UNIVERSITAT JAUME·I

Jornades de Foment de la Investigació

ECPC: UNA MIRADA AL DISCURSO PARLAMENTARIO EUROPEO DESDE LOS ESTUDIOS TRADUCTOLÓGICOS DE CORPUS

Autors

Noemí MARÍN
José Manuel MARTÍNEZ

ÍNDICE

RESUMEN

Esta comunicación se propone presentar el proyecto ECPC (European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos). En primer lugar, se ofrece una relación de los integrantes del grupo ECPC afiliados a la Universitat Jaume I (España); la Dublin City University y el Trinity College Dublin (Irlanda); y a la University of Birmingham, la University of Edinburgh y la University of Manchester (Reino Unido). A continuación, se rastrean los antecedentes teóricos del proyecto: la Lingüística de Corpus (CL) y los Estudios Traductológicos de Corpus[1] (CTS). El siguiente punto se centra en la descripción del Archivo ECPC. Posteriormente, se da cuenta de la actividad desarrollada para crear dicho Archivo. A continuación se presentan los principales objetivos investigadores en torno a este proyecto. Y, finalmente, se apuntan los principales retos a los que tendrá que enfrentarse el grupo en un futuro.

PALABRAS CLAVE: Lingüística de corpus, Estudios Traductológicos de Corpus, Archivo, discurso parlamentario europeo, didáctica de la traducción.

EL GRUPO ECPC[2]

El proyecto que presentamos en este artículo no tendría razón de ser sin el elenco de investigadores que su coordinadora, la Dra. María Calzada Pérez, ha logrado reunir en torno al mismo. Es por ello que consideramos oportuno dedicar este primer apartado a presentar, aunque sólo sea de manera sucinta, a los miembros que integran el grupo ECPC.

El grupo ECPC surge en el año 2005, cuando el proyecto presentado por la Dra. María Calzada Pérez recibe financiación por parte del Ministerio de Educación y Ciencia para su puesta en marcha. Ella es desde ese momento la encargada de coordinar el proyecto desde la Universitat Jaume I de Castellón. Hasta entonces la Dra. María Calzada Pérez se había dedicado al estudio de los discursos del Parlamento Europeo desde el punto de vista de la transitividad, y si bien no abandona su objeto de estudio, sí que adopta una nueva metodología investigadora. Su mano derecha, en cuanto a inspiración teórica se refiere, no es otra que la Dra. Mona Baker (University of Manchester), una investigadora de reconocido prestigio en el ámbito de la traductología y la primera en importar la metodología de corpus al campo de la traducción. Junto a esta, destacan la Dra. Dorothy Kenny (Dublin City University), una investigadora muy potente en el campo de los estudios de traducción basados en corpus, así como la Dra. Gabriela Saldanha (University of Birmingham) y la Dra. Marion Winters (University of Edinburh), dos investigadoras muy incipientes en este ámbito, pero muy prometedoras. Todas ellas, salvo la Dra. Marion Winter, junto al Dr. Saturnino Luz (Trinity College Dublin), especialista en informática, ya trabajaron conjuntamente en un proyecto anterior denominado Translational

1.- Esta comunicación se realiza en el marco del proyecto ECPC (European Parliamentary Comparable and Parallel Corpora / Corpus comparables y paralelos de discursos parlamentarios europeos) financiado por el Ministerio de Educación y Ciencia (HUM2005-03756/FILO).

2.- La página web del grupo ECPC es <http:www.ecpc.uji.es>. Por el momento, se encuentra disponible únicamente en inglés, aunque ya se está trabajando en las versiones española y catalana.

English Corpus (TEC), de características similares a ECPC. Los restantes componentes del grupo, de la Universitat Jaume I de Castellón, son la Dra. Rosa Agost, experta en traductología, y los doctorandos Noemí Marín y José Manuel Martínez[3], que están llevando a cabo su trabajo de investigación en el seno de este grupo.

ANTECEDENTES: LA LINGÜÍSTICA DE CORPUS Y LOS ESTUDIOS TRADUCTOLÓGICOS DE CORPUS

> Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered possible but still lunatic. Today it is very popular. (Sinclair, 1991:1)

Estas palabras de John Sinclair, uno de los grandes especialistas en el ámbito de los corpus modernos, resumen claramente la vertiginosa evolución que han experimentado los estudios de corpus, desde sus inicios, cuando nada hacía sospechar su éxito futuro, hasta nuestros días, momento en el que su popularidad es ciertamente indiscutible. Pero, para percibir esta evolución como tal, antes debemos remontarnos al año 1967, cuando W. Nelson Francis y Henry Kučera, de la *Brown University* (Providence, Rhode Island), compilaron el primer corpus electrónico de la historia, el Brown Corpus, de un millón de palabras de inglés estadounidense, que ha permitido la creación de corpus posteriores. Unos años más tarde, concretamente en 1978, Geoffrey Leech (Universidad de Lancaster), Stig Johansson (Universidad de Oslo), y Knut Hofland (Centro Informático Noruego para las Humanidades de Bergen) crean la contrapartida británica del Brown Corpus, el Lancaster-Oslo Bergen Corpus (LOB), también de un millón de palabras.

En los Estudios Traductológicos de Corpus (Corpus-based Translation Studies, CTS) cabe destacar como pionera a Mona Baker, quien a mediados de la década de los 90, acompañada por Dorothy Kenny, Sara Laviosa, Lynne Bowker y Jennifer Pearson, todas ellas piezas indiscutibles en el mapa actual de los CTS, importó la metodología de la LC a los CTS, que se han convertido en un enfoque fundamental para investigar en este campo.

En la actualidad, la LC atraviesa un momento de creciente popularidad. De hecho, tanto en el campo de la LC como en el de los CTS, existe una gama, cada vez más variada, de corpus que cumplen propósitos bien distintos. De entre las diferentes aplicaciones que se conocen, cabe destacar las que han resultado de la fusión de metodología de los CTS con el trasfondo teórico propuesto desde los Estudios Descriptivos de la Traducción (Descriptive Translation Studies, DTS), iniciados por Toury (1995), ya que los DTS se han servido de la metodología de los CTS para apuntalar normas traductoras como la simplificación, la explicitación o la normalización, entre otras.

En el caso concreto de ECPC, la finalidad última es, en términos generales, conocer más a fondo el género del discurso parlamentario, y, en concreto, el género del discurso traducido del Parlamento Europeo mediante la metodología que nos ofrecen los Estudios Traductológicos de Corpus.

3.- José Manuel Martínez Martínez es beneficiario del plan de promoción de la investigación de la Universitat Jaume I gracias a una beca predoctoral (PREDOC/2006/29).

Así pues, tras presentar de manera sucinta el grupo ECPC y esbozar la trascendencia actual de los estudios de Lingüística de Corpus (LC) y de los Estudios Traductológicos de Corpus (CTS), el siguiente apartado se centra en describir los rasgos más relevantes del productivo fruto de ECPC, su Archivo.

EL ARCHIVO ECPC

Antes de pasar a describir en qué consiste el Archivo ECPC, creemos conveniente detenernos brevemente en la definición de corpus que proponen Bowker y Pearson (2002: 9). Según estas autoras, un corpus «can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria». No obstante, un año más tarde Bowker (2003: 170) va más allá e introduce el término *Archivo* para referirse al conjunto de «different type of corpora [which] can be extracted depending on the nature of the study being undertaken at any time». Así pues, basándonos en estas dos definiciones, podemos decir, a grandes rasgos, que ECPC es un Archivo de discursos parlamentarios compuesto por diferentes corpus parciales recopilados en formato electrónico atendiendo a diferentes criterios. A continuación, se muestran los diferentes corpus que componen el Archivo ECPC:

Corpus 1: Discursos procedentes de la Cámara Baja británica (House of Commons, HC)
Corpus 2: Discursos procedentes de la Cámara Baja española (Congreso de los Diputados, CD)
Corpus 3: Discursos originales en inglés procedentes del Parlamento Europeo (PE/EP)
Corpus 4: Discursos originales en español procedentes del Parlamento Europeo (PE/EP)
Corpus 5: Discursos traducidos en inglés procedentes del Parlamento Europeo (PE/EP)
Corpus 6: Discursos traducidos en español procedentes del Parlamento Europeo (PE/EP)

Ante la ingente cantidad y tipos distintos de corpus que han ido proliferando en los últimos años en el ámbito de los Estudios de Traducción, se hace necesario adoptar una clasificación coherente con los mismos. Esto es precisamente lo que hace Laviosa (2002), donde compone una minuciosa taxonomía de los diferentes tipos de corpus electrónicos con el objetivo de brindar a los investigadores de este campo una descripción común de su objeto de estudio y remediar, al mismo tiempo, la falta de sistematicidad existente hasta entonces en la terminología utilizada para clasificar los diferentes tipos de corpus. La tipología propuesta por esta autora consta de cuatro niveles jerárquicos principales. En el primer nivel, Laviosa (2002: 34-35) localiza 6 apartados, el primero de los cuales recoge los corpus de textos completos, de extractos, los mixtos (formados por textos completos y extractos) y los corpus monitorizados o «monitor corpus», y que la traductóloga (2002: 34-35) describe de la siguiente forma: «A monitor corpus is made up of full texts which are scanned on an ongoing basis so that the corpus is continuously updated.». En el segundo apartado de este primer nivel, Laviosa (2002: 35) identifica corpus diacrónicos y sincrónicos; en el tercer apartado sitúa los generales y los terminológicos; en el cuarto, los monolingües, bilingües y multilingües; en el quinto se centra en la lengua (o lenguas) del corpus; y en el sexto apartado (con el que se cierra el primer nivel de la taxonomía), distingue entre corpus escritos, orales o mixtos (mezcla de las dos modalidades anteriores).

En el segundo nivel taxonómico, Laviosa (2002: 36) divide los corpus monolingües en simples (recopilación de textos en una única lengua) y comparables (formados por dos corpus monolingües simples, uno formado por textos originales en una lengua A y el otro formado por textos traducidos en esa misma lengua A); los bilingües quedan clasificados en: paralelos (textos originales en lengua A y sus traducciones en lengua B) y comparables (textos originales en lengua A y textos originales en lengua B); y los multilingües se agrupan en: paralelos (textos originales en lenguas diversas con sus respectivas traducciones) y comparables (textos originales en diferentes lenguas). Como destaca esta traductóloga (2002: 36), en la LC el término *corpus comparable* se utiliza para referirse a un «bi/multilingual corpus made up of two or more sets of texts from the same subject domain(s)».

En el tercer nivel de la taxonomía de Laviosa (2002), los corpus simples se bifurcan en traductores y no traductores; los corpus bilingües paralelos son monodireccionales y bidireccionales; y los corpus multilingües paralelos se componen de una única lengua origen, de dos lenguas origen y de varias lenguas origen.

En el cuarto y último nivel, Laviosa (2002) profundiza en los corpus traductores y también encuentra corpus de una, dos o más lenguas de partida, que aún pueden subdividirse, atendiendo al modo de la traducción (escrito u oral, en cuyo caso sería interpretación); al método de la traducción (humana, automática y asistida por ordenador); a la dirección de la traducción (directa o inversa); al estatus del traductor (profesional o aprendiz); y al estatus de la traducción misma (publicada o no publicada).

A tenor de esta clasificación y, teniendo en cuenta los diferentes corpus parciales que componen el Archivo ECPC, podemos decir que ECPC reúne corpus sincrónicos de textos completos, escritos y especializados en el género parlamentario que, tomados de manera individual, son monolingües, en inglés (corpus 1, 3 y 5) y en español (corpus 2, 4 y 6). A su vez, los corpus monolingües en inglés se dividen en corpus simples de textos originales (corpus 1 y 3) y en corpus simples de textos traducidos (corpus 5). Todos estos corpus simples se pueden combinar entre sí creando, de este modo, corpus monolingües comparables (1-5 y 3-5). Del mismo modo, los corpus monolingües en español se dividen en corpus simples de textos originales (corpus 2 y 4) y en corpus simples de textos traducidos (corpus 6), que también pueden combinarse entre sí dando lugar a corpus monolingües comparables (2-6 y 4-6). Otras posibles combinaciones son las que dan como resultado corpus bilingües, que, a su vez, pueden ser paralelos bidireccionales (3-6 y 4-5) o comparables (1-2, 1-4, 2-3 y 3-4), siendo ésta última, como puede observarse, la combinación más fructífera.

Cabe señalar, no obstante, que todos estos estudios contrastivos no podrían desarrollarse sin una fase previa de limpieza y tratamiento de los distintos corpus, que permitirán la recuperación automática de ciertos tipos de información requeridos por los investigadores y, en general, para poder explotar al máximo las posibilidades de ECPC. En el siguiente apartado se detalla la metodología que ha guiado el arduo proceso de etiquetado automático.

¿QUÉ HEMOS HECHO?

En la actualidad, el grupo ECPC ha alcanzado varios objetivos fundamentales que constituyen la base para el trabajo investigador que se desarrollará en una fase posterior. A continuación, se enumeran los objetivos cumplidos:

- La creación de un Archivo electrónico que ya ha sido descrito en el apartado 3 del presente artículo y que se compone de los discursos parlamentarios emitidos durante 2005 en la Cámara de los Comunes, el Congreso de los Diputados y el Parlamento Europeo (versiones en español e inglés) tal y como aparecen reflejados en los respectivos diarios de sesiones.

  Para ello se descargaron dichos diarios en formato electrónico (documentos en html) desde las páginas web de cada una de las cámaras para facilitar su procesado en fases posteriores.
- El siguiente paso ha consistido en etiquetar automáticamente los documentos para incluir información metatextual como, por ejemplo, el partido político al que pertenece cada ponente que interviene, el sexo, la lengua en la que se ha dirigido a la cámara, etc. Para ello se ha aprovechado la información que aparecía en el propio texto.

  El procedimiento suponía analizar de forma independiente el diario de sesiones de cada parlamento para identificar patrones en el código fuente que permitiesen codificar la información metatextual de forma que se pudiese utilizar con fines investigadores. Una vez hecho esto se escribió una secuencia (o macro) de búsquedas y reemplazos complejos basados en expresiones regulares que permitía identificar las partes del texto útiles, codificar esa información en etiquetas xml, deshacerse del código html innecesario y formatear los documentos de acuerdo a una definición de tipo de documento DTD para que los textos resultantes fuesen documentos xml bien formados. Estos documentos permiten restringir las búsquedas de fenómenos lingüísticos atendiendo a parámetros metatextuales relevantes, como el partido político, el país al que representa el parlamentario, el sexo, etc., y ver si distintas poblaciones se expresan de forma diferente, y lo que es más importante, si se observan tendencias asociadas a estas variables en la traducción.
- Una vez realizado lo anterior los documentos quedan listos para su consulta. Para llevar a cabo el análisis de los documentos es necesaria una herramienta de consulta del corpus. Debido a que uno de los objetivos últimos es poner el corpus a disposición de la comunidad investigadora para su libre acceso a través de Internet, la herramienta para analizar el corpus debía ser libre y basada en web. Así pues, el Dr. Saturnino Luz, miembro del proyecto ECPC, desarrolló un generador de concordancias bautizado como ConcECPC. La versión 1.0 de este software se basa en el trabajo pionero que él mismo realizó para el proyecto TEC (Translational English Corpus). Esta herramienta basada en una estructura cliente-servidor está escrita en java y permite seleccionar subcorpus a partir de la información metatextual codificada en xml, generar concordancias a partir de los subcorpus seleccionados y consultar toda la información metatextual asociada a la concordancia mostrada.

¿PARA QUÉ?

El fin último de todo este trabajo es profundizar en el estudio de los géneros parlamentarios y sus subgéneros en general y centrarse en los discursos emitidos en el Parlamento Europeo con la ayuda de la metodología de los Estudios Traductológicos de Corpus.

Los objetivos investigadores de este proyecto se enmarcan dentro de los Estudios Descriptivos de la Traducción por un lado, mientras que habría otros objetivos relacionados con la aplicación de los recursos generados por el grupo en otros campos como la didáctica de la traducción o disciplinas como el Procesamiento del Lenguaje Natural.

En cuanto a los primeros se podrían destacar los siguientes:

- Examinar el grado de similitud/diferencia entre los discursos de distintos parlamentos nacionales (como The House of Commons o el Congreso de los Diputados) y el Parlamento Europeo.
- Evaluar el grado de autonomía del Parlamento Europeo (en comparación con los parlamentos nacionales) en cuanto a conceptos, fenómenos y comportamiento (tanto en un plano general como meramente lingüístico).
- Comparar la representación de la «identidad europea» en cada uno de los parlamentos.
- Contrastar textos originales y traducidos (con sus mensajes locucionarios, ilocucionarios y perlocucionarios).
- Estudiar las normas y universales de traducción así como el estilo.
- Establecer el grado de (macro-micro) similitud/diferencia entre los discursos en diferentes variedades del inglés (género).
- Establecer el grado de (macro-micro) similitud/diferencia entre parlamentos nacionales y el PE (género).
- Estudiar cómo se tratan los temas clave en cada uno de los distintos parlamentos (discurso e ideología).

Pero como señalábamos al principio de este apartado el resultado de esta investigación, al igual que los recursos generados, pueden aplicarse en distintos campos:

- Crear recursos para traductores, como el corpus en sí mismo, que puede convertirse en una herramienta de ayuda a la redacción, puede generar archivos tmx para alimentar las memorias de traducción que emplea el software de traducción asistida, etc.
- Puede servir como recurso de referencia para cualquier actividad relacionada con el Procesamiento del Lenguaje Natural, tal como la traducción automática, la extracción terminológica, data mining, etc.
- Y finalmente, pero no menos importante, la didáctica de la traducción.

Desarrollaremos este último punto un poco más por su impacto en el entorno universitario en el que la investigación y la docencia se encuentran íntimamente relacionados.

Bernardini, Stewart y Zanettin (2003:1) bautizan como «Applied Corpus-Based Translation Studies» a la rama resultante de fusionar los Estudios Traductológicos de Corpus con los Estudios Aplicados de Traducción.

Los tres investigadores, radicados en Italia, afirman (Zanettin et al., 2003) que existen al menos tres áreas relacionadas con la formación de traductores en las que resulta muy interesante utilizar los corpus electrónicos:

1. Estudios traductológicos de corpus
2. Herramientas de traducción asistida por ordenador
3. Corpus en la enseñanza aprendizaje de lenguas

Investigadoras de la didáctica de la traducción como Hurtado (2001) afirman que «la competencia traductora es el sistema subyacente de conocimientos, habilidades, destrezas y actitudes necesarios para traducir». Dicha competencia se caracteriza por ser un conocimiento principalmente operativo en el que las estrategias juegan un papel fundamental y que está compuesta por una serie de subcompetencias, a saber: 1) competencia lingüística en las dos lenguas; 2) competencia extralingüística; 3) competencia de transferencia; 4) competencia instrumental y profesional; 5) competencia psicofisiológica y 6) competencia estratégica.

Los corpus se han utilizado tanto por docentes como discentes como herramienta de apoyo para la adquisición de algunas de estas subcompetencias:

a. Competencia lingüística en las dos lenguas. Por ejemplo, los corpus pueden ser un apoyo para la producción del texto meta, en especial si la lengua de destino no es la lengua materna del traductor que los utilizará para resolver dudas de carácter gramatical, colocaciones, convenciones textuales, etc.
b. Competencia extralingüística. Los corpus pueden ser un instrumento valioso para aumentar los conocimientos temáticos de ámbitos específicos desconocidos para el traductor.
c. Competencia instrumental. Trabajar con corpus permite al alumno que se familiarice con los principios básicos de funcionamiento de la mayoría de herramientas lingüísticas empleadas en traducción (extractores terminológicos, memorias de traducción, herramientas de documentación, etc.).
d. Competencia estratégica. En general, el manejo de corpus puede servir para sortear los problemas encontrados en el desarrollo del proceso traductor y salir airoso de situaciones comprometidas.

Kelly (2005:74) insiste en situar la competencia «instrumental» en un lugar prominente. Según ella:

> It is obvious that professional translators must be familiar with translation technologies, how to use them, and also be able to appraise how they affect the translation process.

Además, englobadas dentro de la competencia instrumental, Kelly (2005) señala como fundamentales tres áreas de conocimiento: «communication and documentary research», «linguistic tools and resources» y «translation tools». En definitiva, esta autora (Kelly, 2005:75) aboga por que el alumno entienda los principios básicos de funcionamiento de estas herramientas en lugar de que aprenda a manejar un producto determinado:

> It is probably more appropriate on training programmes to help students to learn and understand the basics of translation memory technology in general, without necessarily learning any one particular commercial programme.

También los docentes pueden beneficiarse del uso de corpus para enseñar traducción. Baker (1995: 233) subraya, desde un principio, la utilidad de este recurso:

> Multilingual corpora, and the kind of insights they provide on the typical behaviour of so-called "equivalent" items and structures in various languages, can be extremely useful in developing teaching materials for translators and in computer-aided translator training.

Otras autoras coinciden en señalar la necesidad de que los estudiantes aprendan a manejar corpus para enfrentarse con las herramientas más competitivas al mundo profesional que les espera una vez finalicen sus estudios. Así, Pearson (2000:93) sostiene la siguiente opinión:

> We believe, therefore, that it is essential for our students to understand the need for personal glossaries and to develop a methodology for compiling glossaries which they will be able to use in their future careers, whether they are working as translators, terminologists or interpreters.

Por su parte, Kübler (2003:41) abunda en la misma dirección con este comentario:

> Learning to use corpora and corpus-query tools can give future translators the technical skills that were usually not associated with translation, but which seem to be more and more necessary, especially in technical translation.

Por último, Johansson, en el capítulo que cierra el libro de Zanettin et al. (2003), reflexiona sobre el uso de los corpus para la enseñanza y la investigación y concluye con la idea que gracias al uso de corpus:

> Learning becomes a form of research, far from rote learning. And there is just a short step in moving on to the real world of language use and new translation tasks, with a sharpened sense of observation, prepared to meet the unknown.

A todo esto podemos añadir, centrándonos en el archivo ECPC, que los textos que lo componen reúnen ciertas características que pueden resultar útiles como herramienta de aprendizaje en el aula de traducción general y como transición a textos de traducción especializada. Y es que el discurso parlamentario es un género textual rico y repleto de matices. En primer lugar, se encuentra a caballo entre el texto escrito y el oral. En realidad, y siguiendo la conocida clasificación de Gregory y Carroll (1978), podría caracterizarse como un texto escrito para ser leído. Facilitar al alumno la exposición a la oralidad y a la textualidad y enfrentarlo a la detección y valoración de dichas cualidades es indispensable para todo alumno que se está formando como traductor e intérprete. En segundo lugar, el discurso parlamentario presenta un grado de especialización moderado ya que suele (o puede) abordar cuestiones técnicas (relacionadas, por ejemplo, con el ámbito de las telecomunicaciones, el medio ambiente, etc.) cuyo contenido debe estar, no obstante, al alcance de la

comprensión de profanos como la mayoría de los políticos y de los votantes. Es, por tanto, un ventana desde la que el alumno puede practicar «textos generales» acercándose a una especialización controlada, que lo preparará para los estadios más autónomos de su aprendizaje. En tercer lugar, el discurso parlamentario presenta una complejidad retórica intrincada, en la que la cuestión ideológica desempeña un papel destacable. De esta manera, el alumno aprenderá a ser consciente de que las palabras se enmarcan en un entorno concreto, que poseen una finalidad retórica determinada y, a menudo, persiguen objetivos ideológicos insoslayables. En cuarto lugar, y como bien sabemos, en toda clase universitaria, deberían tratarse contenidos transversales y podemos aprovechar la variada temática de los textos de ECPC para hacerlo. Finalmente, ECPC, se compone tanto de corpus paralelos (los del parlamento Europeo) como de otros comparables (parlamentos nacionales con el subcorpus del Parlamento Europeo correspondiente a su lengua). Esta estructura puede permitir el desarrollo de múltiples explotaciones (como ya se ha señalado más arriba).
Por todo lo anteriormente expuesto, ECPC promete ser un instrumento valioso para innovar en la formación de traductores e intérpretes.

Y EN UN FUTURO...

En cuanto a la parte técnica, las tareas que el grupo ECPC todavía tiene que acometer son:

1. Revisión del corpus ya recopilado, correspondiente al año 2005, para detectar posibles errores y carencias del etiquetado.
2. Ampliación del corpus
   a. Descargar los discursos correspondientes al período 2004-2008 de los distintos parlamentos.
   b. Añadir al par de lenguas inglés-español muestras de alemán y francés (versiones del PE en alemán y francés; discursos del Bundestag alemán y el Nationalrat austriaco, y discursos de la Assemblée Nationale francesa, la Chambre des Représentants belga y la Chambre des Députés luxemburguesa).
3. Alineación de las versiones del PE
   a. Para ello se empleará un software desarrollado por la Universidad de Bergen y la Universidad de Oslo llamado TCA2. Se trata de un programa que alinea texto origen y texto meta y con el que el usuario puede intereactuar para verificar/corregir propuestas de alineación cuando el programa detecta casos que se salen de un umbral prefijado de seguridad. Utiliza un método que combina medidas estadísticas (como la gran mayoría de alineadores) con elementos ancla (como números, nombres propios y listas de equivalencias fijas) que proporcionan mayor robustez al sistema. Este software es una evolución de TCA que sirvió para alinear el English Norwegian Parallel Corpus (Hofland y Johansson, 1998).
   b. Formatos: XCES vs TMX o ambos
      El formato de codificación de estos textos alineados será uno de los dos estándares desarrollados para tal fin como XCES o TMX. La ventaja del primero es que es más sólido cuando de codificar información lingüística se trata, mientras que el segundo es el estándar que se emplea en la industria para alimentar memorias de traducción. (Gómez Guinovart, 2005).

4. Enriquecimiento del etiquetado
    a. Inclusión de más datos

       Como ya se ha comentado anteriormente, la información que se ha etiquetado es la que aparecía en el propio texto original. En algunos casos la información que se ha podido extraer del propio texto es insuficiente. Para ello habrá que incorporar información, procedente de bases de datos, al corpus, para completarlo y homogeneizar la información proporcionada en cada corpus.
    b. POS

       Se realizará un etiquetado morfológico, puesto que un gran número de investigadores así lo demanda. No obstante, en principio esto no sería necesario dada la metodología empleada por los investigadores miembros de ECPC.
    c. Etiquetado temático

       Cuando las tecnologías asociadas a la web semántica estén más desarrolladas tal vez sea posible etiquetar por temas las distintas intervenciones parlamentarias para afinar aún más la investigación.
5. Desarrollo de la herramienta ConcECPC 2.0
    a. Incorporar la posibilidad de generar concordancias paralelas bilingües e incluso multilingües.
    b. Integrar la Suite «R» como un plugin para realizar cálculos estadísticos y representarlos de forma gráfica.
    c. Hacer posible la búsqueda de colocaciones, clusters y n-grams.
    d. Generar listados de frecuencias para poder comparar otros corpus con ECPC (bien como corpus de estudio o como corpus de referencia).
6. Desarrollo del portal de consulta del corpus

En cuanto a la parte teórica, ya se están desarrollando una serie de parámetros contrastivos que se pondrán en práctica para realizar estudios de naturaleza contrastiva.

BIBLIOGRAFÍA

Bowker, L. (1998): «Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study», *Meta*, vol. 43, núm. 4, pp. 631-651.

Bowker, L. (2003) «Corpus-based Applications for Translator Training: Exploring possibilities», en Granger, S., J. Lerot y S. Petch-tyson (2003): *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Ámsterdam/Filadelfia, Rodopi, pp. 169-183.

Bowker, L. y J. Pearson (2002): *Working with Specialized Language. A practical Guide to Using Corpora*, Londres y Nueva York, Routledge.

Calzada Pérez, M. (2005): «Corpus electrónicos como herramientas de documentación y formación para traductores», en Sales Salvador, D. (2005): *La Biblioteca de Babel: Documentarse para traducir*, Granada, Comares, Colección Interlingua, pp. 163-199.

Calzada Pérez, M. (2005): «Vygotsky, Wordsmith y GENTT: Corpus y herramientas electrónicas en la clase de traducción», en García Izquierdo, I. (2005): *El género textual y la traducción. Reflexiones teóricas y aplicaciones pedagógicas*, Berna, Peter Lang, 2005, pp. 189-206.

Calzada Pérez, M. y S. Luz (2006): «ECPC: Technology as a tool to study the (linguistic) functioning of national and trans-national European parliaments» *Journal of Technology, Knowledge and Society*, 5(2), pp. 53-62.

Gómez Guinovart, X. (2005): «Procesamiento y aplicaciones de los corpus paralelos», *Novática*, 175, pp. 50-54.

Gregory, M y S. Carroll (1978): *Language and Situation. Language Varieties and their Social Contexts*, Londres, Routledge & Kegan Paul.

Hofland, K. y S. Johansson (1998): «The Translation Corpus Aligner: A program for automatic alignment of parallel texts», en Johansson, S. y S. Oksefjell (1998): *Corpora and Cross-linguistic research: Theory, Method, and Case Studies*, Ámsterdam/Atlanta, Rodopi, pp. 87-101.

Hurtado Albir, A. (2001): *Traducción y Traductología. Introducción a la traductología*, Madrid, Cátedra.

Johansson, S. (2003): «Reflections on Corpora and their uses in Cross-linguistic Research», en Zanettin, F. et al. (2003): *Corpora in translator education*, Manchester, UK; Northampton, MA, St. Jerome, pp. 135-144.

Kelly, D. (2005): *A handbook for translator trainers: a guide to reflective practice*, Manchester, St. Jerome Publishing.

Kübler, N. (2003): «Corpora and LSP Training», en Zanettin, F. et al. (2003): *Corpora in translator education*, Manchester, UK; Northampton, MA, St. Jerome, pp. 25-42.

Laviosa, S. (1998): «The Corpus-based Approach: A New Paradigm in Translation Studies», *Meta*, vol. 43, núm. 4, pp. 474-479.

Laviosa, S. (2002): *Corpus-based Translation Studies. Theory, Findings and Applications*, Ámsterdam/ Nueva York, Rodopi.

Luz, S. y Baker M. (2000): «TEC: A toolkit and API for distributed corpus processing», en Bird, S y G. Simmons (eds.) (2000): *Proceedings of Exploration-2000: Workshop on Web-Based Language Documentation and Description*, Filadelfia, University of Pennsylvania, pp. 108-112.

Pearson, J. (2000): «Teaching terminology using electronic resources», en Botley, S., T. McEnery et al. (2000): *Multilingual corpora in teaching and research*, Ámsterdam, Rodopi, pp. 92-105.

Sinclair, J. (1991): *Corpus, concordance and collocation*, Oxford, OUP.

Toury, G. (1985): «A Rationale for Descriptive Translation Studies», en Hermans, T. (ed.) (1985): *The manipulation of literature*, Londres, Croom Hlem, pp. 24-38.

Toury, G. (1995): *Descriptive Translation Studies and Beyond*, Ámsterdam/Filadelfia, John Benjamins Publ Company.

Zanettin, F., S. Bernardini et al. (2003): *Corpora in translator education*. Manchester, UK; Northampton, MA, St. Jerome.