ECPC: el discurso parlamentario europeo desde la perspectiva de los estudios traductológicos de corpus*

José Manuel Martínez Martínez
Universität des Saarlandes
j.martinez@mx.uni-saarland.de

Iris Serrat Roozen
Universidad de Valencia
iris.serrat@uv.es

Resumen

Este artículo presenta la labor investigadora del grupo ECPC, que ha diseñado y creado un Archivo de discursos parlamentarios europeos con el fin de estudiar dicho género y la hipotética influencia de la traducción en la construcción de la identidad europea. La investigación se ha restringido al Parlamento Europeo (mediante la construcción de un corpus paralelo —EN y ES— con las versiones en inglés y español) y a dos parlamentos nacionales, la House of Commons británica (HC) y el Congreso de los Diputados español CD), que constituyen sendos corpus comparables. El Archivo contiene los discursos recogidos en las actas de las sesiones plenarias celebradas a lo largo de la VI legislatura del Parlamento Europeo (2004-2009) en cada una de las cámaras anteriormente mencionadas.

Palabras clave

Estudios Traductológicos de Corpus, metodología, discurso político, debate parlamentario, Parlamento Europeo, House of Commons, Congreso de los Diputados, Análisis Crítico del Discurso, corpus comparable, corpus paralelo

Abstract

This paper presents the main outcome of the ECPC research group: an archive of European parliamentary speeches created to study this genre and the hypothetical influence of translation in the construction of European identity. The archive is made up of, on the one hand, a parallel corpus containing the English and Spanish versions of the European Parliament proceedings, and on the other hand, two comparable corpora —one containing the proceedings of the House of Commons for English and the proceedings of the Congreso de los Diputados for Spanish. The archive contains the speeches delivered in the plenary sittings held during the 6th term of the European Parliament (2004-2009) before each of the above mentioned Houses.

Keywords

Corpus-based Translation Studies, methodology, political discourse, parliamentary debate, European Parliament, House of Commons, Congreso de los Diputados, Critical Discourse Analysis, comparable corpus, parallel corpus

1 Grupo ECPC

El grupo de investigación ECPC[1] (European Parliamentary Comparable and Parallel Corpora) inició oficialmente su andadura en el año 2005, cuando recibió financiación por parte del Ministerio de Educación y Ciencia para el proyecto Corpus comparables y paralelos de discursos parlamentarios europeos, con referencia HUM 2005-03756/FILO.

En la actualidad, el grupo está consolidando su actividad investigadora gracias a la financiación del Ministerio de Ciencia e Innovación bajo el nuevo título Ampliación y profundización de ECPC y de ConcECPC 1.0: avances teórico-descriptivos e innovaciones tecnológicas, con referencia FFI2008-01610/FILO.

Este proyecto se enmarca dentro de los Estudios Traductológicos de Corpus, de manera que la investigación en el seno del grupo se realiza a partir de corpus comparables y paralelos los cuales están compuestos por discursos del Parlamento Europeo (EN para la versión en inglés y ES para la versión en español), así como de los parlamentos nacionales de España (Congreso de los Diputados – CD) y del Reino Unido (Cámara de los Comunes/House of Commons – HC).

*El Archivo ECPC se ha creado en el marco del proyecto Ampliación y Profundización de ECPC y de ConcECPC 1.0 (FFI2008-01610/FILO) financiado por el Ministerio de Ciencia e Innovación durante los años 2009-2011.

[1]Web con información acerca de ECPC: http://www.ecpc.uji.es

2 Antecedentes

2.1 Lingüística de corpus (LC)

El proyecto ECPC es heredero del trabajo que iniciaran en el ámbito de la Lingüística de Corpus expertos como Henry Kučera y W. Nelson Francis (Brown University, Providence, Rhode Island), quienes compilaron por primera vez un corpus electrónico, el Brown Corpus, formado por un millón de palabras de inglés estadounidense. También su homólogo británico, el Lancaster-Oslo/Bergen Corpus (LOB), creado por Geoffrey Leech (Lancaster University), Stig Johansson (Universitetet i Oslo) y Knut Hofland (Universitetet i Bergen) es un referente ineludible.

2.2 Estudios Traductológicos de Corpus (Corpus-based Translation Studies, CTS)

ECPC se nutre asimismo de las aportaciones realizadas por parte de los Estudios Traductológicos de Corpus, como es el Translational English Corpus (TEC)[2], elaborado bajo la dirección de Mona Baker (Baker, 2004), que ha servido de objeto de estudio para otras investigadoras como Kenny (2001), Laviosa (2002) y, posteriormente, Saldanha (2004) y Winters (2004). TEC puede consultarse a través de la interfaz web desarrollada por Luz (2000). Sin embargo, este corpus monolingüe sólo contiene traducciones. ECPC, da un paso más y ofrece también los textos originales (lo que conforma el corpus paralelo EN-ES) y textos producidos por hablantes nativos (lo que da lugar a sendos corpus comparables en español CD, e inglés HC). Precisamente, esta arquitectura que caracteriza al Archivo ECPC, un corpus paralelo más corpus comparables, se ha tomado del English-Nowegian Parallel Corpus (ENPC), la aportación del tándem Johansson y Oksefjell (2000).

En una línea muy similar pero con el fin de estudiar la interpretación en el Parlamento Europeo se enmarca el Eropean Parliament Interpreting Corpus (EPIC)[3] (Sandrelli, Bendazzoli, y Russo, 2010). Su corpus, de un tamaño más modesto que ECPC, recoge transcripciones de los discursos originales y sus interpretaciones en inglés, español e italiano en las que se ha anotado información relacionada con los oradores, rasgos propios de la interpretación y se han etiquetado los textos morfosintácticamente. Además es posible consultarlo utilizando una interfaz web creada al efecto.

2.3 Procesamiento del Lenguaje Natural (PLN)

ECPC también se ha inspirado en propuestas procedentes del PLN como el corpus Europarl (Koehn, 2005). Este Archivo, compuesto por las actas del Parlamento Europeo en 11 idiomas alineadas al nivel de la frase, fue creado con el fin de entrenar sistemas estadísticos de traducción automática. Se ha seguido una metodología similar en cuanto a la obtención del material en bruto y su preparación para el alineado. No obstante, dado que los objetivos investigadores son diferentes, ha sido preciso adaptar esa propuesta y ampliarla tal y como se describe en el apartado 5 de metodología, poniendo un acento especial en la obtención y anotación de información metatextual acerca de los oradores no contenida en los textos y en la revisión del etiquetado y el alineado del corpus paralelo.

Tiedemann y Nygaard (2004) aprovechan el corpus recopilado por Koehn y lo incluyen en su colección de corpus libres, OPUS[4]. Los dos investigadores procesaron este material y lo pusieron a disposición de la comunidad científica, que puede descargarlo en diferentes formatos o consultarlo mediante una interfaz web que utiliza el Corpus Workbench (CWB). Otros grupos de investigación también han hecho accesible este corpus (o versiones más recientes) mediante interfaces web[5][6].

3 Objetivos

3.1 Objetivo principal

El objetivo principal del grupo ECPC es conocer y profundizar en el estudio del discurso parlamentario como género textual con especial atención a la influencia de la traducción en dicho género.

[2]Web con información acerca de TEC y acceso a la herramienta para su consulta: `http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm`

[3]Se puede obtener más información y consultar el corpus en `http://sslmitdev-online.sslmit.unibo.it/corpora/corporaproject.php?path=E.P.I.C.`

[4]Interfaz web para la versión 3 de Europarl basada en el CWB `http://opus.lingfil.uu.se/bin/opuscqp.pl?corpus=Europarl3`

[5]La versión 5 del mismo corpus accesible gracias al proyecto Per-Fide en `http://per-fide.di.uminho.pt/query/`

[6]Los desarrolladores del CWB también ofrecen una interfaz web para la versión 3 del Europarl en `http://linglit193.linglit.tu-darmstadt.de/CQP/Europarl/frames-cqp.html`

3.2 Objetivos secundarios

- Crear un Archivo en formato electrónico compuesto por diversos corpus que permita la combinación y comparación de los mismos.
- Desarrollar parámetros para realizar estudios contrastivos a partir de los corpus comparables y paralelos que conforman dicho Archivo, entre los que se destacan los siguientes objetivos investigadores:
  - Examinar el grado de similitud y/o diferencia entre los discursos emitidos en el Parlamento Europeo y aquellos emitidos en parlamentos nacionales de diferentes Estados Miembros (el Congreso de los Diputados español y la Cámara de los Comunes británica).
  - Establecer una comparación de la representación de la identidad europea entre los distintos parlamentos objeto de estudio.
  - Realizar un estudio del discurso y de la ideología en los diferentes parlamentos.
- Difundir tanto el conocimiento derivado de la realización de los estudios contrastivos anteriores como los resultados del proyecto.
- Desarrollar herramientas de análisis accesibles vía web, con el fin de permitir la consulta del Archivo y la replicación de los estudios realizados. Dichas herramientas deben facilitar la generación en línea de concordancias monolingües y paralelas y la obtención de información estadística relevante para describir y comparar los fenómenos estudiados.
- Elaborar recursos de ayuda a la traducción como memorias de traducción, glosarios, etc.
- Originar un recurso de referencia para las actividades relacionadas con el Procesamiento del Lenguaje Natural (como la traducción automática o la extracción terminológica, entre otras).
- Diseñar propuestas didácticas en torno a la traducción del discurso parlamentario como género textual dirigidas tanto a estudiantes de traducción como a traductores profesionales.

4 Descripción del Archivo

El Archivo de discursos parlamentarios ECPC está compuesto por diferentes corpus que han sido recopilados en formato electrónico. Estos corpus son:

- Discursos procedentes de la Cámara Baja británica (House of Commons, HC)
- Discursos procedentes de la Cámara Baja española (Congreso de los Diputados, CD)
- Discursos en español procedentes del Parlamento Europeo (ES)
- Discursos en inglés procedentes del Parlamento Europeo (EN)

La muestra descargada contiene los discursos emitidos a lo largo del periodo correspondiente a la VI legislatura del Parlamento Europeo (20 de julio de 2004 al 30 de julio de 2009).

Los textos procedentes del Parlamento Europeo reúnen una serie de características que los hacen únicos y que conviene señalar. En primer lugar, la lengua en la que se expresan los oradores a menudo no es su lengua materna, por lo que podemos encontrarnos ante textos producidos por hablantes no nativos. En segundo lugar, las actas no son transcripciones literales de lo dicho por el orador, ni de las interpretaciones. Toda intervención pasa por un proceso de edición que consiste en:

1. La transcripción y corrección del texto oral en la lengua original a partir de la grabación de la sesión siguiendo una serie de normas[7] (el resultado se conoce como versión "arcoíris").
2. La traducción de cada intervención al inglés y, posteriormente, la traducción desde la versión inglesa al resto de lenguas oficiales.

Los rasgos anteriormente descritos dificultan la distinción entre textos influidos por la traducción de aquellos que no lo han sido. Aunque en la mayoría de casos se conoce la lengua original en la que el orador dio su discurso no podemos tener la certeza de si el orador habló en su lengua materna o en una segunda lengua. Sí que podemos saber, sin embargo, si una intervención se ha visto afectada en mayor o menor medida por los procesos de traducción que se realizan en el seno de la Dirección General de Traducción del Parlamento Europeo desde la emisión del discurso oral hasta su publicación en la página web.

Es decir, si en la versión española de las actas encontramos una intervención en español, podemos afirmar que la versión publicada no ha sido mediada por un proceso de traducción (en verde en el cuadro 1). Si la lengua original fue el

[7] Véase `http://www.europarl.europa.eu/transl_es/plataforma/pagina/guia/cre_normas.htm`

inglés, podemos afirmar que se trata de una traducción directa del inglés al español (en amarillo). Sin embargo, para las demás intervenciones pronunciadas en una lengua distinta al español o al inglés, sólo podemos decir que se trata de una intervención que ha sido mediada por un proceso de traducción indirecta en el que el inglés ha sido la lengua pivote (en rojo).

Audio	Arcoiris	EN	ES	IT	DE
EN	EN	EN	ES	IT	DE
ES	ES	EN	ES	IT	DE
IT	IT	EN	ES	IT	DE
DE	DE	EN	ES	IT	DE

Cuadro 1: Grado de mediación interlingüística en las actas del Parlamento Europeo publicadas en Internet

En cuanto al tamaño, para el Archivo, que comprende las actas desde 2004 a 2009 de cada uno de los Parlamentos estudiados, el número de *tokens* por cada corpus puede apreciarse en el cuadro 2:

Corpus	# Tokens
EN	21 737 797
ES	22 685 242
HC	47 712 000
CD	23 734 230

Cuadro 2: Tamaño del corpus ECPC para el periodo 2004-2009

5 Metodología

El grupo ECPC utiliza la metodología de los Estudios Traductológicos de Corpus para estudiar el discurso parlamentario como género textual. Al aplicar dicha metodología de trabajo hemos pretendido superar el tradicional dilema: primar o bien la calidad o bien el tamaño del Archivo. Para obtener un corpus de un tamaño suficiente y que respondiese a los fines perseguidos por el proyecto hubo que ir más allá de los métodos de etiquetado manual tradicionales en nuestro ámbito. A continuación describimos someramente el proceso seguido en este proyecto para obtener nuestro Archivo que se compone de 6 fases: 1) recopilación; 2) almacenamiento; 3) transformación en XML; 4) enriquecimiento de los textos con metadatos sobre los oradores; 5) control de calidad y; 6) alineado de los corpus paralelos.

5.1 Recopilación del Archivo

Estos discursos se han descargado de los diarios de sesiones en formato electrónico (documentos en HTML) accesibles en las respectivas páginas web de cada uno de los diferentes parlamentos. Se eligió el formato HTML frente al PDF pues el primero se puede manipular más fácilmente y el texto está más limpio que en el caso del segundo lo cual facilitó su procesamiento posterior. Por otra parte, para automatizar esta fase se utilizó un *web crawler* al que se le suministró un listado con todas las direcciones que debía descargar. Finalmente se obtuvo un único documento para cada sesión plenaria que contenía todas las intervenciones realizadas durante ese día.

5.2 Almacenamiento del material

Una vez obtenidos los textos que configuran nuestro Archivo, se hizo necesario utilizar un sistema que nos permitiera almacenar toda esta información y que además, dada la naturaleza del grupo con miembros radicados en distintas universidades europeas, estuviese disponible y al alcance de todos ellos. En concreto, buscamos una solución que nos permitiese: 1) acceder al material desde cualquier lugar; 2) poder compartirlo con el resto de miembros; 3) contar con un historial de versiones y; 4) un sistema de control de cambios.

La tecnología elegida para cubrir estas necesidades fue un repositorio *Subversion*, de gran éxito en el ámbito del desarrollo de software.

De este modo todos los investigadores tenían acceso a una copia central alojada en un servidor corporativo de la Universitat Jaume I a partir de la cual podían generar una copia local sobre la que realizar los cambios y posteriormente compartirlos con el resto de miembros. En todo momento, el sistema permitía registrar un historial con todas las modificaciones y gestionar los cambios pudiendo comprobar en qué habían consistido los mismos, deshacerlos, etc.

5.3 Transformación en XML

En este fase del proyecto el objetivo consistió en estructurar, limpiar y anotar automáticamente información metatextual contenida en los mismos textos recopilados. Para ello se identificaron patrones formales en el código fuente HTML de las páginas descargadas previamente, se escribieron expresiones regulares para anotar la información deseada en XML mediante operaciones de búsqueda y reemplazo y, finalmente, se encadenaron en *scripts* de Perl para procesar los documentos de cada subcorpus por lotes. Se eligió el formato XML frente al convencional texto plano para poder describir con precisión la estructura de los documentos, anotar información

sobre los participantes en la situación comunicativa y facilitar el procesamiento del corpus en fases posteriores.

La transformación en XML de los textos y el etiquetado de la información deseada no fueron tareas sencillas porque a lo largo del tiempo el formato de las actas ha ido sufriendo leves modificaciones, por un lado y, por otro, las marcas HTML no siempre seguían de forma sistemática los patrones generales que se habían identificado inicialmente, introduciendo ruido en el XML resultante. Estos hechos obligaron a pilotar y corregir los *scripts* para mejorar el rendimiento de los mismos.

Finalmente, en esta primera fase se pudo extraer la información siguiente:

- En cuanto a los textos: orden del día, encabezados para cada punto del orden del día, intervenciones, la lengua (o lenguas) de cada intervención, el modo en que cada intervención fue presentada (oral o por escrito), comentarios acerca de los procedimientos de las cámaras o acciones y/o reacciones de los oradores.
- En cuanto a los oradores: nombre, grupo parlamentario y cargo.

5.4 Enriquecimiento del etiquetado

A continuación se procedió a enriquecer el etiquetado obtenido en la fase anterior con más información acerca de los oradores que habían participado en cada debate. Dado que ya se habían anotado los nombres de los oradores responsables de cada intervención se pudieron añadir más datos acerca de los mismos tales como: tratamiento, afiliación política, sexo, fecha de nacimiento, lugar de nacimiento y país de procedencia. Estos datos se suministraron a partir de una base de datos creada al efecto con información extraída o bien de las webs oficiales de las distintas cámaras (CD y EN/ES) o bien proporcionada por los servicios de documentación (HC).

Para obtener esa información adicional, en el caso del CD y el EN/ES, se volvió a emplear un *web crawler* para descargar la página web en HTML que contenía la información personal oficial de cada diputado tal y como las proporcionaba la cámara correspondiente. Se extrajo la información necesaria mediante un *script* de Perl basado en búsquedas de patrones y se estructuró de forma tabulada. Para hacer efectiva la incorporación de la nueva información otro *script* de Perl leía las etiquetas que señalaban los nombres de los oradores en las actas en XML, buscaba en la base de datos dicho nombre y si lo encontraba recuperaba el resto de información y la incorporaba al texto en forma de etiquetas XML. En el caso del HC se modeló la información proporcionada en forma de hojas de cálculo de modo que pudiese ser utilizada por el mismo *script* al que nos acabamos de referir.

En la eventualidad de que un orador no apareciese en la base de datos porque no pertenecía al organismo objeto de estudio (tales como miembros del Gobierno, expertos, Comisarios, miembros del Consejo, etc.) se completó su información con los datos proporcionados en las páginas web oficiales de la institución a la que pertenecían y cuando no fue posible se acudió a fuentes secundarias como la Wikipedia.

5.5 Control de calidad

Acorde con la idea de obtener un corpus de cientos de millones de *tokens* sin sacrificar la calidad, se impuso un método de revisión para comprobar que el etiquetado realizado era lo suficientemente preciso y ayudaba a extraer toda la información contenida en los textos.

De nuevo el marcado en XML de los textos volvió a ser una ventaja pues permitió fácilmente comprobar si los documentos estaban bien formados y si seguían las especificaciones recogidas en la DTD que describía el juego de etiquetas empleado en nuestro corpus.

Para el CD y el HC se realizó un único control de calidad en este punto. Sin embargo, para las versiones inglesa y española del Parlamento Europeo, se realizaron dos. La primera antes del paso descrito en el apartado 5.4, consistente en comprobar que el XML estaba bien formado y era válido, junto con la comparación de la estructura del etiquetado de cada bitexto. Esta comprobación extra se realizó para detectar cualquier tipo de error generado en la fase 5.4 que pudiese afectar al alineado, como intervenciones que habían quedado sin detectar, frases de texto etiquetadas como comentarios o notas, etc. La segunda revisión se realizó tras enriquecer el etiquetado limitándose a comprobar la corrección del XML en términos de validez y forma.

5.6 Alineado de los corpus paralelos

Para el alineado de las versiones española e inglesa de los debates del Parlamento Europeo se empleó como gestor y editor del alineado InterText server[8], creado por Pavel Vondřička. Esta

[8] InterText server `http://wanthalf.saga.cz/intertext#ITserver`

herramienta, desarrollada en el marco del proyecto InterCorp, está concebida precisamente para gestionar el alineado de corpus paralelos multilingües, es muy flexible en cuanto al formato de entrada de los textos, siempre que estén anotados en XML, y soporta la codificación de caracteres Unicode. Esta herramienta es un sitio web dinámico basado en PHP y MySQL. Esta arquitectura permite que tanto los administradores y coordinadores del flujo de trabajo como los editores encargados de la revisión del alineado puedan trabajar en línea desde distintos lugares utilizando una GUI intuitiva. Además, la herramienta cuenta con una CLI muy útil que permite realizar las principales tareas administrativas por lotes como la importación de las dos versiones de cada bitexto, el alineado y la exportación del resultado.

InterText deja el alineado automático en manos de dos potentes alineadores, HunAlign (Varga et al., 2005) y TCA2 (Hofland y Johansson, 1998), que se pueden integrar en el sistema. El primero emplea un algoritmo estadístico que devuelve con relativa rapidez un alineado aceptable. El segundo se sirve de un enfoque más sofisticado basado en un conjunto de algoritmos que junto con un diccionario de términos sopesa distintas opciones de las que elije aquella que ha obtenido una mejor puntuación. Aunque TCA2 necesita mucho más tiempo para alinear un mismo texto, se optó por este alineador pues el resultado final es más fiable, de cara a su revisión manual. En comparación, HunAlign tiende a crear relaciones de alineado 1:1 erróneas que pueden resultar indetectables cuando se sigue el proceso de revisión que explicamos a continuación.

Para poder alinear los textos en primer lugar se dividió, mediante un *script* de Perl y de forma automática, cada intervención en párrafos y frases, siendo esta última división la unidad mínima de alineado. Posteriormente se importó en InterText todo el corpus utilizando el comando *import* de la CLI y se alineó automáticamente el corpus paralelo con el alineador TCA2. El uso de este alineador en concreto permitió que la labor de revisión y edición del alineado se limitara a aquellos segmentos que no fueran relaciones 1:1 (una unidad del texto original alineada con una unidad del texto traducido).

Para los textos correspondientes a los años 2004, 2006, 2007, 2008 y 2009 se alinearon un total de 238 textos del corpus EN con los correspondientes a la versión ES. El alineado automático produjo un total de 589 665 segmentos, de los cuales un 97,66 % de los casos consistió en relaciones 1:1 y el 2,33 % restante correspondió a otro tipo de relaciones. Sólo se revisaron manualmente por medio de la GUI estas últimas, puesto que InterText permite encontrar las relaciones que no son 1:1 directamente.

En la revisión se comprobó: 1) que la división en frases fuese correcta; 2) que la propuesta de alineado fuese correcta en cuanto al contenido.

En el primer nivel de revisión, si la división era incorrecta debido a que las reglas del *script* se habían topado con un caso no previsto se procedió a separar o fusionar las frases afectadas. Si el problema se había producido por la falta de signos de puntuación que delimitasen el final de la frase se comprobó contrastando con el HTML original si la puntuación se había perdido en alguno de los procesos de transformación anteriores o bien se trataba de una errata del original. Si el error se debía a un fallo introducido durante el procesado de los textos se corrigió tanto aquel elemento que había generado el error como la división. Si el error se debía, por el contrario, a una errata la división se corrigió sin añadir la puntuación que en teoría faltaba en el original.

En el segundo nivel de revisión, se detectaron los siguientes tipos de relaciones:

ES≥1:EN=0 una frase o más en español por ninguna en inglés.

ES=0:EN≥1 ninguna frase en español por una frase o más en inglés.

ES=1:EN>1 una frase en español por varias frases en inglés.

ES>1:EN=1 varias frases en español por una frase en inglés.

ES>1:EN>1 más de una frase en español para más de una frase en inglés.

Rel. 1≠1	no rev.	rev.	no rev.	rev.
ES≥1:EN=0	620	38	4,50 %	0,35 %
ES=0:EN≥1	1126	182	8,17 %	1,69 %
ES=1:EN>1	7517	6808	54,57 %	63,37 %
ES>1:EN=1	4510	3637	32,74 %	33,85 %
ES>1:EN>1	0	77	0 %	0,71 %

Cuadro 3: Relaciones no 1:1, diferencias entre el resultado del alineado automático con TCA2 (no rev.) y la revisión manual posterior (rev.)

Se verificó que para ninguno de los 5 casos la relación no 1:1 se debiese a una propuesta errónea de TCA2. Si se detectaba algún fallo en este sentido (una o más frases asociadas a un segmento contiguo que en realidad pertenecían al segmento objeto de revisión) se corrigió utilizando el editor de InterText. Para los casos 1 y 2 se comparó además con la versión HTML original con el

fin de descartar que la omisión de información en una de las versiones se debiese al procesado previo de los textos. Los fenómenos del tipo 5 recogen relaciones complicadas (Frankenberg-Garcia, Santos, y Silva, 2006, pp. 8-10) como alineados con frases enteras y fracciones y reordenaciones. En nuestro caso, dada la baja incidencia de esta clase de segmentos no hemos realizado intervenciones más allá de lo expuesto con anterioridad. Además, puesto que InterText registra los cambios realizados, no se anotaron aquellos segmentos que sufrieron modificaciones durante el proceso de revisión.

Tras la revisión se obtuvo un total de 589 445 segmentos alineados, de los cuales un 98,14 % consistió en relaciones 1:1 y el 1,82 % restante correspondió a otro tipo de relaciones. Dicha revisión arroja un total de 2809 diferencias en cuanto a la segmentación en frases y las relaciones de alineado propiamente dichas.

Por último, se exportaron los textos alineados en tres formatos distintos:

corresp: se obtiene un documento por idioma, con los identificadores de los elementos alineables actualizados (en nuestro caso las frases) donde la información sobre el alineado se codifica mediante un atributo llamado "corresp"que indica el identificador de las unidades equivalentes en la otra lengua. Este formato es usado como input para Glossa.

segs: se obtiene un documento por idioma, con los identificadores de las frases actualizados donde la información sobre el alineado se codifica en el mismo texto utilizando unos elementos llamados "seg"para delimitar las areas de texto equivalentes en cada versión. Este formato es el que puede utilizar ParaConc como input.

TEI alignment format: se obtiene un documento por idioma, con los identificadores de las frases actualizados, pero los detalles sobre el alineado se almacenan en un tercer documento en formato XML. Este formato permite volver a importar el alineado en InterText y sigue la recomendación recogida en TEI para codificar este tipo de información.

5.7 Desarrollo de software de consulta

En la actualidad Luz (Calzada Pérez y Luz, 2006) se encarga de desarrollar un conjunto de herramientas que posibilitarán la consulta del corpus vía web y que incorpora distintas características como la generación de concordancias monolingües, la selección de distintos subcorpus atendiendo a las variables codificadas como información metatextual y la representación visual de distintos tipos de información lingüística.

Al mismo tiempo Anders Nøklestad trabaja en la adaptación de Glossa (Nygaard et al., 2008) con el fin de facilitar la generación de concordancias paralelas con caracteríscas similares a la herramienta de Luz. Glossa es una interfaz web que emplea el Open Corpus WorkBench (Evert y Hardie, 2011) como motor de búsqueda y gestor del corpus.

6 Aplicaciones

6.1 Aplicación en la investigación

Entre las potenciales aplicaciones investigadoras de este corpus cabe destacar la posibilidad de profundizar en el conocimiento sobre el género de los discursos parlamentarios (tanto del Parlamento Europeo como de otros parlamentos de Estados Miembros) en una línea similar a los trabajos de Partington (2003) y Guerini, Strapparava, y Stock (2008); y además examinar la influencia de la traducción (Calzada Pérez, 2007). Sin embargo, este material gracias el etiquetado metatextual acerca de los oradores puede ser de interés no sólo para estudiosos de la traducción y la lingüística sino también de la sociolingüística, sociología e incluso de los estudios de género.

6.2 Aplicación en la didáctica

Los potenciales beneficiarios de esta herramienta no se limitan al mundo científico pues los profesionales del ámbito de la traducción, los centros de formación de traductores y los aprendices de esta disciplina también podrán consultar el material utilizando una interfaz web de consulta similar a la del BYU-BNC de Mark Davies[9] o Glossa (Nygaard et al., 2008). Este enfoque ya ha sido explotado con éxito en el ámbito de la enseñanza de lenguas (Moreno Jaén, Serrano, y Calzada Pérez, 2010) y de la traducción (Zanettin, Bernardini, y Stewart, 2003; Beeby, Rodríguez Inés, y Sánchez-Gijón, 2009).

7 Perspectivas de futuro

Algunas de las tareas que el grupo ECPC está realizando en la actualidad o que tiene previstas realizar en el futuro son:

[9]BYU-BNC: The British National Corpus `http://corpus.byu.edu/bnc`

1. Clasificación temática de las intervenciones para poder agruparlas en subcorpus "especializados" utilizando el JRC EuroVoc Indexer (Pouliquen, Steinberger, y Ignat, 2003).
2. Ampliación del Archivo con la versión en alemán de los discursos del PE (DE) y su homólogo nacional, el Bundestag alemán (DB) para el mismo periodo de tiempo.
3. Etiquetado morfosintáctico de todos los corpus que componen el Archivo con TreeTagger[10].

Cabe reseñar la creación de dos corpus derivados de la experiencia acumulada en el seno del grupo ECPC: EMPAC y TraDiCorp.

El EMPAC (EuroparlTV Multimedia Parallel Corpus) es un corpus multilingüe de subtítulos de las noticias emitidas en el canal EuroparlTV del Parlamento Europeo. Actualmente el corpus presenta una versión etiquetada del año 2010 en inglés y en español.

El TraDiCorp (Translation Difficulties Corpus) es un corpus paralelo inglés-español de múltiples traducciones de textos de las actas del Parlmento Europeo realizadas por estudiantes de grado y máster de traducción, con problemas de traducción anotados por los mismos estudiantes en el texto original.

Bibliografía

Baker, Mona. 2004. A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, 9(2):167–193.

Beeby, Allison, Patricia Rodríguez Inés, y Pilar Sánchez-Gijón. 2009. *Corpus use and translating: corpus use for learning to translate and learning corpus use to translate*, volumen v. 82 de *Benjamins translation library*. John Benjamins, Amsterdam.

Calzada Pérez, María. 2007. *Transitivity in translating: the interdependence of texture and context*. Peter Lang, Bern/Berlin/Bruxelles/Frankfurt am Main/New York/Oxford/Wien.

Calzada Pérez, María y Saturnino Luz. 2006. ECPC: Technology as a tool to study the (linguistic) functioning of national and transnational European parliaments. *Journal of Technology, Knowledge and Society*, 5(2):53–62.

Evert, Stefan y Andrew Hardie. 2011. Twenty-first century Corpus Workbench : Updating a query architecture for the new millennium. En *Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, UK.

Frankenberg-Garcia, Ana, Diana Santos, y Rosario Silva. 2006. COMPARA: Sentence alignment revision and markup. Informe técnico, Linguateca.

Guerini, M, C Strapparava, y O Stock. 2008. CORPS: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology & Politics*, 5(1):19–32.

Hofland, Knut y Stig Johansson. 1998. The Translation Corpus Aligner: A program for automatic alignment of parallel texts. En Stig Johansson y Signe Oksefjell, editores, *Corpora and Cross-linguistic research: Theory, Method and Case Studies*, volumen 24. Rodopi, Amsterdam; New York, páginas 87–100.

Johansson, Stig y Signe Oksefjell. 2000. The English-Norwegian Parallel Corpus: Current Work And New Directions. En S Botley McEnery A., y A Wilson, editores, *Multilingual corpora in teaching and research*. Rodopi, Amsterdam; Atlanta, páginas 134–147.

Kenny, Dorothy. 2001. *Lexis and creativity in translation: a corpus-based study*. St. Jerome, Manchester.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. En *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, volumen 5, Phuket, Tailandia, Septiembre. Asia-Pacific Association for Machine Translation and Thai Computational Linguistics Laboratory.

Laviosa, Sara. 2002. *Corpus-based translation studies: theory, findings, applications*. Rodopi, Amsterdam; New York.

Luz, Saturnino. 2000. A software toolkit for sharing and accessing corpora over the Internet. En M Gavrilidou G Carayannis S Markantonatou S Piperidis, y G Stainhauer, editores, *Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000*, Athenas, Greece, Mayo. European Language Resources Association (ELRA).

Moreno Jaén, María, Fernando Serrano, y María Calzada Pérez. 2010. *Exploring new paths in language pedagogy: lexis and corpus-based language teaching*. Equinox, London.

[10] TreeTagger `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

Nygaard, L, J Priestley, A Nø klestad, y J B Johannessen. 2008. Glossa: A multilingual, multimodal, configurable user interface. En *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. European Language Resources Association (ELRA), Marrakesh, Morocco, Mayo.

Partington, Alan. 2003. *The Linguistics of Political Argument: the Spin-Doctor and the Wolf-Pack at the White House*. Routledge, London.

Pouliquen, Bruno, Ralf Steinberger, y Camelia Ignat. 2003. Automatic annotation of multilingual text collections with a conceptual thesaurus. En *Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology - Its Potential and Practicalities (EUROLAN 2003)*, Bucarest, Romania.

Saldanha, G. 2004. The Translator's Presence in the Text: A Corpus-based Exploration. En *1st IATIS conference: Translation and the Construction of Identity*, Seul, South Korea, Agosto. International Association for Translation and Interecultural Studies (IATIS).

Sandrelli, A, C Bendazzoli, y M Russo. 2010. European Parliament Interpreting Corpus (EPIC): Methodological Issues and Preliminary Results on Lexical Patterns in Simultaneous Interpreting. *International Journal of Translation Studies*, 22(1-2):167–206.

Tiedemann, J y L Nygaard. 2004. The OPUS corpus–parallel and free. En M T Lino M F Xavier F Ferreira R Costa, y R Silva, editores, *In Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, Mayo. European Language Resources Association (ELRA).

Varga, D, P Halácsy, A Kornai, V Nagy, L Németh, y V Trón. 2005. Parallel corpora for medium density languages. En *Proceedings of the RANLP 2005*, Borovets, Bulgaria, Septiembre.

Winters, Marion. 2004. F. Scott Fitzgerald's Die Schönen und Verdammten – A Corpus based Study of Translators' Style: Modal particles and their influence on the narrative point of view. En *1st IATIS conference: Translation and the Construction of Identity*, Seul, South Korea, Agosto. International Association for Translation and Interecultural Studies (IATIS).

Zanettin, Federico, Silvia Bernardini, y Dominic Stewart, editores. 2003. *Corpora in translator education*. St. Jerome, Manchester, UK ; Northampton, MA.