RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 54 (1), I Sem. 2016, pp. 149-174.

CL ISSN 0033 - 698X

COMPILACIÓN Y ANÁLISIS DE UN CORPUS PARALELO PARA LA INVESTIGACIÓN EN TRADUCCIÓN. PROYECTO CON *DÉJÀ VU, TREETAGGER* E *IMS OPEN CORPUS WORKBENCH*[*]

COMPILATION AND ANALYSIS OF A PARALLEL CORPUS FOR RESEARCH IN TRANSLATION. PROJECT WITH *DÉJÀ VU, TREETAGGER* AND *IMS OPEN CORPUS WORKBENCH*

TERESA MOLÉS-CASES
Universitat Jaume I. Castellón, España
tmoles@uji.es

RESUMEN

Aunque en los últimos años la lingüística de corpus ha experimentado una gran evolución y en la actualidad cuenta con una creciente presencia en proyectos de investigación en torno a estudios de Lingüística y Traducción (por ejemplo: Kübler y Foucou, 2003; Laroche y Langlais, 2010), los procedimientos técnicos más avanzados enfocados a la compilación y explotación de corpus siguen siendo un escollo. El principal propósito de este trabajo es, por tanto, hacer accesible este tipo de información a toda la comunidad investigadora poco experta en la materia. En concreto, presenta la experiencia de creación de un corpus paralelo alineado con *Déjà Vu*, etiquetado lingüísticamente con *TreeTagger*, documentado con *Notepad++* e indexado con *IMS Open Corpus Workbench*. Además, incluye una breve introducción a la exploración y el análisis de corpus con *Corpus Query Processor*, la principal herramienta de *IMS Open Corpus Workbench*.

Palabras clave: Lingüística de corpus; *Déjà Vu; TreeTagger; IMS Open Corpus Workbench.*

[*] Este trabajo ha sido posible gracias a los proyectos "Refinamiento y sistematización del análisis del corpus COVALT a través de su preprocesamiento y ampliación mediante la inclusión de traducciones al castellano" (FFI2012-35239/FILO) del Ministerio de Educación de España y "Los corpus en la enseñanza de la traducción. Ampliación y explotación didáctica del corpus COVALT" (P1.1B2013-44) de la Universitat Jaume I (España) y a una ayuda para movilidad del personal investigador de la Fundació Caixa Castelló-Bancaixa ("Acción 2 del Plan de promoción a la investigación de la Universitat Jaume I para el curso 2012/2013") en la Universität Leipzig (Alemania). Quiero agradecerles a Ulrike Oster, Víctor González, Daniel Renau, Francisco Nevado y a los dos evaluadores anónimos sus consejos y comentarios.

ABSTRACT

Although Corpus linguistics has advanced a great deal in recent years and is now being increasingly more frequently included within research projects regarding Linguistics and Translation (for instance: Kübler & Foucou, 2003; Laroche & Langlais, 2010), the most advanced technical procedures focused on the creation and exploitation of corpora are still a pitfall. The main aim of this paper is, then, to make this kind of information more widely available to the research community with little experience in the field. In particular, it presents the experience of creating a parallel corpus that was aligned with *Déjà Vu*, linguistically tagged with *TreeTagger*, meta-textually tagged with *Notepad++* and indexed with IMS Open Corpus Workbench. Furthermore, it includes a brief introduction to the exploration and analysis of corpora with *Corpus Query Processor*, the main tool of *IMS Open Corpus Workbench*.

Keywords: Corpus linguistics; *Déjà Vu; Tree Tagger; IMS Open Corpus Workbench.*

Recibido: 15.06.2015. Aceptado: 25.01.2016.

1. INTRODUCCIÓN

Existe un amplio consenso en cuanto a que la lingüística de corpus constituye una herramienta óptima para el estudio de fenómenos lingüísticos y traductológicos (véase, por ejemplo: Bernardini, 2004; Aston, 2009; Kübler, 2011). Cuando un investigador dispone del corpus adecuado para sus fines y una interfaz de búsqueda que le permita extraer de él la información específica que busca, los corpus electrónicos no solamente suponen un importante ahorro de tiempo, sino que hacen posibles análisis que serían inviables a partir de un procedimiento manual.

Pero no siempre existe un corpus de las características necesarias y en muchas ocasiones el investigador se ve obligado a diseñar y crear su propio corpus y dotarlo del formato adecuado para consultarlo. Ésta no es una tarea sencilla y el investigador deberá adquirir conocimientos técnicos y metodológicos diversos, que a veces no son fáciles de encontrar, especialmente cuando se trata de corpus complejos de finalidades muy específicas como pueden ser los corpus paralelos o comparables utilizados en los estudios de traducción[1].

Por otra parte, la literatura especializada en torno al uso de los programas para

[1] Un corpus paralelo está formado por textos originales en una lengua A y sus correspondientes traducciones a una lengua B. Un corpus comparable es una colección de textos originales en una lengua A que pertenecen al mismo ámbito de especialidad y tienen un contenido similar a los textos sometidos a traducción a una lengua B, con los que se suelen comparar. Véase Kenning (2010) para obtener más información sobre ambos tipos de corpus.

el procesamiento de los datos a menudo requiere un nivel considerable de conocimientos en lenguaje de programación, como pasa por ejemplo en el caso de algunos manuales para la indexación de corpus (véase, por ejemplo: Christ, Schulze, Hofmann y König, 1999; Evert y OCWB, 2005). Ante esta dificultad, el propósito del presente artículo es eminentemente metodológico, con un enfoque muy práctico, centrándose en las cuestiones que causan más dificultades por su carácter predominantemente técnico (Fig.1):

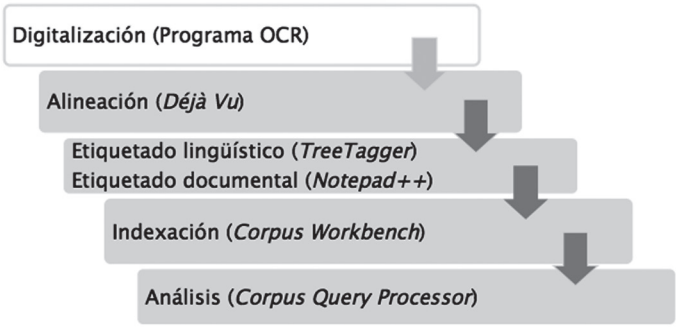

Figura 1. Pasos para compilar un corpus paralelo etiquetado.

En concreto, se describe detalladamente el procedimiento seguido para la creación de un corpus paralelo alemán-español de literatura infantil y juvenil en el sistema operativo Microsoft Windows[2]. Nuestro punto de partida es la consideración de la expresión de la manera de desplazamiento como problema de traducción. Con dicho corpus se persigue estudiar el proceso de traducción de la manera de desplazamiento de una lengua de marco satélite (alemán) a una lengua de marco verbal (español) (Talmy, 1975, 1985, 1991) y proponer un listado de técnicas de traducción adaptado a dicho fenómeno.

2. FUNDAMENTOS TEÓRICOS: LA LINGÜÍSTICA DE CORPUS, LA LINGÜÍSTICA APLICADA Y LA TRADUCCIÓN

Un corpus se puede definir, sucintamente, como una herramienta muy fiable que permite validar de forma empírica teorías a partir de una muestra lingüística real y representativa. Lo que caracteriza especialmente a los corpus, en comparación

[2] Concretamente la versión utilizada ha sido Microsoft Windows XP, Home Edition, 2002, Service Pack 3.

con otros recursos lingüísticos, es que se crean a partir de criterios de diseño específicos y que son representativos del uso real de una lengua o de una variedad de lengua. En palabras de Braun (2005), los datos proporcionados por la lingüística de corpus son:

> realistic, showing language in real use; rich, providing more (and more diversified) information than dictionaries or reference grammars can; illustrative, providing actual patterns of use instead of abstract explanations; up-to-date, revealing trends in language use and evidence for short-term historical change (Braun 2005: 48).

En la literatura no existe todavía consenso en cuanto a si la lingüística de corpus es una metodología – "a way of doing linguistics" (Meyer, 2002: xi) –, una teoría o disciplina – "its unique approach to the study of language […] is firmly based on the integration of four interdependent, equally important elements: data, description, theory, and methodology" (Laviosa, 2002: 8) – o si ésta está a caballo entre la metodología y la teoría: "Corpus linguistics is viewed primarily as a methodology, not a theory. However, this should not be understood simply to imply that corpus linguistics is theory-free. The focus and method of research, as well as the type of corpus selected for a study, is influenced by the theoretical orientation of the researchers, explicit or implicit" (McEnery y Gabrielatos, 2006: 35).

La lingüística de corpus ha gozado de una larga trayectoria en Lingüística Aplicada. Si bien existen algunos estudios de corpus realizados antes de los años cincuenta, el origen de la lingüística de corpus podría establecerse a principios de los años sesenta, tras la creación en 1959 del *Survey of English Usage Corpus* (SEU)[3]. Desde entonces, y gracias a la revolución del *software* y del *hardware*, los corpus han hecho posibles nuevas formas de investigar en Lingüística Aplicada, por ejemplo: estudios cuantitativos, búsqueda por lema, comparación del uso y la norma, entre otros[4]. Los corpus han revolucionado casi todas las disciplinas lingüísticas (McEnery, Xiao y Tono, 2006), como la lexicografía y la gramática, los estudios lingüísticos contrastivos, el estudio de lenguas, la ingeniería lingüística, las tecnologías telemáticas o los estudios de traducción, por poner sólo algunos ejemplos[5].

El uso del corpus en los estudios de traducción es relativamente nuevo, cuenta con un par de décadas. En 1993, Mona Baker propuso usar los corpus comparables monolingües para analizar el proceso de traducción y estudiar las características de la lengua traducida. Desde entonces los corpus han demostrado ser un

[3] El SEU es un corpus de inglés británico hablado y escrito creado por Randolph Quirk en la University College London.

[4] Véase Kennedy (1998) y Corpas Pastor (2008) para un panorama más amplio sobre la evolución histórica de la Lingüística de corpus.

[5] Hunston (2002) recoge las principales aplicaciones de la Lingüística de corpus en Lingüística Aplicada.

recurso excelente no sólo para estudios de corte más descriptivo (por ejemplo: Baker, 1995; Laviosa, 1998; Mauranen y Kujamäki, 2004), sino también para cuestiones de índole más práctica, como la traducción profesional, la traducción automática o la enseñanza de la traducción y la interpretación (por ejemplo: Zanettin, Bernardini y Steward, 2003; Beeby, Rodríguez y Sánchez-Gijón, 2009)[6]. Baker (1995: 227) distingue los tres tipos de corpus que pueden ser de especial interés para el investigador en traducción: el corpus paralelo, el comparable y el bilingüe/multilingüe[7]. Este artículo se centrará en el corpus paralelo, que sirve principalmente para estudiar las estrategias que utilizan los traductores cuando se enfrentan a determinados problemas de traducción y para observar la naturaleza de la lengua traducida.

3. METODOLOGÍA

3.1. Procedimiento

La metodología que presenta este trabajo forma parte de una tesis doctoral desarrollada en el marco de los Estudios de Traducción de la Universitat Jaume I (España). Como se explicó en la introducción, este artículo narra la experiencia de compilación de un corpus paralelo alemán-español de literatura infantil y juvenil que data de 1973 a 2011 y que contiene 18 novelas originales en alemán y sus respectivas traducciones al español, con un total de 916.063 palabras (*tokens*). En la Tabla I se indican las 18 novelas originales que componen la base empírica:

Tabla I. Novelas del corpus paralelo.

NOVELA	AUTOR
Momo	Michael Ende
Vorstadtkrokodile	Max von der Grün
Das Geheimnis des Brunnens	Luise Rinser
Ben liebt Anna	Peter Härtling
Stolperschritte	Mirjam Pressler
Anne will ein Zwilling werden	Paul Maar
Die Wartehalle	Klaus Kordon
Das Fünfmarkstück	Klaus Kordon

[6] Véase Laviosa (2002), Olohan (2004) y O'Keeffe y McCarthy (2010) para obtener un panorama más general sobre la aplicación de la lingüística de corpus en traducción.

[7] Un corpus bilingüe/multilingüe está formado por dos o más corpus monolingües escritos en diferentes lenguas, producidos con el mismo criterio para la misma o diferentes instituciones.

Continuación Tabla I.

Die Unterirdischen	Angela Sommer-Bodenburg
Der neue Pinocchio	Christine Nöstlinger
Die Geschichte von der Schüssel und vom Löffel	Michael Ende
Wenn du dich gruseln willst	Angela Sommer-Bodenburg
Spürnase Jakob-Nachbarkind	Christine Nöstlinger
Als der Weihnachtsmann vom Himmel fiel	Cornelia Funke
Die Zauberschule	Michael Ende
Reise gegen den Wind	Peter Härtling
Der verborgene Schatz	Paul Maar
Rico, Oskar und die Tieferschatten	Andreas Steinhöfel

El objetivo específico de la creación de dicho corpus reside en el interés por estudiar cómo se traduce la expresión de la manera de desplazamiento del alemán al español, dado que, según la teoría de los patrones de lexicalización de Talmy (1975, 1985, 1991), la lengua alemana y la lengua española pertenecen a tipologías diferentes con respecto a cómo codifican el componente Trayectoria. El alemán es una lengua de marco satélite y expresa el Movimiento y la Manera o la Causa a través del verbo principal y la Trayectoria se expresa normalmente a partir de partículas adjuntas a esta forma, a las que Talmy denomina 'satélites'. El español es una lengua de marco verbal y normalmente expresa el Movimiento y la Trayectoria a través del verbo principal, mientras que la Manera o la Causa, en caso de ser relevante, se considera un coevento y se expresa a través de otras formas (por ejemplo, un gerundio).

Estas diferencias tipológicas se atribuyen principalmente a dos circunstancias. Por una parte, a una cuestión discursiva, es decir, en las lenguas de marco verbal se requiere un esfuerzo cognitivo añadido para expresar la Manera (Slobin, 2006) y por tanto generalmente ésta sólo se expresa si es imprescindible para la caracterización del evento. Por otra parte, a una cuestión léxica, es decir, las lenguas de marco verbal tienen un repertorio verbal más reducido y menos expresivo para la lexicalización del componente Manera y en muchas ocasiones esta información debe ser interpretada e inferida del contexto (McNeill, 2000; Özcalışkan y Slobin, 2003). Estas divergencias pueden provocar algunas dificultades en lo que respecta al procesamiento del lenguaje y a la traducción, como por ejemplo, omisión de componentes del evento de movimiento (Filipović, 2007). Nuestro interés reside en valorar si se mantiene, omite, añade o modifica la información relativa a la manera de desplazamiento en cada una de las concordancias extraídas del corpus y finalmente proponer un listado de técnicas de traducción adaptado a dicho fenómeno.

3.2. Compilación del corpus paralelo: digitalización y alineación

• ***Digitalización***

En primera instancia, lo que condiciona las características y procedencia de los textos de un corpus es el objeto de estudio específico para el cual éste es creado. Tras una primera reflexión deberá explorarse si dichos textos están disponibles en formato de papel o electrónico.

Si el material objeto de estudio está sólo disponible en papel, el primer paso es digitalizarlo con un programa que utilice un algoritmo de reconocimiento óptico de caracteres (OCR). Digitalizar es una tarea sencilla que por lo general no causa problemas. No obstante, tras la digitalización es necesario leer y revisar las obras escaneadas. Se suele realizar una tarea de preprocesamiento, donde es recomendable comparar los textos, sobre todo, a nivel estructural, ya que esto evitará posibles problemas posteriores. Es necesario repasar sobre todo si se ha pasado por alto algún número de página, intertítulo, nota al pie, etc. Para una mayor orientación sobre preprocesamiento, véase Bowker y Pearson (2002: 96-97, 100) y Frankenberg-García (2007). Puesto que este proceso carece de complejidad, en el presente artículo se parte detalladamente de la consiguiente fase de alineación[8].

• ***Alineación***

Alinear consiste en "finding correspondences, in bilingual parallel corpora, between textual segments that are translation equivalents" (Kraif, 2002: 273). Se trata de una fase esencial de la creación del corpus paralelo, puesto que consiste en crear vínculos entre los textos origen (TO) y los textos meta (TM) que permiten establecer concordancias bilingües.

En la actualidad encontramos una amplia gama de programas de alineación, de pago (por ejemplo, *Stingray, Trados, Transit, LF Aligner, Bligner, Déjà Vu*) y gratuitos (*Youalign, Geometric Mapping and Alignment, Champollion Tool Kit, Uplug*). En nuestro caso, se ha alineado el corpus con el programa informático *Déjà Vu X2,* un programa de traducción automática y asistida compatible con Windows, que ofrece un asistente de alineación práctico y sencillo, que admite una gran variedad de formatos y que permite alinear varios pares de textos simultáneamente. La versión utilizada ha sido *Déjà Vu 8.0.611.*

[8] Una alternativa a la digitalización es la selección de textos que estén disponibles en formato electrónico en la Web. Por ejemplo, *Project Gutenberg* (http://www.gutenberg.org/) pone a disposición textos en formato digital.

a) Creación del archivo de alineación y de la memoria de traducción

El primer paso es crear un proyecto de alineación a través del asistente de alineación de *Déjà Vu*: *Menú Archivo>Nuevo>Proyecto de alineación*[9]. Tras introducir las lenguas de trabajo y cargar los textos, el programa importa toda la información y crea un bitexto, un texto que contiene información codificada para el propósito de almacenamiento y procesamiento en ordenadores y que presenta dos columnas; la columna de la izquierda se corresponde con el TO y la de la derecha, con el TM (Harris, 1988).

Tras este proceso de alineación automática resulta imperioso realizar una revisión manual. Cabe asegurarse de que todos los pares coinciden a partir de la lectura íntegra de los textos en el asistente de *Déjà Vu*, ya que la traducción no se realiza siempre de una manera predecible o lineal, sino que en ella se pueden producir varios fenómenos, tal y como indican Frankenberg-García y Santos (2003: 4): división de una oración del original en dos o más oraciones en la traducción, fusión de dos o más oraciones del original en la traducción, omisión de información, inserción de elementos no presentes en el texto origen, reordenación de elementos, entre otros. Esta revisión parece especialmente relevante en el caso de los textos literarios, ya que por sus características parecen otorgar más libertad al traductor y le permiten ser más creativo.

Finalmente, un aspecto clave en esta fase es la creación de la memoria de traducción, que no es más que "a database containing paired source text and translation segments, hence it is a type of parallel corpus" (Zanettin, 2012: 169). La alineación es, por tanto, el proceso mediante el cual se alimenta la memoria de traducción.

b) Exportación de la memoria de traducción

Una vez creada y guardada la memoria de traducción, para continuar con el proceso de compilación del corpus y poder etiquetarlo, ésta se debe exportar. Para ello, en *Déjà Vu*, se debe seguir primero la ruta *Menú Archivo> Abrir> memoria_de_traducción*, y seguidamente *Menú Archivo> Exportar> Datos externos.* En este proceso se deben elegir los siguientes parámetros: el formato del archivo exportado[10]; el tipo de campo; el idioma; el elemento delimitador entre TO y TM y la codificación para la exportación de la memoria de traducción[11]. De este modo disponemos finalmente de un corpus paralelo alineado.

[9] Durante el trabajo con *Déjà Vu* se aconseja desactivar cualquier programa antivirus o *firewall*, ya que puede haber interferencia entre ambos programas.

[10] Se aconseja seleccionar *Texto.*

[11] Europeo Occidental (Windows) (ANSI), Unicode (UTF-8), etc. En el caso de lenguas como el francés o el catalán, se recomienda seleccionar el formato de codificación Unicode (UTF-8), que sí reconoce los acentos graves, entre otros símbolos lingüísticos, por ejemplo. En caso de no seleccionar ninguna codificación, el programa selecciona la más adecuada en función de las lenguas de trabajo.

3.3. Etiquetado

Etiquetar un corpus consiste en dotarlo de información lingüística e interpretativa (Leech, 2004). Este apartado se enfoca en el etiquetado lingüístico y documental.

3.3.1. Etiquetado lingüístico

Si bien existen varios tipos de etiquetado lingüístico, como gramatical, léxico, fonético, semántico, pragmático, discursivo y estilístico (Leech, 2004), el presente trabajo se concentra en el etiquetado gramatical y léxico. El primero consiste en identificar la categoría gramatical a partir de etiquetas, y el segundo, en identificar el lema.

Existe una amplia gama de programas de etiquetado. Algunos ejemplos de pago son *Connexor Machinese Syntax* o *STILUS Core* y algunos ejemplos gratuitos son *TreeTagger* o *Freeling*. A la hora de elegir el programa de etiquetado, es recomendable buscar uno que incluya las lenguas que interesan al usuario. En la experiencia descrita en el presente artículo se utilizó la herramienta *TreeTagger,* desarrollada por Helmut Schmid en el *Institute for Computational Linguistics* de la Universität Stuttgart. Se trata de un programa que desempeña las funciones de etiquetador gramatical y léxico[12].

a) Segmentación y preparación de los textos

TreeTagger etiqueta por lengua, esto es, textos monolingües. Por tanto, antes de etiquetar morfológicamente los textos se debe transformar el archivo de alineación bilingüe producido por *Déjà Vu* (la memoria exportada) en dos archivos de texto monolingües. Este proceso no está exento de complejidad, ya que se debe recurrir a la línea de comandos[13]. Además, se precisa de un *script* en el lenguaje de programación Perl que convierta el archivo bilingüe alineado en dos archivos monolingües[14]. A continuación, se explican detalladamente los pasos que se deben seguir para segmentar los textos, como fase previa a la fase de etiquetado:

[12] *TreeTagger* (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) dispone de versiones para Linux, Mac y Windows y es compatible con las siguientes lenguas: inglés, francés, alemán, español, italiano, búlgaro, holandés, ruso, griego, portugués y chino.

[13] La línea de comandos es un accesorio del sistema operativo Windows que ocupa comandos de MS-DOS. Su acceso varía en función de la versión del sistema operativo de Windows, pero suele ubicarse en la carpeta "Accesorios". También se puede activar a partir de la introducción de las iniciales CMD en la opción "Ejecutar".

[14] En nuestro caso, hemos recurrido al *script segmentador.pl*, que además de segmentar los textos, añade los atributos estructurales <s> y </s> de inicio y fin de frase necesarios para indexar el corpus en una fase posterior. Se trata de un *script* implementado por el grupo de investigación COVALT de la Universitat Jaume I. Dicho *script* está disponible bajo petición.

- Descargar el programa *Strawberry Perl*[15].
- Copiar el *script* de Perl (*segmentador.pl*) y la memoria de traducción exportada (*ejemplo.txt*) al directorio *bin*, en *C:\cd strawberry\cd perl\bin*>.
- Activar la línea de comandos. De forma predeterminada aparecerá lo siguiente: *C:\Users:\Username.* Finalmente, para activar el *script* y segmentar el documento se tienen que insertar en la línea de comandos las instrucciones que se muestran a continuación (Fig. 2):

```
cd..
cd..
cd C:\strawberry\perl\bin
copy ejemplo.txt ejemplo.alg
perl segmentador.pl ejemplo
```

Figura 2. Segmentación del archivo bilingüe con el *script segmentador.pl.*

En este proceso se crean dos archivos, el archivo separado en la lengua origen (*ejemplo.to*) y el archivo separado en la lengua meta (*ejemplo.tm*). Estos archivos se encuentran en la carpeta *bin*, dentro del directorio *perl* y a la vez dentro de la carpeta *strawberry*, en el disco *C*. Por ejemplo, veamos el aspecto de los dos archivos separados tras esta fase (Fig. 3):

```
[…]
<s>
Er krabbelte unter dem Tisch hervor.
</s>
<s>
»Matilda? Emmanuel? Alles in Ordnung?«
</s>
[…]
```

```
[…]
<s>
Salió a gatas de debajo de la mesa.
</s>
<s>
–¿Matilda? ¿Emmanuel? ¿Estáis bien?–
</s>
[…]
```

Figura 3. Archivos segmentados y con los atributos estructurales de inicio y fin de frase.

[15] *Strawberry Perl* es una distribución del lenguaje de programación Perl para la interfaz de Windows (http://strawberryperl.com/releases.html).

b) Etiquetado gramatical y léxico

La siguiente fase es etiquetar los textos con *TreeTagger*. Se puede etiquetar bien cada texto de cada lengua individualmente, bien todos los textos de cada lengua de forma simultánea. El segundo procedimiento es más rápido y sobre todo aconsejable para corpus paralelos de gran tamaño, pero requiere, antes del etiquetado, fusionar en un solo documento todos los textos de cada lengua.

Una vez activado el *TreeTagger* se selecciona la lengua en la que se desea etiquetar, en el lateral izquierdo de la interfaz. Después, se carga el archivo a etiquetar (*Input file*) y se le asigna un nombre al archivo etiquetado (*Output file*). Para evitar que *TreeTagger* etiquete los atributos estructurales <s> y </s> cabe activar la opción *SGML tags present*. Finalmente, se activa el comando *Run*. *TreeTagger* crea un documento de texto plano por cada lengua con un aspecto como el que se muestra a continuación (Fig. 4).

```
<s>
Salió   VLfin   salir
a       PREP    a
gatas   NC      gata
de      PREP    de
debajo  ADV     debajo
de      PREP    de
la      ART     el
mesa    NC      mesa
.       FS      .
</s>
```

Figura 4. Archivo etiquetado con *TreeTagger.*

Se observa un texto con tres columnas separadas por tabulaciones. Se trata de los atributos posicionales *word, pos* y *lemma*, respectivamente. La primera columna representa las palabras o *tokens*; la segunda, la categoría gramatical a partir de etiquetas que podemos encontrar en los *tagsets* de *TreeTagger* para cada lengua[16]; y la tercera, el lema.

Finalmente, cabe comentar que *TreeTagger* presenta un pequeño margen de error en sus resultados, ya que no identifica el lema de algunas palabras y lo indica como "<unknown>". Éste y otros errores son comunes en la mayoría de los programas de este tipo: "Because of the complex and ambiguous nature of language, even a relatively simple annotation task such as POS-tagging can only be done

[16] Disponibles en: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

automatically with up to 95% to 98% accuracy" (Leech, 2004). Al respecto existen opiniones enfrentadas en torno a si una posterior revisión manual del etiquetado sería o no necesaria. Por ejemplo, Sinclair (1992) se manifiesta contrario a la postedición humana y prioriza un análisis automático en su integridad: "Analysis should be restricted to what the machine can do without human checking, or intervention" (1992: 381). Al contrario, Kahrel, Barnett y Leech (1997) opinan que "ultimately it is the human being's mental interpretation that enables us to evaluate the quality of annotation" (1997: 244). Nuestra opinión al respecto es que cada investigador deberá valorar si su corpus puede tener ese pequeño margen de error en función del tamaño principalmente, pero también en función de qué se estudie, cómo, en qué circunstancias, etc.

3.3.2. Etiquetado documental

El etiquetado documental proporciona información principalmente bibliográfica sobre cada uno de los textos del corpus. Se trata de un etiquetado XML (*eXtensible Markup Language*) con metadatos: "A common metalanguage by which electronic texts of all kinds can be stored, transmitted and displayed by different users" (Zanettin, 2012: 80). En otras palabras, se trata de atribuir un encabezamiento a cada texto, que será la marca de identidad para que la interfaz de indexación lo identifique y por tanto lo relacione con su texto paralelo alineado.

El encabezamiento puede contener cuanta información se desee. Su principal objeto es identificar el texto, por lo que los atributos más comunes son: autor, fecha de publicación, título, editorial, colección, etc. Este tipo de etiquetado suele añadirse de forma manual con un editor de XML. Con frecuencia el encabezamiento se añade después del etiquetado lingüístico, principalmente para evitar que el etiquetador lo etiquete, lo que también se puede evitar seleccionando la opción *SGML tags present* de *TreeTagger*, tal y como hemos comentado en el apartado anterior. En nuestro caso hemos recurrido al editor *Notepad*++[17]. Los metadatos elegidos para el encabezamiento de cada texto deberán añadirse junto con los atributos estructurales <text> y </text> de inicio y fin de texto. Veamos un ejemplo (Fig. 5):

[17] *Notepad*++ es un editor de texto gratuito con soporte para varios lenguajes de programación que tiene únicamente soporte para Windows (http://notepad-plus-plus.org). Otros editores de XML similares son *Oxygen* y *Jedit,* por ejemplo.

```
<text id="TM1" title="Lasoperayelcazo" language="de"
target_language="es" author="MichaelEnde" author_sex="M"
translator="RosanaTerzi" translator_sex="F" Publisher="SM"
year="1996" words="5.698">

[…]
<s>
Vacilante   ADJ     vacilante
continuó    VLfin   continuar
caminando   VLger   caminar
a           PREP    a
tientas     NC      tienta
.           FS      .
</s>
[…]
</text>
```

Figura 5. Archivo con etiquetado documental y lingüístico.

Tras estos pasos se dispone del corpus paralelo etiquetado lingüísticamente y documentado. Ahora bien, antes de poder explorarlo, es necesario indexarlo.

3.4. Indexación del corpus

Indexar un corpus significa "encode the corpus in a special binary format, write a registry file and create indexes for efficient access" (Evert, 2011) con la finalidad de analizarlo e interrogarlo a partir de búsquedas que permitan estudiar fenómenos y comportamientos lingüísticos específicos. Antes de indexar el corpus, se debe seleccionar la interfaz que lo albergará y que servirá para la indexación de éste.

En el presente trabajo hemos recurrido a la interfaz *IMS Open Corpus Workbench* (CWB) en su versión 3.0, un sofisticado conjunto de herramientas de análisis de corpus de código abierto diseñado por el *Institut für Maschinelle Sprachverarbeitung* de Stuttgart que permite explorar corpus etiquetados de gran tamaño. Su característica más sobresaliente es el *Corpus Query Processor* (CQP), un *software* de búsqueda y recuperación avanzada que permite realizar búsquedas muy complejas.

CWB posee un lenguaje de consulta específico y usa un formato de entrada en forma vertical. Por ejemplo, usa archivos de texto en los que cada *token* se encuentra al principio de una nueva línea, seguido por la categoría gramatical y la lematización correspondientes en forma tabular, tal y como se muestra en el siguiente ejemplo (Fig. 6):

```
<corpus>

<text id="TM1" title="Lasoperayelcazo" language="de"
target_language="es" author="MichaelEnde" author_sex="M"
translator="RosanaTerzi" translator_sex="F" Publisher="SM"
year="1996" words="5.698">

[…]
<s>
Vacilante   ADJ     vacilante
continuó    VLfin   continuar
caminando   VLger   caminar
a           PREP    a
tientas     NC      tienta
.           FS      .
</s>
[…]
</text>

</corpus>
```

Figura 6. Estructura de los corpus indexados con *IMS Open Corpus Workbench.*

Como se desprende de la imagen, además de los atributos estructurales <s>, </s>, <text> y </text> que ya se han visto en apartados anteriores de este artículo, CWB precisa de un tercer atributo estructural que indica inicio y fin del corpus, es decir, <corpus> y </corpus>. En el apartado 4.1.2 se indica en qué momento del proceso se debe introducir dicho atributo estructural.

Existen varias opciones a la hora de indexar un corpus con CWB[18]. En esta sección se explica con detalle el proceso de indexación de un corpus paralelo con CWB para un ordenador con el sistema operativo de Windows.

La distribución de CWB para Windows se encuentra en pruebas *beta,* por lo que puede no presentar la misma estabilidad que tiene en las versiones definitivas de otros sistemas operativos como Linux, Mac o Solaris. Por tanto, hasta que exista una distribución definitiva de CWB para Windows, este artículo presenta una alternativa para usuarios de este sistema operativo que deseen indexar su corpus

[18] Por ejemplo: desde la web, desde un ordenador o un servidor con un sistema operativo compatible con CWB (como Linux, Mac o Solaris, cuyas versiones de CWB ya son definitivas), desde un ordenador con Windows (cuya versión de CWB está en pruebas *beta*), etc. La selección de una u otra vía estará condicionada fundamentalmente por los conocimientos del usuario en lenguaje de programación Unix (un sistema operativo portable, multitarea y multiusuario desarrollado en 1969) y por el sistema operativo desde el que éste trabaje. En el sitio web *IMS Open Corpus Workbench* (http://cwb.sourceforge.net) encontramos información detallada al respecto.

con CWB y contar con la estabilidad de una versión definitiva. Dicha alternativa consiste en acceder a las funcionalidades de CWB a través de la conexión a un servidor compatible con CWB. En nuestro caso hemos recurrido a un servidor Linux[19]. A continuación, se explica este proceso detalladamente[20].

3.4.1. Pasos previos a la indexación

En este apartado se da cuenta de una serie de requisitos técnicos y de organización que cabe tener presentes antes de acceder a CWB.

a) Requisitos técnicos

En primer lugar, es imprescindible acceder al servidor Linux mediante un programa de transferencia de archivos. Cabe tener datos de acceso como usuario administrador del servidor: nombre del servidor, nombre de usuario, contraseña, protocolo de transferencia de archivos y número de puerto. En el presente caso se accedió a un servidor Linux a través del programa de código abierto *WinSCP*[21]. Tras la instalación y la validación de los datos de usuario administrador, aparece una ventana cuya parte izquierda muestra la carpeta local del ordenador (la de Windows), y cuya parte derecha muestra la cuenta del servidor (la de Linux).

Seguidamente, es necesario otorgar permisos a la información que se desea cargar al servidor, es decir, al corpus. Para ello, se recurrió a la instalación del programa *PuTTY* y a la validación con los datos de acceso al servidor[22]. Cabe comentar que al escribir la contraseña en *PuTTY*, ésta no se visualiza; además el ratón se bloquea, por lo que tras escribirla es necesario insertar un salto de línea a modo de activación.

Una vez instalados estos programas, es necesario instalar en el servidor la distribución de CWB correspondiente al sistema operativo del servidor (en nuestro caso, Linux) y la interfaz de lenguaje de Perl[23].

[19] Si bien las universidades y los centros de investigación suelen dar acceso a este tipo de servicios a sus usuarios, con una sencilla búsqueda en la web, encontramos múltiples proveedores de servidores Linux o de otros sistemas operativos compatibles con CWB.

[20] CWB cuenta con una lista de correo de la Università di Bologna en la que se resuelven al instante dudas acerca de este programa y sus funcionalidades (http://devel.sslmit.unibo.it/mailman/listinfo/cwb).

[21] *WinSCP* es un programa de transferencia de archivos (http://winscp.net/eng/download.php). Otro programa similar es *FileZilla*.

[22] PuTTY es un emulador de terminal, un programa que permite conectar con máquinas y ejecutar programas remotamente. PuTTY se conecta como cliente a múltiples protocolos, como SSH, Telnet o Rlogin (http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html).

[23] Los paquetes de instalación se encuentran en: http://cwb.sourceforge.net/download.php. En cuanto al proceso de instalación hay información detallada en: http://cwb.sourceforge.net/install.php. No obstante, a continuación se explica sucintamente cómo se instalaría la distribución de

b) Preparación de la carpeta raíz y de los textos

Antes de acceder a CWB es necesario preparar los textos o corpus para su codificación e indexación. Los sistemas Unix tienen una estructura de árbol, donde hay una carpeta raíz desde la que se ramifica el resto de carpetas. Para indexar el corpus se deberá trabajar siempre dentro de la carpeta raíz, que será la que albergará el corpus. Por tanto, en primer lugar se debe crear una carpeta en el servidor: por ejemplo, *corpus,* en *home/user_name*. Una vez creada, se accederá a ella a partir de *PuTTY* con el comando *cd,* es decir, con la indicación de la ruta de la carpeta: *cd corpus*[24].

Indistintamente del número de textos que contenga el corpus paralelo, éste sólo debe contener dos documentos de texto: un documento con los textos originales y otro documento con los textos meta. Por ello, es necesario unir todos los TO y TM en dos documentos. Para ello, dentro de la carpeta *corpus* deben crearse dos subcarpetas y desde la carpeta local se arrastran los textos originales (*original1.txt, original2.txt,* etc.) a una carpeta y los textos traducidos (*traducido1.txt, traducido2.txt,* etc.) a otra. Tras ello hay que desplazarse al interior de cada una de las subcarpetas con el comando *cd*, tal y como se ha visto en el caso anterior. Una vez dentro de cada subcarpeta, para fusionar los textos se puede seguir el siguiente atajo: *cat *.txt > original.txt y cat *.txt> traducido.txt*[25]. De este modo, el resultado será un documento en cada lengua del corpus que contendrá todos los textos etiquetados en esa lengua. En este artículo, a efectos prácticos, se denominarán *original.txt* y *traducido.txt.* A continuación, se aconseja reubicar estos documentos en la carpeta raíz *corpus*, en el primer nivel, simplemente por cuestiones de practicidad.

Una vez se dispone de dos documentos finales, cabe editarlos e insertar manualmente las etiquetas de inicio y cierre del corpus (<corpus> y </corpus>) para que la estructura del corpus se corresponda finalmente con la que puede interpretar el CWB, tal y como se ha mostrado previamente (Fig. 6).

CWB en un servidor Linux, lo que es extensible también a la instalación de la interfaz del lenguaje de Perl. En primer lugar, debe descargarse al ordenador Windows el paquete instalador correspondiente al sistema operativo del servidor, en el caso en cuestión, Linux (*cwb-3.0.0-linux-i386.tar.gz*), y seguidamente debe cargarse al servidor con un programa que implemente la transferencia de archivos vía SSH, por ejemplo, *WinSCP.* A continuación se debe acceder con *PuTTY* al servidor y se debe descomprimir el paquete instalador con el comando de descompresión *tar* (*tar zxvf cwb-3.0.0-linux-i386.tar.gz*). Después hay que ejecutar el instalador, esto es, en primer lugar cabe entrar en la carpeta adecuada *($ cd cwb-3.0.0-linux-i386/)* y seguidamente instalar con el comando *sudo* el programa instalador (*sudo ./install-cwb.sh*), que siempre tiene .sh como extensión. Al final del proceso el sistema requiere la contraseña del servidor.

[24] El comando *cd* significa 'change directory'.

[25] Este comando fusiona todos los archivos con extensión .txt. El asterisco indica que pueden tener cualquier nombre y que se ordenarán alfabéticamente.

3.4.2. Proceso de indexación

Para indexar el corpus con CWB es necesario crear o disponer de una herramienta que tenga la función de indexar el corpus con CWB[26]. Debido a la complejidad de indexación con dicha herramienta, en nuestro caso se usó un *wrapper* de *Perl* del *script* original de indexación de CWB. Los pasos que se describen a continuación se corresponden con dicho *script* [27].

En primer lugar, el script *cwbify_standard_corpus.pl* se debe arrastrar o copiar a la carpeta raíz, *corpus*. A partir de este momento se trabajará por línea de comandos, esto es, con *PuTTY*. La Tabla II indica exactamente qué comandos se deben introducir para continuar el proceso de indexación:

Tabla II. Activación del *script* de indexación.

1. Activar el *script* de Perl con *pico* (*Pine Composer*), un editor de textos para Unix y sistemas basados en Unix	`pico cwbify_standard_corpus.pl;`
2. Desplazarse hacia abajo con la flecha del teclado y comprobar los atributos del encabezamiento. El atributo *id* es el atributo que viene por defecto en CWB. En caso de querer añadir más atributos cabría teclearlos (todo en la misma línea, sin espacios). Los atributos deben escribirse en inglés y no se permiten espacios en el nombre de un único atributo	`s_attributes("text:0+id+title+language +target_language+author+author_sex +translator+translator_sex+publisher +year+words");`
3. Cerrar *pico*	`control+x;`
4. Guardar los datos con los comandos que proporciona la ventana, con el teclado (Y = yes; N= no)	`y/n`
5. Una vez el *script* contiene los metadatos del encabezamiento y una vez cerrado *pico*, es necesario hacer ejecutable el *script*	`chmod +x cwbify_standard_corpus.pl;`

El siguiente paso es indexar cada uno de los corpus, *original.txt* y *traducido.txt,* por separado. En primer lugar, es necesario indicar la lengua del corpus (*langcode: -l*), el directorio donde se guardará (*dir: -d*), el nombre (*cwbname: -c*) y la

[26] La herramienta original de indexación de CWB es *cwb-encode* y viene por defecto con la distribución estándar del *Corpus Workbench*. En las siguientes direcciones se pueden seguir los pasos de indexación con esta herramienta: http://cwb.sourceforge.net/files/CWB_Encoding_Tutorial/CWB_Encoding_Tutorial.html y http://cwb.sslmit.unibo.it/doku.php?id=users:indexing_a_corpus.

[27] Se trata del *script cwbify_standard_corpus.pl*, creado por Marco Baroni, de la Università degli Studi di Trento y Adriano Ferraresi, de la Università di Bologna. Dicho *script* está disponible bajo petición.

descripción (*descname: -n*), así como indicar en qué archivo se encuentra[28]. A continuación se presentan los comandos que se debe introducir[29]. Para ello, en primer lugar, con el fin de indexar el corpus original debe introducirse en PuTTY, lo que se indica a continuación (Fig. 7)[30]:

```
./cwbify_standard_corpus.pl -l de
-d /home/user_name/corpus/corpus_data_DE
-c DE_CORPUS -n "The German source texts"
original.txt
```

Figura 7. Comandos de indexación (para los textos originales).

Seguidamente, para indexar el corpus traducido se debe repetir este procedimiento y adaptar los parámetros mencionados en el párrafo anterior a las características del corpus en cuestión. De este modo, ya se dispone de los textos originales y traducidos indexados. El siguiente paso es alinearlos, tal y como se indica en la Tabla III. Esta acción se realiza dos veces, una por cada lengua:

Tabla III. Alineación de los corpus indexados.

1. Alinear el TO con el TM (obsérvese que en este caso el nombre de los corpus se escribe en minúscula.)	`cwb-align de_corpus es_corpus s`
2. Crear el archivo de registro del corpus con *pico*	`pico /usr/local/share/cwb/registry/de_corpus`
3. Desplazarse hasta el final del documento con la flecha del teclado y añadir el comando de la columna de la derecha, lo que significa que el *de_corpus* estará alineado con el *es_corpus*	`ALIGNED es_corpus`

[28] La lengua del corpus se debe indicar en formato ISO en minúsculas (en, de, it, etc.); el nombre del corpus se debe escribir íntegramente en mayúsculas; las características del corpus se deben escribir entre comillas.

[29] Obsérvese que estos comandos se han adaptado a la indexación de un corpus paralelo alemán-español, por tanto, dicha información deberá adaptarse a la combinación lingüística del corpus que se desee indexar según los parámetros que explica la anterior nota al pie, al igual que *user_name* deberá ser sustituido por el nombre real del usuario del servidor o, en general toda la ruta *(/home/user_name/corpus*) por la ruta real correspondiente.

[30] Todo en la misma línea, separado por espacios.

Continuación Tabla III.

4. Tras crear el archivo de registro, se cierra *pico*	`control+x`
5. Guardar los cambios (Y= yes; N= no)	`y/n`
6. Codificar la alineación. Se puede comprobar con el comando *ls* (list) si se ha creado el archivo *out.align* en la carpeta raíz	`cwb-align-encode -D out.align`

Una vez alineado el corpus paralelo de la lengua original (alemán) a la lengua meta (español), es necesario alinear el corpus en la otra dirección, es decir, de la lengua meta (español) a la lengua original (alemán). Para ello basta con repetir las instrucciones de la Tabla III e intercambiar los nombres de los corpus, esto es, en vez de *de_corpus*, *es_corpus* y viceversa. Finalmente, se dispone de un corpus paralelo etiquetado indexado con CWB[31]. Seguidamente, ya es posible explorarlo con *Corpus Query Processor.*

3.5. Exploración del corpus paralelo: búsquedas con *Corpus Query Processor*

Como ya se comentó en párrafos anteriores, CQP es uno de los principales componentes de CWB, y su característica más importante es que permite realizar búsquedas de carácter muy complejo[32]. Sus principales ventajas son: la integración de un número ilimitado de etiquetas a nivel de palabra, metadatos y etiquetas estructurales (en forma de etiquetas XML de inicio y cierre) en sus búsquedas y la capacidad de realizar búsquedas muy generales en corpus grandes y gestionar de forma eficaz mucha información, que se guarda en macrolibrerías para una recuperación posterior (Baroni, 2005; Hoffmann y Evert, 2006).

A continuación se indican algunos comandos de **búsqueda** que se deben introducir en *PuTTY* y que, además de servir como práctica para CQP, ayudarán a comprobar si el corpus se ha indexado correctamente (Tabla IV). No obstante, para interrogar el corpus con CQP se recomienda estudiar el manual de usuario de CQP[33], puesto que "a corpus is still only a resource, and will only show us what we are capable of finding" (Saldanha, 2009: 3).

[31] Una vez superado este proceso, la siguiente fase es estudiar las opciones existentes de cara al archivo y la distribución del corpus. Al respecto Wynne (2004) propone una serie de pautas que constituyen una hoja de ruta idónea.

[32] Otras herramientas similares son *WordSmith Tools* (Scott, 1999), *Monoconc Pro* (2000) y *Word Sketch Engine* (Kilgarriff, Rychly, Smrz y Tugwell, 2004).

[33] Disponible en: http://cwb.sourceforge.net/files/CQP_Tutorial.pdf.

Tabla IV. Principales comandos de búsqueda de *Corpus Query Processor.*

Activar CQP	`cqp –e;`	CQP se puede activar a través de PuTTY.
Visualizar los corpus disponibles en el servidor e indicar el corpus que se desea consultar	`show corpora;` `DE_CORPUS;`	Tras introducir el comando *show corpora* se visualizan todos los corpus disponibles en el servidor. Seguidamente se debe escribir el nombre del corpus que se desea consultar.
Visualizar el corpus con su corpus paralelo alineado	`show +es_corpus;` `set C es_corpus;`	De forma predeterminada CQP mostrará el corpus sin alinear.

Cambiar el tamaño de visualización del contexto	`set Context 50;`
	`set Context 10 words;`
	`set Context s;`
		Si sólo se especifica un número, se refiere a caracteres; words, a tokens; s, a oraciones.

Buscar por palabra, por categoría gramatical o por lema	`[word = "ging"];`
	`[pos = "VVINF"];`
	`[lem = "kriechen"];`
		La primera búsqueda volcará únicamente los datos correspondientes a la forma *ging* del verbo *gehen*. La segunda búsqueda dará como resultado todos los verbos (excepto modales y auxiliares) que aparecen en la forma de infinitivo. Para buscar por categoría gramatical cabe tener en cuenta el tagset de cada lengua. La tercera búsqueda volcará todas las formas del verbo *kriechen*. En función de la versión de CQP, la abreviación para lema puede ser "lem" o "lemma".

Contar por lema o por palabra	`count by lem;` `count by word;`	La primera búsqueda mostrará la frecuencia de cada lema (si no se especifica ninguna búsqueda, el sistema entiende que es sobre la última búsqueda realizada). La segunda búsqueda mostrará la frecuencia de cada palabra.

Nombrar y guardar una búsqueda	`Klettern = [lem = "klettern"];`
	`show named;`
	`cat Klettern;`
	`set DataDirectory ".";`
	`show named;`
	`save Klettern;`
	`set PrintOptions hdr;`
	`cat Klettern > "Klettern";`
	`cat Klettern > "klettern.txt";`
		La búsqueda se puede guardar en *home* o en la carpeta específica del corpus, por ejemplo *corpus*. Esto dependerá de si accedemos a CQP desde *home* del administrador, o desde otra carpeta.
		Klettern = Nombre que se le da a la búsqueda.
		show named = Para ver los nombres de todas las búsquedas.
		cat Klettern = Para visualizar la búsqueda.
		set PrintOptions hdr = Para mostrar los términos de búsqueda en el encabezamiento del documento de búsqueda.

Buscar una secuencia	`[lem = "ir"][] "hacia";` `[lem = "ir"][] {2} "hacia";`	Por ejemplo, "fue brincando hacia" (1 palabra) o "fue dando tumbos hacia" (2 palabras).
Interrumpir búsqueda	`q`	Las búsquedas se pueden interrumpir en cualquier momento.

4. RESULTADOS

Existen varios patrones lingüísticos que pueden expresar manera de desplazamiento en alemán. No obstante, debido a la necesidad de delimitar el análisis de datos, se consideró el verbo alemán de manera de desplazamiento como unidad de búsqueda en la parte alemana del corpus paralelo y consiguientemente el evento que lo incluye como unidad de análisis. Si bien es cierto que tomar como unidad de búsqueda el verbo de manera de desplazamiento no representa toda la expre-

sión de la manera de desplazamiento existente en la parte alemana del corpus, en concreto, y en la lengua alemana, en general, sí da cuenta de una parte muy significativa de este fenómeno, al ser éste el patrón lingüístico predominante para expresar dicho fenómeno.

Tras extraer los verbos del subcorpus alemán ([pos = "VV.*"] y del subcorpus español ([pos = "VL.*"])[34], se ordenaron por lema en función de la frecuencia (count by lem). Consiguientemente, se realizó una lectura manual de dicho listado de frecuencias para extraer los verbos que expresan manera de desplazamiento en cada lengua y se constató que el subcorpus alemán contiene 86 verbos de manera de desplazamiento (lemas) y el subcorpus español, 50. Estos resultados corroboran las afirmaciones de McNeill (2000) y Özcalışkan y Slobin (2003) en cuanto a que las lenguas de marco verbal tienen menor riqueza léxica, esto es, un repertorio verbal más reducido[35].

Para extraer la base empírica objeto de estudio, es decir, los eventos de manera de desplazamiento en alemán y sus traducciones al español, se extrajeron las concordancias[36] de los 86 verbos alemanes de manera de desplazamiento (por ejemplo: [lem = ".*klettern"][37]) alineadas con sus respectivas traducciones y se guardó cada una de las búsquedas en un documento de texto para una posterior exploración. Finalmente, se analizaron las concordancias y se establecieron conclusiones en cuanto a las técnicas de traducción de la manera de desplazamiento del alemán al español observadas en el corpus paralelo (Molés-Cases, 2015).

5. CONCLUSIONES

Como se ha visto hasta aquí, el proceso de alineación, etiquetado, indexación y exploración de un corpus paralelo no es una empresa fácil. No obstante, debido a la importancia de la lingüística de corpus para la investigación en Lingüística y en Traducción cualquier dificultad en el proceso de creación de corpus debería asumirse como menor y transitoria, ya que los resultados volcados al final del proceso permiten un análisis minucioso y fiable de un gran número de fenómenos. Los

[34] Los símbolos ".*" funcionan a modo de *wild card* o comodín. Estas expresiones regulares ordenan a CQP buscar todos los verbos en alemán y en español respectivamente en cualquier forma verbal. Las etiquetas específicas dependen del *tagset* de cada lengua. Según el *tagset* de *TreeTagger* para alemán, "VV" indica cualquier verbo, excepto modales y auxiliares. Según el *tagset* de *TreeTagger* para español, "VL" indica cualquier verbo, excepto los verbos *ser, estar, tener, haber* y los verbos modales.

[35] Cabe tener en cuenta que esta diferencia en la riqueza del lexicón verbal se refiere a la manera de desplazamiento y que en español puede compensarse a través del contexto u otros recursos, como: adverbios, adjetivos, etc.

[36] Una concordancia es "a collection of the occurrences of a word-form, each in its own textual environment" (Sinclair, 1991: 32).

[37] Este comando ordena a CQP buscar las concordancias de todas las formas verbales (incluso aquellas que contengan prefijos) del lema *klettern*.

procesos de alineación y etiquetado destacan por su larga duración, no tanto por su complejidad. No obstante, la principal dificultad de este proceso reside en la indexación del corpus.

Por una parte, el proceso de indexación con CWB puede resultar laborioso, ya que requiere una adquisición paulatina de conocimientos técnicos de programación a usuarios generalmente sin experiencia y sin habilidades técnicas previas, así como el estudio del manual de búsquedas de CQP y la práctica de las múltiples funciones que éste ofrece. Además, hay que resaltar la sensibilidad de CWB, esto es, la mínima diferencia con respecto a las instrucciones interrumpiría el proceso, lo que conllevaría una gran pérdida de tiempo.

En contrapartida, la única habilidad técnica que se requiere aprender es la adquisición de conocimientos de Unix, y pese a su dificultad, debe considerarse como una fase transitoria y necesaria para la gestión y el análisis del corpus. "Fortunately (and contrary to many people's fears), programming for linguistic research questions does not require a special aptitude in computer science or mathematics. The basic tool you need is knowledge of a language that a computer understands" (Biber, Conrad y Reppen, 1998: 256). Así, a pesar del esfuerzo inicial, es muy recomendable indexar un corpus con CWB, puesto que éste ofrece datos exactos acordes con la rigurosidad de su proceso de indexación. Se trata de un programa de gestión de corpus muy útil, veloz, flexible y de gran capacidad que además incluye un procesador capaz de realizar búsquedas muy complejas y específicas, lo que permite el estudio de fenómenos y comportamientos lingüísticos y traductológicos de forma minuciosa y flexible y, a su vez, ofrece un amplio abanico de posibilidades de análisis. Al final, con la práctica se puede adquirir una actitud mecánica y personalizada hasta cierto punto, y es aquí donde el esfuerzo inicial tiene su recompensa.

REFERENCIAS

Aston, Guy. (2009). Foreword. En Allison Beeby, Patricia Rodríguez-Inés y Pilar Sánchez-Gijón (Eds.). *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam/Philadelphia: John Benjamins, x.

Baker, Mona. (1993). Corpus Linguistics and Translation Studies – Implications and Applications. En Mona Baker, Gill Francis y Elena Tognini-Bonelli (Eds.). *Text and Technology: In honour of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 233-252.

————————. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, 7(2), 223-243.

Baroni, Marco. (2005). Esplorare un corpus con CWB. Disponible en: http://sslmit.unibo.it/~baroni/termsett/05_1/cwb_cqp.pdf. [Consulta: 12/06/2012].

Beeby, A., Rodríguez, P., Sánchez-Gijón, P. (Eds.). (2009). *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam/Philadelphia: John Benjamins.

Bernardini, Silvia. (2004). Corpora in the classroom: An overview and some reflections on future developments. En John M. Sinclair (Ed.). *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins, 15-36.

Biber, Douglas, Susan Conrad y Randi Reppen. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, Lynne, Jennifer Pearson. (2002). *Working with Specialized Language. A practical guide to using corpora*. Londres: Routledge.

Braun, Sabine. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, 17(1), 47-64.

Christ, Oliver, Bruno M. Schulze, Anja Hofmann y Esther König. (1999). *The IMS Corpus Workbench. Corpus Query Processor. User's Manual*. Stuttgart: University of Stuttgart. Disponible en http://corpora.dslo.unibo.it/TCORIS/cqpman.pdf. [Consulta: 01/03/2012].

Corpas Pastor, Gloria (2008). *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfurt am Main: Peter Lang.

Evert, Stefan y OCWB (2005). The CQP Query Language Tutorial. Disponible en: http://cwb.sourceforge.net/files/CQP_Tutorial.pdf. [Consulta: 28/08/2012].

————————. (2011). Index a corpus with CWB. Disponible en: http://cwb.sslmit.unibo.it/doku.php?id=users:indexing_a_corpus. [Consulta: 02/04/2012].

Filipović, Luna. (2007). *Talking about Motion: A Crosslinguistic Investigation of Lexicalization Patterns.* Amsterdam/Philadelphia: John Benjamins.

Frankenberg-García, Ana. (2007). Compara: Optical Character Recognition (OCR) editing and preliminary markup instructions. Disponible en: http://comum.rcaap.pt/bitstream/123456789/334/1/OCR.pdf. [Consulta: 15/11/2012].

Frankenberg-García, Ana y Diana Santos. (2003). Introducing COMPARA, the Portuguese-English parallel corpus. En Federico Zanettin, Silvia Bernardini y Dan Stewart (Eds.). *Corpora in Translator Education*. Manchester: St. Jerome, 71-87.

Harris, Brian. (1988). Bi-text, a New Concept in Translation Theory. *Language Monthly* 54, 8-11.

Hoffmann, Sebastian y Stefan Evert. (2006). BNCweb (CQP edition) - the marriage of two corpus tools. En Sabine Braun, Kurt Kohn y Joybrato Mukherjee (Eds.). *Corpus technology and language pedagogy? New resources, new tools, new methods.* Frankfurt am Main: Peter Lang, 177-195.

Hunston, Susan. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kahrel, Peter, Ruthanna Barnett y Geoffrey Leech. (1997). Towards cross-linguistic standards or guidelines for the annotation of corpora. En Roger Garside, Geoffrey Leech y Anthony McEnery (Eds.): *Corpus Annotation. Linguistic Information from Computer Text Corpora*. Nueva York: Longman, 231-242.

Kennedy, Graeme (1998). *An Introduction to Corpus Linguistics*. London/New York: Longman.

Kenning, Marie-Madeleine. (2010). What are parallel and comparable corpora and who can we use them? En Anne O'Keeffee y Michael McCarthy (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge: Nueva York. 487-500.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). *The Sketch Engine*. Eleventh EURALEX International Congress, Lorient, Francia.

Kraif, Olivier. (2002). Translation Alignment and Lexical Correspondence. En Bengt Altenberg y Sylviane Granger (Eds.). *Lexis in contrast. Corpus-based Approach*. Amsterdam/Philadelphia: John Benjamins, 271-90.

Kübler, Natalie. (2011). Working with different corpora in translation teaching. En Ana Frankenberg-García, Lynne Flowerdew y Guy Aston (Eds.). *New Trends in Corpora and Language Learning*. London: Continuum, 62-80.

Kübler, Natalie y Pierre-Yves Foucou. (2003). Teaching English verbs with bilingual corpora: Examples in the field of computer science. En Sylviane Granger, Jacques Lerot y Stephanie Petch-Tyson (Eds.). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam: Rodopi, 185-206.

Laroche, Audrey y Philippe Langlais. (2010). Revisiting Context-based Projection Methods for Term-Translation Spotting in Comparable Corpora. XXIII International Conference on Computational Linguistics (COLING), Beijing, China. Stroudsburg: Association for Computational Linguistics, 617-625.

Laviosa, Sara. (1998). *The corpus-based approach: a new paradigm in Translation Studies*. Meta 43 (4), 474-479.

————————. (2002). *Corpus-based Translation Studies*. Amsterdam/New York: Rodopi.

Leech, Geoffrey. (2004). Adding Linguistic Annotation. En Martin Wynne (Ed.). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxford Books. Disponible en: http://www.ahds.ac.uk/__print__/guides/linguistic-corpora/chapter2.htm. [Consulta: 01/07/2013].

Mauranen, Anna y Pekka Kujamäki. (2004). *Translation Universals. Do they exist?* Amsterdam/Philadelphia: John Benjamins.

McEnery, Tony y Costas Gabrielatos. (2006). English corpus linguistics. En Bas Aarts y April McMahon (Eds.). *The Handbook of English Linguistics*. Oxford: Blackwell, 33-71.

McEnery, Tony; Richard Xiao y Yukio Tono. (2006). *Corpus-based Language Studies. An advanced resource book.* London/New York: Routledge.

McNeill, David. (2000). Analogic/Analytic representations and cross-linguistic differences in thinking for speaking. *Cognitive Linguistics* 11, 43-60.

Meyer, Charles F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Molés-Cases, T. (2015). La 'saliencia' de la manera en los eventos de desplazamiento. Propuesta de técnicas de traducción. En Iraide Ibarretxe-Antuñano y Alberto Hijazo-Gascón (Eds.). *New horizons in the study of motion: bringing together applied and theoretical perspectives*. Cambridge: Cambridge Scholars.

O'Keeffe, Anne y Michael McCarthy. (2010). *The Routledge Handbook of Corpus Linguistics*. Routledge: Nueva York.

Olohan, Maeve (2004). *Introducing Corpora in Translation Studies*. New York: Routledge.

Özcalışkan, Şeyda y Dan Slobin. (2003). Codability Effects on the Expression of Manner of Motion in Turkish and English. En A. Sumru Özsoy, Didar Akar, Mine Nakipoğlu-Demiralp, E. Eser Erguvanlı-Taylan y Ayhan Aksu-Koç (Eds.). *Studies in Turkish linguistics*. Estambul: Boğaziçi University Press, 259-270.

Saldanha, Gabriela. (2009). Principles of corpus linguistics and their application to translation studies research. *Tradumàtica* 7, 1-7.

Scott, Mike. (1999). *WordSmith Tools*. Oxford: OUP.

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

Sinclair, John M. (1992). The automatic analysis of corpora. En Jan Starvik (Ed.). *Directions in Corpus Linguistics. Proceedings of Nobel Symposium*. Berlín/Nueva York: Mouton de Gruyter, 379-397.

Slobin, Dan. (2006). What makes manner of motion salient? Explorations in linguistic typology, discourse, and cognition. En Maya Hickmann y Stephane Robert (Eds.). *Space in Languages: Linguistic Systems and Cognitive Categories*. Amsterdam: John Benjamins, 59-81.

Talmy, Leonard. (1975). Semantics and syntax of motion. En John P. Kimball (Ed.). *Syntax and Semantics*. Nueva York: Academic Press, 181-238.

————————. (1985). Lexicalization patterns: Semantic structures in lexical forms. *Language typology and syntactic description*, 3(99), 36-149.

————————. (1991). Path to realization: A typology of event conflation. En *Proceedings of the Seventeenth Annual Meeting of Berkeley Linguistics Society*. Berkeley, 480-519.

Wynne, Martin. (2004). Archiving, distribution and preservation. En Martin Wynne (Ed.). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxford Books. Disponible en: http://www.ahds.ac.uk/__print__/guides/lin-

guistic-corpora/chapter6.htm. [Consulta: 13/07/2013].
Zanettin, Federico. (2012). *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome.
Zanettin, Federico; Silvia Bernardini y Dominic Steward. (2003). *Corpora in translator education*, Manchester/Northampton, MA: St. Jerome.