos, de los traductores y de la dirección de la traducción. Los textos en los que las versiones alemana y española son traducciones de una tercera lengua (las lenguas representadas son inglés, francés, italiano y sueco) son particularmente complejos para su alineación, ya que han sido objeto de dos procesos de traducción independientes.

Después del proceso automático, como hemos dicho, se valida manualmente esta alineación. Solo de esta manera es posible conseguir los resultados pretendidos (una tasa de error inferior al 0,5%). Para ello procedemos en dos fases. Primero localizamos bisegmentos en los que un segmento está vacío, es decir, los segmentos no emparejados en el texto fuente o en el meta. Puede tratarse de una alineación errónea, producto de una mala segmentación, como en el ejemplo de la tabla 2, en cuyo caso habría que reordenar los segmentos, pero también puede deberse a que el texto no haya sido traducido o se haya insertado texto en la traducción, en cuyo caso lo señalamos mediante un sistema de notación específico. Finalmente, con el fin de minimizar la cantidad de comprobación manual, nos centramos en bisegmentos en los que, debido a las correlaciones de longitud entre el segmento fuente y el meta, es más probable que haya errores. Para identificarlos, calculamos el cociente de la suma de caracteres del bisegmento y la diferencia de caracteres entre el segmento fuente y el meta. Entonces aplicamos esta proporción para ordenar los bisegmentos. Los errores tienden a ocurrir en aquellos en los que la diferencia de caracteres entre uno y otro segmentos es proporcionalmente mayor. Es en estos segmentos donde se realiza una revisión más exhaustiva. La comprobación manual de los resultados de la alineación puede hacerse de esta manera más eficientemente, el proceso es menos costoso en mano de obra y requiere menos tiempo. Este procedimiento es un compromiso entre lo que sería deseable —la revisión exhaustiva de todos los bisegmentos— y lo que es factible —la revisión de los bisegmentos con una mayor probabilidad de estar mal alineados—, y, con todo, es capaz de asegurar un alto nivel de precisión. Los textos manualmente validados se añaden al corpus y, además, en un futuro se utilizarán como archivos de entrenamiento del LF-Aligner a fin de mejorar sus resultados.

6. ANOTACIÓN EN LOS CORPUS PARALELOS

Tal como señalan McEnery & Hardie (2012: 29), los corpus contienen típicamente tres tipos de información no textual que se les añade para facilitar la investigación de los datos contenidos: los metadatos, el marcado textual y la anotación lingüística.

Los metadatos suministran información externa sobre los textos, como la lengua, la información bibliográfica o el tamaño a fin de, posteriormente, recuperar información relevante del corpus. El marcado textual señala la división interna de los textos, como partes o capítulos. La anotación lingüística, por último, se refiere a un cierto análisis lingüístico de los datos codificado dentro del corpus.

6.1. Los metadatos y el marcado textual

La lista de metadatos de PaGeS consta de tres partes: datos relacionados con la obra original, datos relacionados con la traducción y datos relacionados con el proceso de revisión.

Cada uno de los textos incluidos en el corpus recibe un identificador único que se utiliza como nombre de archivo y proporciona información relacionada con el idioma original y el idioma de la versión. La lista de metadatos incluye información sobre el autor y/o traductor, título, fecha, información sobre la publicación, lengua original y la lengua de la versión, género, así como información sobre el proceso de revisión y estadísticas básicas de cada obra, como el número de caracteres, tokens y segmentos que contiene.

Estas etiquetas de metadatos adicionales se adjuntan individualmente a los archivos de texto único y estos archivos adjuntos de metadatos se almacenan localmente junto con cada documento de texto individual. Por el momento, estos metadatos se guardan en un formato tabular, pero pretendemos, para una mayor estandarización, convertirlos siguiendo las directrices del esquema de metadatos más ampliamente utilizado, el TEI (Text Encoding Initiative P5, Versión 2.9.1)[12], que es una aplicación del lenguaje de marcado extensible XML para el marcado de los metadatos.

El marcado textual para señalizar las divisiones internas de los textos tales como partes, capítulos o subcapítulos, se inserta en el bitexto manualmente. El objetivo principal de este procedimiento es facilitar la localización de la cadena de texto dentro de la obra. Se ha prescindido en aquellos casos en que estaba disponible del marcado de las páginas, ya que en gran parte de las obras se ha usado directamente la edición digital y se han recogido las divisiones mencionadas de partes o capítulos que son siempre constantes en cada edición y en el texto fuente y meta.

[12] *Vid*. <http://www.tei-c.org/Guidelines/P5/>.

6.2. Anotación lingüística: lematización y etiquetado

Después de alinearlos y revisarlos manualmente, los textos han de ser de nuevo separados por lenguas para proceder a su anotación lingüística. En PaGeS consiste en lematización, esto es, la agrupación de las formas flexivas de las palabras por lemas y en su etiquetación en clases de palabras (PoS-tagging). Con la lematización, ya implementada en la versión actual 1.0, consultando cualquier forma se obtienen ocurrencias de todo el paradigma, lo que economiza sustancialmente las búsquedas. En cuanto a la etiquetación, la tarea de cualquier sistema automático consta de dos partes: búsqueda de etiquetas y desambiguación posterior. En la búsqueda de etiquetas, se determina el conjunto de las potenciales etiquetas para el token dado. En el paso siguiente de desambiguación, la lista de etiquetas posibles se reduce a la etiqueta correcta para esta instancia particular, basándose en información morfológica y contextual sobre la distribución de clases de palabras.

Tras varias pruebas se adoptaron los etiquetadores que dieron un resultado más preciso: para el alemán, TreeTagger[13], desarrollado por H. Schmid (1995), y, para el español, FreeLing[14], desarrollado por L. Padró (2011). El hecho de usar dos etiquetadores diferentes implica un trabajo previo de nivelación, puesto que tanto las etiquetas como su nivel de detalle difieren considerablemente.

FreeLing usa el etiquetado EAGLES[15] que consta de 577 etiquetas para el español, que contienen información flexiva como el género, número, gradación o tiempo y persona para los verbos. Estas etiquetas se agrupan en 12 categorías básicas: adjetivos, adverbios, determinantes, nombres, verbos, pronombres, conjunciones, interjecciones, preposiciones, signos de puntuación, numerales y fechas y horas. Por el contrario, el inventario de etiquetas de TreeTagger para el alemán se reduce a 54, agrupadas en 11 clases de palabras (nombres, verbos, artículo, adjetivos, pronombres, cardinales, adverbios, conjunciones, adposiciones, interjecciones y partículas) a las que se añaden tres etiquetas para extranjerismos, signos de puntuación y resto.

Como se puede observar en la tabla 3, las etiquetas de FreeLing consisten en etiquetas de longitud variable donde cada carácter corresponde a una característica morfológica. El primer carácter de la etiqueta designa la categoría (PoS). La categoría determina la longitud de la etiqueta y la interpretación de cada carácter. Las características que no son aplicables o no se especifican para una palabra en particular se marcan como 0. A diferencia de estas, las etiquetas de TreeTagger son más generales, sin información flexiva y, ocasionalmente con alguna información sintáctico-distribucional que no ofrece FreeLing. Un ejemplo de ello son los adjetivos que TreeTagger divide en dos clases: atributivos (ADJA) y apositivos o adverbiales (ADJD). La tabla 3 muestra las diferencias en la etiquetación entre ambos programas.

---

13 *Vid*. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger>.

14 *Vid*. <http://nlp.lsi.upc.edu/freeling>.

15 EAGLES es una iniciativa de la Comisión Europea para la elaboración de las directrices EAGLES, que establecen recomendaciones en relación con las labores de procesamiento de lenguaje natural, entre otros sobre anotación y marcado de corpus. Actualmente está coordinada por el Consorzio Pisa Ricerche, Pisa, Italia, <http://www.ilc.cnr.it/EAGLES/browse.html>. Para el repertorio completo de las etiquetas usadas para el español, *vid.* <http://nlp.lsi.upc.edu/freeling-old/doc/tagsets/tagset-es.html>.

Español: FreeLing			Alemán: TreeTagger		
Token	**Etiqueta**	**Lema**	**Token**	**Etiqueta**	**Lema**
Fuera	RG	fuera	Draußen	ADV	draußen
hacía	VMII3S0	hacer	war	VAFIN	sein
una	DI0FS0	uno	ein	ART	eine
mañana	NCCS000	mañana	grauer	ADJA	grau
fría	AQ0FS00	frío	kalter	ADJA	kalt
y	CC	y	November-morgen	NN	November-morgen
gris	AQ0CS00	gris	,	$,	,
de	SP	de	und	KON	und
noviembre	W	[??/11/??]	es	PPER	es
,	Fc	,	regnete	VVFIN	regnen
y	CC	y	in	APPR	in
llovía	VMII3S0	llover	Strömen	NN	Strom|Strömen
a_cántaros	RG	a_cántaros	.	$.	.
.	Fp	.			

Tabla 3. Etiquetado de FreeLing y TreeTagger.

Para elaborar un sistema de búsquedas de etiquetas común a ambas lenguas, hemos desarrollado un inventario de etiquetas para PaGeS, que recoge el mínimo común denominador de ambos etiquetadores. Sobre 10 clases de palabras básicas: adjetivo (ADJ), adverbio (ADV), preposición, (PREP) artículo (ART), cardinal (CARD), conjunción (CONJ), sustantivo (N), pronombre (PRON), verbo (V) e interjección (ITJ) se puede refinar la búsqueda a las especificaciones comunes a ambos etiquetadores, como por ejemplo los modos verbales o la diferencia entre nombre común y propio. Una vez que se implemente este etiquetado, será posible realizar búsquedas no sólo por cadenas de texto, sino también por categorías de clases de palabras o también combinar ambas búsquedas.

7. PRESENTACIÓN Y EXPLOTACIÓN

Una vez que a los archivos de texto se les ha añadido la anotación lingüística, son indexados y gestionados por el indexador y motor de búsqueda Apache-Solr (<http://lucene.apache.org/solr/>), una potente y muy rápida plataforma de búsqueda de código abierto escrita en Java, que cubre una amplia gama de funcionalidades.

Uno de los mayores desafíos a la hora de la confección de un corpus radica en su accesibilidad. De hecho, los corpus bilingües que hemos presentado, como ya se ha indicado, no ofrecen una interfaz con posibilidad de búsqueda en línea. La facilidad e inmediatez de uso de un corpus, junto con distintas posibilidades de búsqueda, es la clave para su éxito. En la confección de un corpus paralelo hay una imperiosa exigencia de conciliar las necesidades de los distintos grupos de usuarios a los que hemos aludido en el apartado 2, esto es, ofrecer un potente motor de búsqueda a usuarios exigentes como los investigadores, pero sin perder, por ofrecer una máscara demasiado compleja, a otros usuarios que simplemente quieren consultar el uso de una palabra o expresión, o encontrar rápidamente un equivalente de traducción. A tal fin, hemos diseñado una búsqueda a tres niveles. En primer lugar, está la búsqueda simple o standard cuya interfaz web provisional se muestra en la figura 3. Aquí el usuario

simplemente tiene que introducir la palabra o palabras en español o alemán. Estas búsquedas son por defecto lematizadas, a no ser que se inserte la palabra entre comillas, con lo cual se restringe la búsqueda a la palabra exacta. Si se hace una búsqueda multipalabra, se obtienen contextos en que aparecen todas las palabras introducidas lematizadas, siempre que se encuentren a una distancia máxima de cinco palabras. Así se obtiene con celeridad una lista de resultados con su traducción, los textos originales en la columna izquierda, las traducciones en la derecha. La cadena buscada se resalta en negrita y se muestra junto con un pasaje de texto anterior y posterior, tal como se ve en la figura 3. Los resultados pueden ser posteriormente exportados a archivos de texto o Excel y descargados localmente

Figura 3. Interfaz de usuario provisional de PaGeS.

Si se pincha en [source], se llega a otra página donde se muestran las indicaciones bibliográficas del pasaje y donde se puede ampliar el contexto de la palabra buscada, tal como muestra la figura 4.

Figura 4. Ampliación de contexto e información bibliográfica.

En la búsqueda avanzada, aún no implementada, se ofrece la posibilidad de refinar la búsqueda aplicando una serie de filtros, que se presentan mediante menús desplegables. Estos filtros son los siguientes: obra, autor, fecha de publicación, original o traducción y la

búsqueda por el equivalente de traducción. Está previsto que los resultados puedan ser ordenados siguiendo distintos criterios.

Por último, está la búsqueda más compleja, que se hace directamente sobre la interfaz de la búsqueda standard, pero usando el lenguaje propio del Solr, que soporta la búsqueda con expresiones regulares regex[16]. Obviamente aquí se requiere un mayor esfuerzo por parte del usuario que ha de estar familiarizado con este lenguaje.

8. TRABAJO FUTURO Y NOTAS FINALES

Este artículo describe los pasos y aborda diferentes aspectos que han de ser tenidos en cuenta en la construcción de un corpus paralelo bilingüe para múltiples propósitos, ejemplificado en la creación del corpus PaGeS, un corpus alemán/español, que, aunque originalmente creado para fines de investigación lingüística, pretende cubrir un amplio abanico de usos.

Dado que es un proyecto en curso, continuamos, por una parte, añadiendo nuevas obras para llegar a la cantidad y equilibrio mencionados. Para ampliar la variedad de género textual, tenemos previsto incorporar algunos textos periodísticos, aunque siempre representarán una proporción pequeña, dada la limitada disponibilidad de recursos bilingües en este campo. Por otro lado, estamos implementando otras funcionalidades como la búsqueda por etiquetas y la alineación de palabras. Con respecto a la alineación de palabras, se están probando varias herramientas (Giza ++, Nattools y Berkeley aligner), pero aún no se ha tomado una decisión definitiva. Otra mejora prevista se relaciona con la lematización de los verbos alemanes con prefijos separables, utilizando un script de M. Volk (Volk *et al*. 2014).

A más largo plazo estamos sopesando cambiar el motor de búsqueda. Solr, al tratarse de una plataforma general, no específica para corpus, presenta ciertas limitaciones a la hora de proceder a búsquedas complejas, combinando distintos parámetros. Tales consultas solo pueden ser procesadas por un sistema de consulta específico que permita consultas sobre distintos niveles de anotación lingüística. Esta es la razón por la que estamos explorando el uso de Corpus Workbench, una de las herramientas de código abierto más conocidas para gestionar y consultar corpus con anotaciones lingüísticas, desarrollado en el IMS de la Universidad de Stuttgart.

A pesar de la existencia de otros corpus paralelos, PaGeS presenta una serie de características distintivas, entre las que cabe destacar: el tipo de textos utilizado de gran variedad léxica y gramatical, la alta calidad de originales y traducciones, el equilibrio en la bidireccionalidad, la revisión manual de los procesos automáticos, la anotación y el etiquetado lingüístico y, por último, su disponibilidad en línea y su facilidad de uso. Todas estas características lo convierten en un recurso adecuado para múltiples aplicaciones, desde la investigación contrastiva, pasando por los estudios de traducción hasta el aprendizaje y enseñanza de lenguas extranjeras.

[16] *Vid*. <http://www.openjems.com/solr-regex-tutorial/>.

REFERENCIAS BIBLIOGRÁFICAS

BAKER, M. (1996): "Corpus-based translation studies: The challenges that lie ahead". En H. Somers (ed.): *Terminology, LSP and Translation*. Amsterdam: Benjamins, 175-86.

BERNARDINI, S. (2004): "Corpora in the Classroom: An Overview and Some Reflections on Future. Developments". En J. Sinclair (ed.): *How to Use Corpora in Language Teaching.* Amsterdam: John Benjamins, 15-36.

BRAUNE, F. & A. FRASER (2010): "Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora". *COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics*: Posters. Beijing, China, 81-9.

BROWN, P. F., J. C. LAI & R. L. MERCER (1991): "Aligning Sentences in Parallel Corpora". *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics, ACL '91*. Stroudsburg, PA: ACL, 169-76.

DOVAL, I. (2016): "Bilingual Parallel Corpora for Linguistic Research". *EPiC Series in Language and Linguistics* 1, 88-96.

GALE, W. A. & K. W. CHURCH (1993): "A Program for Aligning Sentences in Bilingual Corpora". *Computational Linguistics* 19/1, 75-102.

HAJLAOUI, N. *et al.* (2014): "DCEP-Digital Corpus of the European Parliament". *Proceedings LREC 2014 (Language Resources and Evaluation Conference). Reykjavik, Iceland. Mai 26-31, 2014*, 3164-71. En línea: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/943_Paper.pdf>.

HEID, U. (2008): "Corpus linguistics and lexicography". En Lüdeling & Kytö (2008: 131-53).

LÜDELING, A. & M. KYTÖ (2008): *Corpus Linguistics. An International Handbook*. Vol. 1. *Handbücher zur Sprach- und Kommunikationswissenschat*. Berlin: Walter de Gruyter.

HILL, T. (2011): *El verano de los juguetes muertos*. Barcelona Penguin Random House. [*Der Sommer der toten Puppen*. Berlin: Suhrkamp, 2013.]

JOHANSSON, S. (2007a): *Seeing through Multilingual Corpora: On the use of corpora in contrastive studies.* Amsterdam: John Benjamins.

JOHANSSON, S. (2007b): "Using Corpora: From Learning to Research". En E. Hidalgo, L. Quereda & J. Santana (eds.): *Corpora in the Foreign Language Classroom.* Amsterdam: Rodopi, 17-30.

KOEHN, P., (2005): "EuroParl, A parallel corpus for statistical machine translation". *Proceedings of the machine translation summit*. Phuket: AAMT, 79-86. En línea: <http://www.statmt.org/europarl/>.

KAY, M. & M. RÖSCHEISEN (1993): "Text-translation Alignment". *Computational Linguistics* 19.1, 121-142.

MCENERY, T. & A. HARDIE (2012): *Corpus Linguistics*. Cambridge: Cambridge University Press.

PADRÓ, L. (2011): "Analizadores Multilingües en FreeLing". *Linguamatica* 3/2, 13-20.

RÖMER, U. (2008): "Corpora and language teaching". En Lüdeling & Kytö (2008: 112-31).

SCHMID, H. (1995): "Improvements in Part-of-Speech Tagging with an Application to German". *Proceedings of the ACL SIGDAT-Workshop*. Dublin, 47-50. En línea:<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf>.

STEINBERGER R. *et al.* (2014): "An overview of the European Union's highly multilingual parallel corpora". *Language Resources and Evaluation Journal* 48/4, 679-707.

TIEDEMANN, J. (2011): *Bitext Alignment*. Toronto: Morgan & Claypool.

TIEDEMANN, J. (2012): "Parallel Data, Tools and Interfaces in OPUS". *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC-2012). Paris: ELRA, 2214-8. En línea: <www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf>.

VARGA, D. *et al.* (2007): "Parallel corpora for medium density languages". En N. Nicolov *et al*. (eds.): *Recent Advances in Natural Language Processing IV*. Amsterdam: John Benjamins, 590-6.

VOLK, M., J. GRAËN, & E. CALLEGARO (2014): "Innovations in Parallel Corpus Search Tools". En N. Calzolari *et al.* (eds.): *Proceedings LREC 2014*, 3172-8. En línea: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/504_Paper.pdf>.

WETZEL, D. & F. BOND (2012): "Enriching parallel corpora for statistical machine translation with semantic negation rephrasing". En M. Carpuat, L. Specia & D. Wu (eds.): *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Stroudsburg: ACL, 20-9. En línea: <http://aclweb.org/anthology/W12-4203>.

ZHEKOVA, D. *et al.* (2016): "Alignment and Application of Russian-German Multi-Target Parallel Corpora for Linguistic Analysis and Literary Studies". *MATLIT* 4/1, 45-61. En línea: <http://dx.doi.org/10.14195/2182-8830_4-1_3>.