

**GLORIA CORPAS PASTOR**

## **CORPUS, TECNOLOGÍA Y TRADUCCIÓN**

### **1. Introducción**

No es casualidad que la Lingüística de Corpus floreciese especialmente en el contexto europeo. Recordemos que la investigación en tecnologías lingüísticas (o “industrias de la lengua”) ha sido el marchamo de las políticas científicas europeas.<sup>1</sup> Desde ahí se ha favorecido la investigación en tecnologías lingüísticas como forma de salvaguardar, por un lado, la diversidad cultural y el multilingüismo de Europa, y, al mismo tiempo, superar las barreras y dificultades que esto supone para poder alcanzar los objetivos comunes a todos los europeos. Multilingüismo, multiculturalidad, traducción, tecnologías son rasgos inherentes a la sociedad europea actual. Se podría decir, además, que estas características definitorias han contribuido decisivamente al desarrollo de aplicaciones y recursos lingüísticos encaminados dar soporte a las políticas sociales europeas, y sus estribaciones en materia de comercio, educación e investigación.

Si bien las tecnologías lingüísticas y el corpus se han abierto camino desde época muy temprana en las vertientes teóricas y aplicadas de la Lingüística, han sido necesarias varias décadas para que traductores e intérpretes se hayan subido por fin a este carro, que ya iba repleto de investigadores de otras disciplinas afines. En este trabajo realizaremos un breve excursión por lo que ha supuesto la incorporación de tales recursos y herramientas para el ámbito de la traducción y la interpretación, con especial referencia a las tecnologías propias del sec-

---

<sup>1</sup>Para una visión de conjunto sobre las políticas científicas europeas en materia de tecnologías lingüísticas, véase Corpas Pastor (2008).

tor, la formación de profesionales y la investigación en Traductología. Nuestro objetivo es ofrecer al lector no especialista una primera aproximación al tema. No pretendemos, pues, ser exhaustivos, sino solo ilustrativos del panorama actual y de su historia más reciente.

## 2. Tecnologías lingüísticas

Antes de continuar, se hace necesario aclarar tres conceptos clave en torno a los cuales gira esta sección. En primer lugar, entendemos por *corpus* una colección de textos en soporte digital, que ha sido compilada conforme a unos criterios de diseño apropiados que le confieren representatividad (EAGLES, 1996). En segundo lugar, entendemos por *tecnología lingüística* “la derivada de aplicar los conocimientos sobre la lengua al desarrollo de sistemas informáticos”<sup>2</sup>. Y en tercer lugar, entendemos por *tecnología de la traducción* un tipo de tecnología lingüística, monolingüe o multilingüe, diseñada para formar parte de los entornos de trabajo del traductor (Bowker y Corpas Pastor, 2012/En preparación).

Las tecnologías lingüísticas, también denominadas “industrias de la lengua”, comprenden tanto la investigación teórica como la investigación aplicada, esto es, tecnología y desarrollos (Sager, 1993: 10-18). En el primer caso se refiere a la Lingüística Computacional (LC) y en el segundo, al Procesamiento de Lenguaje Natural (PLN) y los lenguajes documentales y (semi-)artificiales. La Lingüística Computacional teórica<sup>3</sup> se centra en la descripción, modelado y representación computacional de las lenguas naturales que servirá posteriormente de base para la construcción de diccionarios automatizados en soporte electrónico, gramáticas analíticas y sintéticas, etc. En esta vertiente teórica se tienen en cuenta, así mismo, los resultados de la lingüística y la semiótica con objeto de crear nuevos lenguajes documentales y de clasificación, interlenguas e interfaces en lenguaje natural. La otra vertiente está representada por las aplicaciones tecnológicas y se subdivide, a su vez, en dos: a) producción, extracción, almacenamiento y procesamiento de lenguajes ya existentes y b) desarrollo y producción de nuevos lenguajes. Las primeras se refieren al PLN, incluidos los sublenguajes, como los que se utilizan en el desarrollo de videojuegos, en las ciencias de la información, etc., o la corrección y ayuda a la redacción de documentos (corrección ortográfica, gramatical, de estilo, diccionarios en soporte informático y editores de lenguaje controlado). Las segundas se refieren a las lenguas (semi-)artificiales de los sistemas documentales de recuperación y almacenamiento de información, el resumen documental y la consulta a bases de datos en lenguaje natural, y a los lenguajes controlados que se utilizan en la industria aeronáutica o en la investigación en física nuclear, entre otros.

Como hemos señalado más arriba, las políticas científicas europeas en materia de tecnologías lingüísticas han ocupado desde el principio un papel central en los distintos Programas Marco establecidos hasta la fecha. En un primer momento, la investigación básica se centró en los distintos recursos lingüísticos existentes o de nueva creación para las distintas lenguas europeas. Entre estos recursos multilingües destacan los corpus lingüísticos y las herramientas

---

<sup>2</sup> <http://oesi.cervantes.es/oesi/tls.jsp>

<sup>3</sup> Sobre los distintos ámbitos de la Lingüística Computacional y las tecnologías lingüísticas, véase Lavid (2005) y los trabajos reunidos en Mitkov (2003).

para su tratamiento (codificación, análisis, etiquetado y explotación). Precisamente ésta fue la misión del grupo EAGLES,<sup>4</sup> constituido en el seno del programa de Ingeniería Lingüística del IV Programa Marco (1994-1998), cuyo objetivo era establecer normas de codificación y anotación de corpus con vistas a la reutilización de recursos lingüísticos, una de las prioridades de la Comisión en aquellos primeros momentos.

Pero la relevancia del corpus no se redujo a su empleo como simple recurso lingüístico. El diseño, compilación, anotación, gestión y explotación de los corpus trajo consigo consecuencias muy importantes para el desarrollo de la Lingüística Computacional,<sup>5</sup> tanto en lo referente a sus fundamentos teóricos, como a sus aplicaciones y técnicas, además de afectar también al inventario de recursos propios. Desde el punto de vista de la ingeniería lingüística y el PLN, la Lingüística de Corpus tuvo una enorme repercusión para el desarrollo de sistemas de análisis morfosintáctico, lexicones computacionales y diccionarios computerizados,<sup>6</sup> sobre todo en los primeros tiempos.

Además, el corpus ha sido el recurso lingüístico por excelencia que ha determinado, por un lado, el desarrollo de técnicas y herramientas de tratamiento (codificación, análisis, etiquetado y explotación); y, por el otro, ha propiciado una profunda revisión de los fundamentos teóricos, las aplicaciones e, incluso, el inventario de recursos propios. Tal es la importancia de este recurso que en un principio la Lingüística de Corpus se llegó a considerar como una rama de la Lingüística Computacional (cf. Oostdijk, 1991: 2).

Por otra parte, desde el ámbito del PLN se han llevado a cabo multitud de investigaciones y desarrollos que han supuesto, a su vez, un importante avance en los métodos de análisis, gestión y explotación de esos mismos corpus de los que se estaban sirviendo como recurso primario o bien como base de evaluación.<sup>7</sup> La gestión automática de corpus textuales conlleva una serie de operaciones previas, como la atomización, el análisis morfológico y la lematización, la segmentación en frases y la resolución del límite de frase. El etiquetado de corpus trajo consigo también el establecimiento de estándares de anotación, según las recomendaciones de EAGLES.

---

<sup>4</sup> *Expert Advisory Group on Language Engineering Standards* (<http://www.ilc.cnr.it/EAGLES/home.html>).

<sup>5</sup> En el apartado anterior hemos abordado el importante papel que ha tenido el corpus como recurso y como impulsor de los avances en tecnologías lingüísticas de la LC y el PLN. Sin embargo, éstos no han sido los únicos ámbitos de referencia. La metodología de investigación con corpus ha supuesto un revulsivo también para otros estudios lingüísticos, especialmente los de tipo aplicado. Sobre los usos del corpus en lingüística aplicada, véase Hunston (2002: 96 y ss.), McEnery et al. (2006: 80 y ss.), así como los trabajos reunidos en Lüdeling y Kytö (2008-2009).

<sup>6</sup> Según Ooi (1998: 1-2), en la lexicografía basada en corpus convergen la Lingüística Computacional, la Lexicografía Computacional y la Lingüística de Corpus. Si bien la Lingüística Computacional aspira a crear lexicones formales apropiados para la implementación de sistemas de PLN, la Lexicografía Computacional se ocupa de dotar a los diccionarios electrónicos comerciales de un formato suficientemente explícito para poder ser utilizados en Lingüística Computacional o bien busca automatizar las tareas lexicográficas y diseñar diccionarios (asistidos) por ordenador. La Lingüística de Corpus, por su parte, se centra en la compilación electrónica de colecciones representativas de textos extraídos del discurso real. Las tres disciplinas comparten un mismo interés, esto es, utilizar el lenguaje natural para extraer información y construir bases de datos léxicas, almacenar el lexicón computacional y amoldar la descripción lingüística a éste, emplear lexicones enriquecidos para PLN, definir estándares de intercambio léxico y reutilizar recursos, de forma que se puedan unir esfuerzos y optimizar los resultados de la investigación.

<sup>7</sup> Los métodos y técnicas propios de la investigación básica de corpus constituirían la vertiente teórica de la Lingüística de Corpus.

Tales métodos se han aplicado posteriormente en diversos campos de la Lingüística Computacional, concretamente en tareas intermedias como el desarrollo de correctores ortográficos y gramaticales, o la implementación de herramientas para el análisis morfosintáctico del corpus, la desambiguación léxica y la alineación de corpus paralelos, entre otras. Pero también se utiliza el corpus para una variedad de tareas finales, esto es, aplicaciones que hacen uso de las tareas intermedias a fin de conseguir resultados prácticos y de interés general.

La gestión de corpus tiene un papel fundamental en la arquitectura de muchas aplicaciones de PLN, ya sean monolingües o multilingües. Además de los usos referidos hasta el momento, merece la pena señalar otros muy relevantes, como son la generación de lenguaje natural (GLN), la producción de documentación multilingüe y redacción técnica, la extracción de terminología, la clasificación documental, las técnicas de PLN basadas en corpus para el tratamiento de preguntas-respuesta, y otras aplicaciones de la minería de datos, como el análisis de opiniones y sentimientos. También los corpus orales desempeñan una importante función en otras aplicaciones, como la conversión de texto a habla, los sistemas de traducción habla-habla (es decir, la traducción oral automática) y los sistemas de diálogo<sup>8</sup>.

Tanto la LC como el PLN se sirven del corpus para la evaluación de sus aplicaciones y sistemas (McEnery, 2003). El corpus forma parte integrante de los protocolos actuales de evaluación de *software* y una de sus utilidades básicas, es, precisamente, la evaluación de la calidad del *output* de los sistemas de TA (traducción automática). Además, el alineamiento de corpus paralelos es una pieza clave de los sistemas de TA y MT (memorias de traducción). Actualmente se tiende hacia la hibridación de tales sistemas, como es el caso de la traducción automática basada en ejemplos (EBMT),<sup>9</sup> basada en corpus y con una fuerte base estadística.<sup>10</sup> Dichos sistemas presentan similitudes con otras herramientas diseñadas específicamente para traductores, que también hacen uso de corpus bilingües alineados, donde los ejemplos previamente traducidos sirven de modelo al usuario para la traducción en curso. Nos referimos a los sistemas de memorias de traducción, del tipo Trados, Déjà Vu, MultiTrans, Star Transit, WordFast, etc., por citar sólo algunas de las más conocidas y utilizadas. Las memorias de traducción no sólo se valen del corpus, sino que son en sí mismas corpus bilingües o multilingües alienados.

Los sistemas de TA y MT forman parte de las denominadas tecnologías de la traducción, surgidas a la luz de los avances de la Lingüística Computacional, mediante las cuales se automatiza la actividad translativa. Pero los traductores emplean también otros recursos y tecnologías lingüísticas en su actividad cotidiana. Nos referimos a los procesadores de texto, los correctores ortográficos y gramaticales, así como los recursos electrónicos (por ejemplo,

---

<sup>8</sup> Los sistemas de diálogo son aplicaciones informáticas, ligadas a corpus de intercambios conversacionales, capaces de establecer comunicación con seres humanos para realizar algún tipo de actividad. Sobre este tema, véase McTear (2004).

<sup>9</sup> Sobre este tema, véanse los trabajos reunidos en Hansen e Yvon (2010).

<sup>10</sup> Merece la pena señalar el sistema *METIS: Statistical Machine Translation Using Monolingual Corpora I y II* (UE, n° de ref. IST-2001-32775), que aplica una metodología de explotación de corpus comparables para la traducción automática. Se trata de un sistema de traducción basado en corpus, de tipo EBMT y, por tanto, híbrido (estadístico y de reglas), cuya particularidad reside en que no utiliza corpus paralelos alineados, como viene siendo usual, sino corpus monolingües.

las obras de referencia y los repertorios lexicográficos en cd-rom) y las herramientas telemáticas (uso de correo electrónico, de la red mundial, de bases de datos, etc.), que conformarían el nivel básico de automatización. En un nivel superior estarían las herramientas de traducción asistida, esto es, las tradicionales memorias de traducción y las herramientas para capturar datos (reconocimiento óptico de caracteres, sistemas de reconocimiento de voz, sistemas de texto-habla), así como los programas de localización y traducción de páginas web, las herramientas de autor (*authoring tools*), los programas de gestión de terminología (bases de datos, extractores, etc.) y los programas de gestión de corpus (lematizadores, extractores de listas de palabras, generadores de concordancias, etc.). Y, por encima del anterior se situarían los sistemas de traducción automática, junto a los sistemas multilingües de generación de lenguaje natural y de resúmenes. Un último nivel de automatización estaría integrado por los sistemas de gestión de proyectos, divididos en sistemas orientados a empresas, orientados al trabajo colaborativo y orientados al trabajo individual (Sargent, 2007).

Todos esos recursos, muchos de los cuales hacen uso de corpus, son de gran utilidad para realizar distintas tareas asociadas a un encargo de traducción dado. Sin embargo, no todos ellos se consideran herramientas de traducción asistida (TAO) o tecnologías de la traducción. Sólo se considera como tales aquellas herramientas diseñadas específicamente para la traducción propiamente dicha, a diferencia de otras herramientas o aplicaciones pensadas para otros usos generales y que sí entrarían en la categoría de aplicaciones telemáticas o de PLN (como, por ejemplo, los procesadores de texto, los correctores ortográficos, el correo electrónico o los sistemas de gestión de proyectos) o de Lingüística Computacional (diccionarios electrónicos, sistemas de generación de lenguaje natural, sistemas de generación de resúmenes, herramientas de autor, etc.).

Es decir, por regla general, cuando hablamos de tecnologías de la traducción<sup>11</sup> nos referimos, concretamente, a sistemas integrados en un entorno de trabajo del traductor (*translator's workstation* o *workbench*), de forma que los distintos componentes interactúan entre sí o bien permiten su utilización secuencial y en cadena (Somers, 2003b). Dichos entornos constan como mínimo de un sistema de memoria de traducción y un gestor de terminología asociado. Adicionalmente pueden contar con sistemas de extracción y reconocimiento terminológicos, sistemas de gestión de corpus y concordancias, sistemas de localización, e, incluso sistemas de traducción automática y de gestión de proyectos.

Con variaciones, éste sería un entorno de trabajo típico. En primer lugar, el módulo de análisis compara la similitud entre el texto que se ha de traducir y los corpus ya existentes. Ello permite, entre otras cosas, estimar la dificultad del trabajo y calcular el presupuesto en base a ello. Las memorias de traducción hacen uso de corpus paralelos alineados que rastrean una base de datos de bitextos para mostrar al traductor segmentos coincidentes previamente traducidos y reutilizables. Cuando en la base de datos no se encuentran coincidencias, el módulo de traducción automática proporciona una posible traducción, normalmente basada en una búsqueda estadística que utiliza corpus paralelos y comparables. Al mismo tiempo, los programas de concordancias permiten al traductor realizar búsquedas por debajo del nivel de

---

<sup>11</sup> Sobre tecnologías de la traducción véase Austerlühl (2001), Bowker (2002), Quah (2005) y los trabajos reunidos en Somers (2003a).

sentencia tanto en los corpus paralelos que alimentan las memorias como en cualquier corpus bilingüe previamente alineado. Los extractores de terminología suelen operar de forma similar, con la salvedad de que sugieren candidatos a términos y términos equivalentes automáticamente. Los gestores de terminología van almacenando la información terminológica, la cual se introduce en bases de datos a la cual tiene acceso el traductor, y a partir de las cuales se realiza el reconocimiento y propagación automática de terminología por el sistema. El módulo de control de calidad comprueba la precisión y corrección terminológica con respecto a las especificaciones del cliente, la consistencia interna o el uso de lenguajes controlados. Este módulo también puede incluir correctores ortotipográficos y gramaticales. Finalmente, el módulo de gestión de proyectos permite mantener información actualizada en todo momento sobre clientes, trabajos concretos y archivos asociados a éstos.

Todos estos módulos y sistemas que integran el entorno de trabajo del traductor hacen uso en diversa medida del corpus y en diversa medida también necesitan de intervención humana, según su distinto grado de automatización. Tales herramientas no se conciben como sustitutivas de la traducción humana, como se pretendía tras la II Guerra Mundial, sino como ayuda a la traducción humana. El interés por las tecnologías de la traducción surge por la demanda laboral, cuya fisonomía ha cambiado tras los procesos de globalización e internacionalización recientes, así como por las exigencias de rapidez y calidad a bajo coste de los mercados. Además, la documentación técnica multilingüe, la localización de contenidos digitales (páginas web, aplicaciones informáticas, etc.), la automatización del aseguramiento y el control de la calidad, etc., hacen que en algunos casos el uso de determinadas tecnologías de la traducción se convierta casi en un imperativo. De hecho, el mercado laboral demanda profesionales bien formados en tecnologías lingüísticas multilingües orientadas a ayudar al traductor en su quehacer diario.

A pesar de la necesidad real de preparar profesionales con un buen conocimiento y dominio de las tecnologías propias del sector, la formación académica de traductores e intérpretes ha ido, tradicionalmente, muy a la zaga. Una iniciativa pionera fue el proyecto *Language Engineering for Translators* (LETRAC),<sup>12</sup> subvencionado por la DGXIII de la Comisión Europea de enero de 1998 a marzo de 1999. Entre los contenidos propuestos por LETRAC se encontraban, precisamente, la gestión y el procesamiento de corpus lingüísticos. Y se recomendaba que la formación en tecnologías lingüísticas comience por la gestión de corpus como fase previa, pues éste forma parte de las arquitecturas de tales sistemas.

Por otro lado, la propia Norma europea UNE-EN 15038:2006 sobre calidad en la prestación de servicios de traducción incluye las tecnologías de la traducción como parte integrante de la gestión de proyectos. Dicha Norma, además, plantea la necesidad de que los traductores

---

<sup>12</sup> En su seno se confeccionó un listado de elementos curriculares relacionados con las tecnologías lingüísticas para adecuar los estudios de traducción al mercado laboral. Además de módulos sobre los componentes básicos de un ordenador, los distintos tipos de hardware y software necesarios para la gestión de datos, en las recomendaciones finales de LETRAC se proponía formar al alumno en el manejo de los programas de edición, las telecomunicaciones, Internet y los lenguajes de programación, y, en un nivel superior, en la gestión de proyectos, los lenguajes controlados en entornos industriales, los programas de gestión terminológica, la traducción automática, las memorias de traducción, la localización y la Lingüística de Corpus. URL: <http://www.iai.uni-sb.de/iaien/en/letrac.htm>.

desarrollen un alto grado de competencia tecnológica, que es aquella que se refiere a la capacidad y la técnica requeridas para la preparación y producción profesional de traducciones.

El uso de tecnologías de la traducción —entre las cuales figura, como ya se ha dicho, la compilación, gestión y explotación de corpus— supone un alto grado de competencia tecnológica por parte de sus usuarios, que, a su vez, contribuye a la profundización en las cuatro competencias restantes que la norma mencionada considera que deben estar presentes en un traductor profesional, a saber: a) la competencia traductora;<sup>13</sup> b) la competencia lingüística y textual en la lengua de origen y en la lengua de destino;<sup>14</sup> c) la competencia cultural;<sup>15</sup> y d) la competencia documental, adquisición y procesamiento de la información, esto es, la capacidad de adquirir de forma eficaz el conocimiento necesario para comprender el texto de origen y producir el texto de destino.<sup>16</sup> La competencia documental requiere también experiencia en el uso de recursos documentales y la capacidad de desarrollar estrategias adecuadas para el uso eficaz de las fuentes de información disponibles. El uso de corpus para traducción e interpretación, especialmente el tipo *ad hoc*, requiere, además, un alto grado de *competencia tecnológica informacional*<sup>17</sup> por parte del traductor-usuario-compilador.

### 3. Formación e investigación

En los últimos años se está prestando una considerable atención a las tecnologías de la traducción y al corpus como parte de las competencias y resultados de aprendizaje reflejados en los planes de estudios.<sup>18</sup> De ello y de las estribaciones del corpus en las investigaciones de tipo descriptivo nos vamos a ocupar en esta sección.<sup>19</sup>

---

<sup>13</sup> “Competencia traductora: la competencia traductora es la capacidad para traducir textos al nivel requerido. Incluye la capacidad de valorar los problemas de comprensión y producción textual, así como la capacidad para verter el texto a la lengua de destino de acuerdo con lo convenido entre el cliente y el proveedor de servicios de traducción, y de justificar los resultados.” (UNE-EN 15038:2006).

<sup>14</sup> “Competencia lingüística y textual en la lengua de origen y en la lengua de destino: la competencia lingüística y textual incluye la capacidad para comprender la lengua de origen y dominar la lengua de destino. La competencia textual requiere el conocimiento de las convenciones del tipo de texto para una gama lo más amplia posible de textos en lenguaje estándar y especializado e incluye la capacidad para aplicar estos conocimientos en la producción de textos.” (UNE-EN 15038:2006).

<sup>15</sup> “Competencia cultural: la competencia cultural incluye la capacidad para utilizar la información sobre convenciones locales, normas de comportamiento y sistemas de valores que caracterizan las culturas de origen y de destino.” (UNE-EN 15038:2006).

<sup>16</sup> “Competencia documental, adquisición y procesamiento de la información: la competencia documental incluye la capacidad de adquirir de forma eficaz el conocimiento de las convenciones del tipo de texto para una gama lo más amplia posible de textos en lenguaje estándar y especializado adicional, necesario para comprender el texto de origen y producir el texto de destino. La competencia documental requiere también experiencia en el uso de recursos documentales y la capacidad de desarrollar estrategias adecuadas para el uso eficaz de las fuentes de información disponibles.” (UNE-EN 15038:2006).

<sup>17</sup> La competencia tecnológica y la competencia documental están, pues, muy relacionadas, especialmente cuando se hace uso de recursos documentales electrónicos y fuentes de información en línea. De hecho, sería preferible hablar de *competencia tecnológica informacional* para referirnos a la capacidad de seleccionar y manejar de forma adecuada las tecnologías de la traducción y la documentación electrónica necesarias para la preparación, producción, gestión y evaluación profesional de las producciones de traductores e intérpretes.

<sup>18</sup> Consúltense los trabajos de Torres del Rey (2003, 2005) y Bouillon y Starlander (2009) acerca de la formación tecnológica de traductores e intérpretes.

<sup>19</sup> Kruger, Wallmach y Munday (2011) recogen una actualizada selección de trabajos sobre aplicaciones del corpus en traducción.

Los usos “pedagógicos” del corpus en Traducción abogan por esta metodología de análisis para el desarrollo integral de la competencia traductora, la mejora de la docencia y la adecuación al mercado laboral. Son herederos de otros desarrollos de la Lingüística Aplicada, que son anteriores en el tiempo. Una de las aplicaciones más relevantes y conocidas del corpus ha sido la producción de gramáticas y diccionarios para la enseñanza de lenguas. El enfoque novedoso de los diccionarios y gramáticas COBUILD (Collins Birmingham University International Language Database), creados en la década de los ochenta en la Universidad de Birmingham bajo la dirección de John M. Sinclair,<sup>20</sup> consolidaron al corpus como materia prima para la descripción y análisis del uso real de la lengua en el discurso. En estrecha relación con lo anterior está la utilización del corpus para la enseñanza y, más concretamente, para la enseñanza de lenguas asistida por ordenador, en inglés *computer-assisted language learning* (CALL).

Las primeras aplicaciones del corpus para la enseñanza de idiomas tuvieron lugar en la Universidad de Birmingham a principios de los noventa, de la mano de Tim Johns, el padre del enfoque DDL (*data-driven learning*). Johns aplicó las técnicas de explotación de corpus y el análisis de concordancias para enseñar EFL (*English as a Foreign Language*) a los estudiantes internacionales que acudían a la citada Universidad. En sus comienzos, el DDL estuvo inspirado en los trabajos desarrollados por el equipo COBUILD en el ámbito de la enseñanza de lenguas. Concretamente nos referimos al Enfoque Léxico (*Lexical Approach*), propuesto por Sinclair y Renouf (1988), y que se concretaría en el innovador *Collins COBUILD English Course* (Willis y Willis, 1989), cuya selección de materiales y secuenciación de contenidos estaba basada en el análisis del corpus BCET. Poco después, Willis (1990) convertiría esta novedosa corriente didáctica en un auténtico enfoque teórico.

El Enfoque Léxico y el DDL compartían un mismo objetivo: proporcionar al alumno un entorno real de uso que le permitiese contextualizar y, de paso, desmitificar la lengua a través de los datos lingüísticos que aporta el corpus. Ambos enfoques se diferencian, sin embargo, en la forma en la que se utilizan dichos datos. Para el equipo COBUILD, el corpus permite describir el uso real de la lengua y ofrece, por tanto, un material de primera para la investigación y el diseño curricular, que, a su vez, se traduce en una mejora de la enseñanza y de las obras de referencia correspondientes. En el DDL, por el contrario, el trabajo de corpus se introduce directamente en el aula, pues se parte de la hipótesis de que el descubrimiento autónomo a partir de los datos reales que arroja el corpus motiva al alumno, facilita la deducción de significado a partir del contexto y refuerza el proceso de enseñanza-aprendizaje.<sup>21</sup>

---

<sup>20</sup> Este proyecto tuvo su origen en uno anterior, denominado *English Lexical Studies*, iniciado también por Sinclair en la Universidad de Edimburgo, a principios de los sesenta, y desarrollado posteriormente en la Universidad de Birmingham con el nombre de *Birmingham Collection of English Texts* (BCET), que más tarde pasaría a formar parte del *Bank of English* (BoE).

<sup>21</sup> El uso de corpus en la enseñanza de idiomas permite ampliar por descubrimiento los conocimientos que se poseen sobre una lengua y, al mismo tiempo, ayuda a consolidar los ya adquiridos, aumenta la consciencia lingüística y favorece el aprendizaje autónomo, sí como el aprendizaje incidental y casual, estimulando la curiosidad intelectual de los estudiantes. Así mismo, el uso de corpus contribuye al desarrollo global de la competencia discursiva y socio-comunicativa en L2 y mejora las destrezas orales y escritas, tanto en referencia a la lengua general como a la especializada (cf., por ejemplo, Gavioli, 2005; y los trabajos reunidos por Ghadessy *et al.*, 2001, y Connor y Upton, 2004).

El enfoque DDL se inspira en la teoría de la inferencia estadística consistente en aquellos métodos mediante los cuales se realizan generalizaciones acerca de la muestra de observaciones experimentales con la que se trabaja. Aplicado a la enseñanza, se trata de que los alumnos infieran y contrasten las generalizaciones extraídas de concordancias relativas a un determinado fenómeno lingüístico. Ello se puede llevar a cabo mediante procedimientos inductivos o deductivos. En el primer caso, el alumno observa, clasifica, generaliza y construye hipótesis explicativas a partir de los patrones que descubre en las líneas de concordancia analizadas. Es lo que se denomina “aprendizaje por descubrimiento” (en inglés, *discovery learning*). En el segundo caso, el alumno parte de generalizaciones previamente adquiridas para, posteriormente, clasificar los datos procedentes del corpus, contrastarlos con las reglas ya aprendidas y, como consecuencia de ello, consolidar y precisar aún más los conocimientos adquiridos.<sup>22</sup>

Que el corpus se haya “colado” en las aulas de traducción no es de extrañar, habida cuenta que, como ya se ha dicho, desde finales de los ochenta se empleaba este recurso en la enseñanza de idiomas. La enseñanza de la traducción y la interpretación con corpus ha sido, pues, otro exponente más de esa tendencia generalizada. Recordemos que en Italia el enfoque DDL contó con un importante foco de desarrollo. Concretamente en la Scuola Superiore di Lingue Moderne per Interpreti e Traduttori de la Universidad de Bolonia en Forlì (Italia) se comenzó a trabajar en esta línea desde mediados de los noventa (cf. Aston, 1995), aunque la producción científica correspondiente no alcanzó su cenit hasta unos años más tarde (cf. los trabajos de Bernardini, 2000, Bernardini y Zanettin, 2000; y Aston, 200c). Desde la Facultad de Traducción de Bolonia se utilizaba corpus para la enseñanza de segundas lenguas, de cursos de especialidad y, posteriormente, para la traducción inversa.

La enseñanza de la traducción con corpus es un área bien consolidada en el seno de la disciplina,<sup>23</sup> que es reflejo de la investigación realizada en otros ámbitos afines y que aprovecha resultados previos conseguidos con esta misma metodología en el campo de la enseñanza de lenguas y el análisis de errores. Un ejemplo relativamente reciente lo tenemos en el proyecto *Multilingual eLearning in LANGuage Engineering* (MeLLANGE, Leonardo da Vinci, 2004-2007), orientado al empleo de corpus para la formación de los futuros profesionales de la traducción. En concreto, en este proyecto se pretende crear una metodología colaborativa para la elaboración de materiales didácticos virtuales basados en corpus tanto para la enseñanza de la traducción como para la gestión de recursos lingüísticos y las tecnologías lingüísticas.

---

<sup>22</sup> Los últimos desarrollos del enfoque DDL han utilizado otros tipos de corpus también. Por ejemplo, Johns (2002) menciona explícitamente el uso de corpus bilingües, formados por textos originales y sus traducciones a otras lenguas (esto es, corpus paralelos), en el contexto del denominado *reciprocal learning* o sistema de aprendizaje en tándem (*tandem learning*). También se han utilizado corpus monolingües de L2, es decir, textos producidos por estudiantes de segundas lenguas. En este caso se busca la comparación de las características de la producción oral y escrita de los estudiantes de segundas lenguas con las características de esos mismos tipos de textos, pero producidos por hablantes nativos en circunstancias comunicativas prototípicas. Quizá el máximo exponente en este ámbito sea Sylviane Granger, de la Universidad de Louvain (Bélgica), y sus trabajos sobre el ICLE (*International Corpus of Learner English*).

<sup>23</sup> El interés suscitado por las posibles aplicaciones del corpus para la enseñanza de la traducción se ha visto reflejado también en los congresos CULT (*Corpus Use and Learning to Translate*), que se vienen celebrando desde finales de los noventa y que, hasta la fecha, han contado con tres ediciones: 1997 (Bertinoro), 2000 (Bertinoro) y 2004 (Barcelona).

Así pues, se ofrecen módulos de enseñanza de la traducción a distancia mediante corpus monolingües y bilingües.

Los corpus creados por traductores constituyen uno de los tipos de corpus que pueden ser utilizados de forma didáctica, según la clasificación de Bowker (2003). Se asemejan a los corpus de aprendizaje (*learner corpus*), los cuales contienen producciones de los alumnos y son utilizados en la enseñanza de segundas lenguas. En segundo lugar está el corpus creado para traductores. Suelen ser corpus virtuales *ad hoc*. Este tipo de corpus no contiene traducciones, sino textos que sirven de recurso a los traductores para adquirir conocimiento sobre un campo de especialidad concreto, así como la terminología y las convenciones estilísticas propias de la especialidad. Un tercer tipo es el corpus comparable, que sirve de referencia al profesor a la hora de evaluar la calidad de las traducciones realizadas por los estudiantes y poder proporcionar a los alumnos un *feedback* objetivo que les ayude a mejorar y progresar en su formación. Así mismo, permite observar de manera indirecta el comportamiento traductor de los alumnos.

Los corpus comparables están compuestos por corpus de textos no traducidos, monolingües —en la lengua meta y equiparables al TO—, o bilingües (en las lengua meta y de origen), que son equiparables al TO y entre sí, por cuanto presentan un diseño y composición similares. Cuando el corpus comparable se compila en lengua meta, de acuerdo con las especificaciones o necesidades de un encargo, proyecto o bloque de traducción concreto, se denomina corpus *ad hoc* (cf. Corpas Pastor, 2001). Generalmente se trata de corpus “virtuales”, especializados y de pequeño tamaño, compilados a partir de textos procedentes de la red Internet. El diseño de este tipo de corpus responde a la necesidad de cubrir las necesidades informativas del traductor con respecto a un determinado encargo y en cualquier dirección.<sup>24</sup>

Los corpus comparables proporcionan a los usuarios un inventario amplio de formas convencionales de expresar unidades de sentido y funciones específicas en un determinado registro y forma textual en la lengua de origen y en la lengua meta, las cuales se pueden considerar equivalentes (en diverso grado). Por otra parte, se muestran claramente superiores a otros recursos tradicionales, como los diccionarios, pues están al día, contienen neologismos, ofrecen información contextualizada sobre los términos, etc.

Pero los beneficios de utilizar corpus comparables para la enseñanza-aprendizaje de la traducción van más allá del puro trabajo terminológico. Afectan positivamente tanto a la comprensión del texto de origen como la producción del texto meta. Por un lado, los alumnos consiguen producir traducciones más idiomáticas y cercanas a la norma de la cultura meta. El corpus permite, pues, satisfacer las necesidades documentales específicas derivadas del pro-

---

<sup>24</sup> La compilación y explotación de corpus especializados monolingües ha resultado ser de enorme utilidad para la traducción inversa (Stewart, 2000; Corpas Pastor, 2001, 2002; Gavioli, 2005: 114 y ss.) y la traducción directa (Zanettin, 2001; Varantola, 2003; Corpas Pastor, 2004b, 2007); al igual que los corpus comparables bilingües resultan especialmente eficaces para la didáctica de la traducción en ambas direcciones (Friedbicher y Friedbicher, 1997; Gavioli y Zanettin, 2000; Corpas Pastor, 2004a). Son muchos los trabajos que recalcan los beneficios didácticos del corpus para el proceso de enseñanza-aprendizaje (cf. Bowker, 1998, 2000 y 2003; Rabadán y Fernández Nistal, 2002; Adab, 2004; Bernardini, 2000; o las monografías editadas por Bernardini y Zanettin, 2000; Zanettin, Bernardini y Stewart, 2003; y Beeby, Sánchez Gijón y Rodríguez Inés, 2009).

yecto concreto (conceptuales, terminológicas, estilísticas, discursivas, etc.), proporciona modelos en la producción del texto meta, ayuda a resolver problemas, a tomar decisiones y a validar las soluciones adoptadas en el proceso de la traducción. Evidentemente, este mismo tipo de información puede ser utilizada también para la evaluación objetiva de encargos de traducción.

Por otro lado, la compilación de corpus *ad hoc* es en sí misma una completa actividad de enseñanza-aprendizaje, pues supone seguir un protocolo que, en líneas generales, comprende el análisis del texto origen orientado a su traducción a la lengua meta, el establecimiento de descriptores de indización, la selección de los sistemas de recuperación de la información, la evaluación de fuentes y recursos electrónicos, así como la comprobación del grado de correspondencia del corpus compilado con respecto a las características lingüístico-discursivas y textuales del texto de origen, ya que los documentos seleccionados deben ser equiparables a éste en cuanto a la temática, el grado de especialización comunicativa y la variedad textual.

La enseñanza de la traducción especializada (jurídica, socio-económica y científico-técnica) se ha servido no sólo de corpus especializados monolingües comparables, sino también de los generales monolingües, como el BNC; y, sobre todo, de los especializados bilingües, comparables y paralelos, solos o en combinación con los monolingües a los que nos acabamos de referir. Buena parte de las investigaciones con corpus paralelos se basan en los principios de la lingüística contrastiva y de la enseñanza de lenguas para fines específicos. Los corpus paralelos tienen aplicaciones muy importantes para la didáctica de la traducción. Generalmente se usan para extraer términos, colocaciones y patrones sintagmáticos (léxico-gramaticales) equivalentes entre las lenguas implicadas.

La utilización de corpus paralelos alineados permite a los traductores en formación estudiar las estrategias empleadas por los traductores profesionales para resolver tal o cual problema, comprobar las pérdidas y alteraciones de la información que se producen en el contexto de la comunicación mediada, y, a partir de ahí, empezar a desarrollar sus propias estrategias futuras. Sin embargo, y a pesar de su potencial, es un hecho que los corpus paralelos no se emplean con fines pedagógicos tan frecuentemente como los comparables.

Finalmente, y aunque en menor medida, los corpus comparables de textos no traducidos se empiezan a utilizar progresivamente para la didáctica y la práctica profesional de la interpretación. Los principales temas abordados son la evaluación de la calidad, la direccionalidad y los procesos cognitivos que tienen lugar en la interpretación simultánea, la formación autónoma de intérpretes mediante el análisis de corpus, la precisión terminológica y otros aspectos de documentación preparatoria para interpretación de conferencias. Pero, en este caso, trabajos como los de Setton (2002), Chachibaia (2004), Lindquist y Giambruno (2006), o Bendazzoli y colaboradores<sup>25</sup> constituyen por el momento estudios aislados, que no llegan a conformar aún un cuerpo teórico.

Además de la vertiente didáctica de la que nos hemos ocupado hasta ahora, los Estudios de Traducción con corpus cuentan con una vertiente descriptiva. Esta gran corriente de investigación en Traductología ha estado liderada desde sus comienzos por los trabajos de Mona Baker. Esta línea, que se ha ido afianzado paulatinamente hasta ocupar el lugar central que

<sup>25</sup> Cf. Bendazzoli *et al.*, 2004; Bendazzoli y Sandrelli, 2005; Russo, Bendazzoli y Sandrelli, 2006.

tiene hoy día, tiene como eje el estudio de la lengua traducida, como una variedad más, en contraste, no con otras lenguas, sino en comparación con un corpus de textos producidos espontáneamente en la lengua de origen, de características similares al corpus de textos traducidos y, por tanto, equiparable a éste (Baker, 1993).

La metodología de corpus posibilitaba por primera vez el estudio detallado y sin complejos de la traducción como fenómeno y, en consonancia con ello, de las características intrínsecas de la lengua traducida, es decir, aquellos rasgos típicos de la producción de textos traducidos que los diferencian de los textos originales y que se ha dado en denominar *translationese* (por analogía al término *motherese* de la lingüística aplicada). Se parte de la idea de que tales rasgos diferenciales son susceptibles de ser investigados mediante las técnicas de corpus y que, de alguna manera, pueden revelar el comportamiento traductor, y mostrar ciertas tendencias que podrían ser consideradas universales de traducción.

Las propuestas de Baker (1993) aspiran a identificar los rasgos diferenciales de la lengua traducida y, de esta manera, poder caracterizar el comportamiento traductor, así como las restricciones a partir de las cuales opera. La autora enumera algunos de los aspectos de la traducción que se podrían investigar con la ayuda de corpus comparables traducidos. Así, en contraste con un corpus equiparable de textos originales, la lengua traducida se muestra más simple y menos ambigua; y, por consiguiente, más explícita; suele ser más convencional en el uso de formas gramaticales e, incluso, más prototípica. Como consecuencia de todo ello, los rasgos lingüísticos del discurso traducido presentan diferencias de distribución y frecuencia con respecto al discurso no traducido. Se trata de regularidades, rasgos comúnmente observables en todos los textos traducidos, independientemente de cuál sea la lengua del texto origen o de quién haya sido el traductor, que no varían de unas culturas a otras.

Estos rasgos universales están íntimamente relacionados con las leyes de traducción propuestas por Toury (1995), esto es, la ley de estandarización y la ley de interferencia.<sup>26</sup> Los Estudios de Traducción con Corpus trascienden, pues, las relaciones de equivalencia que se puedan establecer entre las palabras y estructuras del texto origen y el texto meta, para centrarse en los constructos teóricos globales —simplificación, explicitación, normalización, convergencia y, recientemente, transferencia—, que caracterizan la lengua traducida y sientan las bases de la rama descriptiva de la disciplina.

La simplificación hace referencia a la tendencia generalizada entre los traductores a producir un texto meta más sencillo y accesible, tanto con respecto a la forma como al fondo. Esta estrategia ya había sido observada anteriormente por autores como Blum-Kulka y Levenston (1983) en una variedad de contextos comunicativos: en las traducciones entre el inglés y el hebreo, en las producciones orales y escritas de los estudiantes de L2 (segunda lengua), en el discurso que utilizan profesores y hablantes nativos cuando se comunican con los hablantes no nativos, y en las lecturas simplificadas que se utilizan en la enseñanza-aprendizaje de L2. De hecho, la noción de *motherese* de la lingüística aplicada hace referencia

---

<sup>26</sup> La ley de estandarización postula que en traducción los textemas tienden a convertirse en repertoremas. Esto quiere decir que las intrincadas relaciones textuales que se establecen en el texto de origen se sustituyen en el texto meta por otras más habituales en la comunidad discursivo-textual meta. La ley de interferencia subraya la tendencia a realizar un trasvase (positivo o negativo) de los rasgos del texto de origen al texto meta.

a la estrategia de simplificación extrema que utilizan los progenitores, y los adultos en general, para comunicarse con los niños pequeños cuando éstos comienzan a adquirir su L1 (lengua materna).

Laviosa (1996, 1997, 1998a, b y c) ha realizado varios estudios empíricos con corpus comparables de textos literarios traducidos y no traducidos en inglés a fin de poder demostrar la naturaleza “universal” de la simplificación. Sin embargo, a pesar de ciertos resultados positivos en el corpus que analiza, en un estudio posterior (Laviosa, 2002), y tras haber realizado un completo estado de la cuestión, la autora se muestra pesimista acerca de la existencia del universal de simplificación. Corpas Pastor (2008), por su parte, refuta claramente dicha hipótesis de universalidad en un estudio en el que aplica técnicas de PLN sobre un corpus de textos científico-técnicos en español.<sup>27</sup>

Según Baker (1996), la simplificación del discurso traducido suele ir acompañada de cierto grado de explicitación (adición y especificación de información), y viceversa. En realidad, los traductores utilizan estrategias de explicitación desde tiempo inmemorial, al igual que ocurre con respecto a las producciones escritas en L2, por lo que la hipótesis de la explicitación podría ser aplicable al discurso mediado en general (Blum-Kulka, 1986). De nuevo, los resultados obtenidos no son concluyentes, pues mientras los trabajos de Baker y seguidores parecen mostrar la existencia de dicho universal (cf. Øverås, 1998; Olohan y Baker, 2000, Olohan, 2004, etc.), otros autores demuestran todo lo contrario, esto es, la existencia de una fuerte tendencia a la implícitación en el texto traducido (cf. Puurtinen, 2003, para el finlandés, y Cheong, 2006, para el coreano).

Baker (1996) considera que la tendencia a la normalización y al conservadurismo constituye un tercer universal. De acuerdo con ello, los textos traducidos suelen ser más prototípicos que los no traducidos: por un lado, reproducen, y hasta exageran, aquellos rasgos que son propios y característicos de la lengua meta; y, por el otro, resultan más conservadores y presentan un menor grado de creatividad que los textos no traducidos. Y ello no se limita a la puntuación o a la gramática, sino que se extiende también por los patrones léxico-combinatorios. Este universal ha sido estudiado especialmente por Kenny (1998, 2001), aunque sus conclusiones son ambivalentes, pues su estudio confirma y, al mismo tiempo, desmiente la supuesta tendencia a la normalidad de los textos traducidos. Quizá ello apunte hacia la naturaleza más bien “normativa” de los supuestos universales de traducción, que debería comprobarse en estudios a larga escala, que incluyan corpus representativos, tanto de textos literarios como no literarios, en una variedad de pares de lenguas, y teniendo en cuenta los usos imperantes, así como las características socio-culturales y temporales de las comunidades lingüístico-discursivas implicadas. Lo cierto es que el panorama es algo dispar en este sentido. Olohan (2004), en cambio, confirma la hipótesis normalizadora sobre un estudio de frecuencias del corpus comparable literario inglés utilizado.

---

<sup>27</sup> Sin embargo, otros estudios sostienen la existencia de tal universal. Éste es el caso Russo, Bendazzoli y Sandrelli (2006), que utilizan un corpus comparable de discursos originales en español y discursos interpretados al español a partir de discursos originales en italiano y en inglés. Los autores concluyen que las producciones de los intérpretes, resultan más simples y menos ambiguas que el discurso producido originalmente.

La convergencia guarda una estrecha relación con la normalización, en tanto ambos rasgos reflejan la tendencia de los textos traducidos a presentar un alto grado de convencionalismo y conservadurismo. Apenas ha habido investigación empírica sobre este tema, por lo que no se dispone de estudios que demuestren fehacientemente la existencia de convergencia en traducción. No obstante, Laviosa (1996, 2002) señala que los textos traducidos suelen presentar menor variación interna y presentan una mayor similitud entre sí, en tanto *convergen*, es decir, tienden hacia un mismo punto situado en el centro del eje. Laviosa (2002: 72-73) caracteriza la convergencia como la tendencia de los textos traducidos a agruparse (*clustering*) en torno a los valores medios de sus propias características internas. A conclusiones similares llegan Olohan (2003) y Mauranen (2004). Corpas Pastor (2008), en cambio, señala que sólo se da convergencia a nivel estilístico, pero no a nivel sintáctico, en su análisis de un corpus de traducciones de textos científico-técnicos en español, en el cual se han utilizado técnicas de PLN y cálculos estadísticos.

La transferencia se une a la lista anterior de universales —simplificación, explicitación, normalización y convergencia—, junto a los cuales constituye un quinto universal “en discordia”. Y curiosamente, al igual que la mayoría de ellos, guarda también una estrecha relación con las características de la producción de segundas lenguas (cf. Levenston, 1979). A mediados de esta década se ha comenzado a estudiar la influencia que pudieran tener los originales en las traducciones, y si dicha influencia pudiera ser de carácter “universal”. Los resultados obtenidos matizan los universales propuestos hasta ahora, relativizándolos y proponiendo nuevos rasgos universales que subsumiremos bajo la denominación única de *transferencia*.

La ley de interferencia a la que nos referíamos más arriba postula el trasvase de rasgos del texto de origen al texto meta como una constante en traducción. Esta ley permite predecir la existencia de transferencia (positiva o negativa) en los textos traducidos. Se produce transferencia negativa cuando el texto traducido se desvía de las prácticas normales, habituales y sancionadas por la comunidad meta. Por otro lado, la transferencia positiva aumenta las posibilidades de que se seleccionen rasgos en el texto meta presentes en el texto de origen. Esto es, el universal de transferencia postula que los textos traducidos abundan en características que tienen correspondencias directas en la lengua meta; y que, por el contrario, aquellas características que son idiosincrásicas de la lengua meta, y, por tanto, únicas, específicas y carentes de correspondencias en la lengua de origen, tienden estar infrarrepresentadas en los textos meta. Diversas investigaciones parecen corroborar la existencia de este quinto universal, como, por ejemplo, los trabajos de Tirkonnen-Condit (2002, 2004), Mauranen (2004), Eskola (2004) sobre un corpus de traducciones literarias al finlandés; o los de Teich (2003) y Corpas Pastor (2008), quienes aplican también técnicas de PLN a un corpus de divulgación (inglés-alemán) y científico-técnicos (inglés-español), respectivamente.

Además del estudio de universales, el análisis de corpus también puede arrojar resultados muy interesantes en cuanto al estudio de las normas de traducción y la caracterización de diversos estilos traductores. Según Baker (1992), el análisis con corpus estaría especialmente indicado para la investigación sobre las normas operativas de Toury, y, más concretamente, en el caso de las matriciales—que gobiernan la distribución del material lingüístico en el texto meta (dan cuenta, pues, de las omisiones, adiciones, sustituciones y transposicio-

nes)—, y de las textuales, que determinan la realización formal del texto meta. Dichas normas, que varían de unas culturas a otras, constituyen regularidades de otro tipo que también se pueden observar con respecto a determinados tipos de traducción en contextos históricos y socio-culturales concretos. Por ejemplo, en japonés se observa una mayor tolerancia hacia el uso de préstamos en traducción, que, por ejemplo, en árabe o en francés (Baker, 1992). La autora cita a Toury y Vanderauwera como ejemplos de autores que parecen haber utilizado tanto corpus comparables monolingües como corpus. Y, descendiendo en la escala de abstracción, Baker (2001) también se refiere a la utilidad del corpus para la caracterización estilística de las producciones de los traductores. En esa misma línea van, por ejemplo, los trabajos de Bosseaux (2001), Olohan (2003, 2004) y Saldanha (2005, 2011/En prensa) sobre la variedad de estilos de los traductores literarios.

#### 4. Conclusión

En las páginas precedentes hemos ofrecido un panorama general sobre la relevancia de las tecnologías lingüísticas en Traducción, con especial referencia al corpus, como recurso y como metodología de análisis e investigación. Las tecnologías lingüísticas, y más concretamente, las tecnologías de la traducción, vienen conformando una de las líneas prioritarias de las políticas europeas desde principios de la década de los ochenta. Las investigaciones sobre corpus iniciadas en el seno de los diferentes Programas Marco lanzados por la Unión Europea se han saldado con un desarrollo espectacular de los principios y técnicas de la Lingüística de Corpus, lo cual, a su vez, ha dinamizado el crecimiento de la Lingüística Computacional teórica y su vertiente más aplicada, el Procesamiento de Lenguaje Natural. Al mismo tiempo, ambas disciplinas han contribuido a que la metodología de corpus alcance su madurez.

Las denominadas “tecnologías de la traducción” son herramientas diseñadas específicamente para ayudar al traductor en su quehacer diario. Entre ellas se cuenta la compilación y gestión de corpus. La capacidad para manejarse con este tipo de tecnologías lingüísticas forma parte de las competencias profesionales del traductor. Si bien durante un tiempo la formación en competencia tecnológica informacional se ha relegado a un segundo plano en los planes de estudios de traducción, lo cierto es que cada vez se presta más atención a este tema.

En este trabajo nos hemos centrado fundamentalmente en la formación de traductores e intérprete mediante corpus. La pedagogía de la traducción ha apostado por enfoques con corpus, como, por ejemplo, el DDL, ya validados en el contexto de la enseñanza de idiomas y el análisis de errores de la Lingüística Aplicada. También la vertiente descriptiva de los Estudios de Traducción ha recogido ideas provenientes del aprendizaje y adquisición de segundas lenguas y de la caracterización de la comunicación mediada, que ha sabido reinterpretar mediante el análisis de corpus. De ahí surgen, por ejemplo, las investigaciones sobre normas, leyes y universales que supuestamente caracterizan la lengua traducida. En este ámbito, la metodología con corpus está empezando a aliarse con los principios y técnicas propias del Procesamiento de Lenguaje Natural. Este maridaje de metodologías está dando lugar a investigaciones de mayor calado y empirismo sobre la naturaleza del texto traducido y el comportamiento traductor. A buen seguro que esta alianza entre los Estudios de Traducción con cor-

pus y el Procesamiento del Lenguaje Natural se salda con resultados sorprendentes, muy diferentes de lo que estábamos acostumbrados a encontrar en los análisis de corpus intuitivos, algo subjetivos y rara vez extrapolables.

## 5. Bibliografía

- Adab, B. 2004. "Evaluating Choices: The Role of Corpora in Translation Choices and Target Text Assessment". En I. Kemble (ed.). 2004. *Using Corpora and Databases in Translation: Proceedings of the Conference held on 14th November 2003 in Portsmouth*. Portsmouth: University of Portsmouth. 1-14.
- Aston, G. 1995. "Corpora in Language Pedagogy: Matching Theory and Practice". En G. Cook y B. Seidlhofer (eds.). 1995. *Principles and Practice in Applied Linguistics*. Oxford: Oxford University Press. 257-270.
- Aston, G. (ed.). 2001. *Learning with Corpora*. Berna, Berlín, Bruselas, Fráncfort, Nueva York, Oxford, Viena: CLUEB.
- Austermühl, F. 2001. *Electronic Tools for Translators*. Mánchester: St. Jerome.
- Baker, M. 1992. *In Other Words: A Course Book on Translation*. Londres: Routledge.
- Baker, M. 1993. "Corpus Linguistics and Translation Studies –Implications and Applications". En M. Baker, M.G. Francis y E. Tognini-Bonelli (eds.). 1993. *Text and Technology: In Honour of John Sinclair*. Ámsterdam y Filadelfia: John Benjamins. 233-250.
- Baker, M. 1996. "Corpus-based Translation Studies: The Challenges that Lie Ahead". En H. Somers (ed.). 1996. *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*. Ámsterdam y Filadelfia: John Benjamins. 175-186.
- Baker, M. 2001. "Towards a Methodology for Investigating the Style of a Literary Translator". *Target*. 12 (2). 241-266.
- Bateman, J. y Zock, M. 2003. "Natural Language Generation". En R. Mitkov (ed.). 2003b. 284-304.
- Beeby, A.; Sánchez Gijón, P. y Rodríguez Inés, P. (eds.). 2009. *Corpus Use and Learning to Translate. Proceedings from the CULT Conference (22-24 January, Barcelona, Spain)*. Ámsterdam y Filadelfia: John Benjamins.
- Bendazzoli, C., Monti, C.; Sandrelli, A.; Russo, M.; Baroni, M.; Bernardini, S.; Mack, G.; Ballardini, E. y Mead, P. 2004. "Towards the Creation of an Electronic Corpus to Study Directionality in Simultaneous Interpreting". En N. Oostdijk, G. Kristoffersen, y G. Sampson (eds). *Compiling and Processing Spoken Language Corpora, LREC 2004 Satellite Workshop, Proceedings of the 4<sup>th</sup> International Conference on Language Resources and Evaluation*. 33-39.
- Bendazzoli, C. y Sandrelli, A. 2005. "An Approach to Corpus-Based Interpreting Studies: Developing EPIC (European Parliament Interpreting Corpus)". *MuTra 2005 – Challenges of Multidimensional Translation: Conference Proceedings*. 1-12.

- Bernardini, S. 2000. *Competence, Capacity, Corpora*. Bologna: Cooperativa Libreria Universitaria Editrice.
- Bernardini, S. y Zanettin, F. (eds.). 2000. *I corpora nella didattica della traduzione. Corpus Use and Learning to Translate*. Bologna: CLUEB.
- Bouillon, P. y Starlander, M. 2009. "Technology in Translator Training and Tools for Translators". En *MT XII: Proceedings of the Twelfth Machine Translation, August 26-30, 2009*. [http://www.mt-archive.info/MTS-2009-Bouillon-ppt.pdf].
- Bowker, L. 1998. "Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study". *META*. 43 (4). 631-651.
- Bowker, L. 2000. "A Corpus-Based Approach to Evaluating Student Translations". *The Translator*. 6 (2). 183-210.
- Bowker, L. 2002. *Computer-Aided Translation Technology: A Practical Introduction*. Ottawa: University of Ottawa Press
- Bowker, L. 2003. "Corpus-Based Applications for Translator Training: Exploring the Possibilities". En S. Granger, J. Lerot y S. Petch-Tyson (eds.). 2003. 169-183.
- Bowker, L. y Bennison, P. 2003. "Student Translation Archive and Student Translation Tracking System: Design, Development and Application". En F. Zanettin, S. Bernardini, y D. Stewart (eds). 2003. 103-117
- Bowker, L. y Corpas Pastor, G. 2012/En preparación. "Translation Technology". En R. Mitkov (ed.) 2012/En preparación. *The Oxford Handbook of Computational Linguistics*. 2ª ed. Oxford y Nueva York: Oxford University Press.
- Blum-Kulka S. 1986. "Shifts in Cohesion and Coherence in Translation". En J. House y S. Blum-Kulka (eds.). 1986. *Interlingual and Intercultural Communication*. Tubinga: Narr. 17-37.
- Blum-Kulka, S. y Levenston, E. A. 1983. "Universals of Lexical Simplification". En C. Faerch y G. Kasper (eds.). 1983. *Strategies in Interlanguage Communication*. Londres: Longman. 119-139.
- Bosseaux, C. 2001. "A Study of the Translator's Voice and Style in the French Translations of Virginia Woolf's *The Waves*". En M. Olohan (ed.). 2001. *CTIS Occasional Papers*. Vol. 1. Mánchester: Centre for Translation & Intercultural Studies, UMIST. 55-75.
- Chachibaia, N. 2004. "A Communicative Approach to Simultaneous Interpreting: a Corpus-based study". En I. Kemble (ed.). 2004. *Using Corpora and Databases in Translation*. Londres: University of Portsmouth. 140-155.
- Cheong, H.-J. 2006. "Target Text Contraction in English-into-Korean Translations: a Contradiction of Presumed Translation Universals?". *Meta*. 51(2). 343-367.
- Comisión de las Comunidades Europeas. 2007. *ICT - INFORMATION AND COMMUNICATION TECHNOLOGIES. A Theme for research and development under the specific programme "Cooperation" implementing the Seventh Framework Programme (2007-2013) of the European Community for research, technological development and demonstration activities Work Programme 2007-08*. Bruselas: CECA-CE-CEEA

- Connor, U. y Upton, T. A. (eds.). 2004. *Discourse in the Professions: Perspectives from Corpus Linguistics*. Studies in Corpus Linguistics. Ámsterdam y Filadelfia: John Benjamins.
- Corpas Pastor, G. 2001. "Compilación de un corpus *ad hoc* para la enseñanza de la traducción inversa especializada". *TRANS. Revista de traductología*. 5. 155-184.
- Corpas Pastor, G. 2002. "Traducir con corpus: de la teoría a la práctica". En J. García Palacios y M. T. Fuentes (eds.). 2002. *Texto, terminología y traducción*. Salamanca: Almar. 189-226.
- Corpas Pastor, G. 2004a. "La traducción de textos médicos especializados a través de recursos electrónicos y corpus virtuales". En L. González y P. Hernández. (eds.). 2004. *Las palabras del traductor. Actas del II Congreso Internacional "El español, lengua de traducción", 20 y 21 de mayo, Toledo 2004*. Bruselas: Comisión Europea/ESLETRA. 137-164.
- Corpas Pastor, G. 2004b. "Localización de recursos y compilación de corpus vía Internet: aplicaciones para la didáctica de la traducción médica especializada". En V. García Yebra y C. Gonzalo García (eds.). 2004. *Manual de documentación y terminología para la traducción especializada*. Colección "Instrumenta Bibliologica". Madrid: Arco/Libros. 223-506.
- Corpas Pastor, G. 2007. "Lost in specialised translation: the corpus as an inexpensive and under-exploited aid for language service providers". En *Translating and the Computer 29. Proceedings of the Twenty-ninth International Conference on Translating and the Computer, 29-30 November 2007, Londres*. Aslib/IMI: Londres. 1-21.
- Corpas Pastor, G. 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma*. (Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation, 49). Fráncfort: Peter Lang.
- EAGLES (Expert Advisory Group on Language Engineering Standards). 1996. "Text corpora Working Group reading Guide". EAGLES Document EAG-TCWG-FR-2. Versión de mayo de 1996. [<http://www.ilc.cnr.it/EAGLES/corpintr/corpintr.html>].
- Eskola, S. 2004. "Unusual Frequencies in Translated Language". En A. Mauranen y P. Kujamäki (eds.). 2004. 83-99.
- Friedbicher, I. y Friedbicher, M. 1997. "The Potential of Domain-Specific Target-Language Corpora for the Translator's Workbench". *Corpus use and learning to translate. Bertinoro, 14-15 November 1997*. [<http://www.sslmit.unibo.it/cultpaps/paps.htm>].
- Ghadessy, M.; Henry, A. y Roseberry, R. L. (eds.). 2001. *Small Corpus Studies and ELT: Theory and Practice*. Ámsterdam y Filadelfia: John Benjamins.
- Gavioli, L. 2005. *Exploring Corpora for ESP Learning*. Ámsterdam y Filadelfia: John Benjamins
- Gavioli, L. y Zanettin, F. 2000. "I corpora bilingui nell'apprendimento della traduzione. Riflessioni su un'esperienza pedagogica". En S. Bernardini y F. Zanettin (eds.). 2000. 61-80.

- Hansen, V. e Yvon, F. (eds.). 2010. *EAMT 2010: Proceedings of the 14th Annual conference of the European Association for Machine Translation, 27-28 May 2010*. [http://www.mt-archive.info/EAMT-2010-TOC.htm]
- Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Kenny, D. 1998. "Creatures of Habit? What Translators usually do with Words". En S. Laviosa (ed.). 1998d. 571-588.
- Kenny, D. 2001. *Lexis and Creativity in Translation*. Mánchester: St. Jerome.
- Kruger, A.; Wallmach, K. y Munday, J. (eds.). 2011/En prensa. *Corpus-based Translation Studies: Research and Applications*, Continuum, London/New York, 2011.
- Lavid, J. 2005. *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Madrid: Cátedra.
- Laviosa, S. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. Tesis doctoral. UMIST: Universidad de Mánchester.
- Laviosa, S. 1998a. "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose". En S. Laviosa (ed.). 1998d. 557-570.
- Laviosa, S. 1998b. "The Corpus-based Approach: a New Paradigm in Translation Studies". En S. Laviosa (ed.). 1998d. 474-479.
- Laviosa, S. 1998c. "The English Comparable Corpus: a Resource and a Methodology". En L. Bowker, M. Cronin, D. Kenny y J. Pearson (eds.). 1998. *Unity in Diversity: Current Trends in Translation Studies*. Mánchester: St. Jerome 101-112.
- Laviosa, S. (ed.). 1998d. *L'Approche Basée sur le corpus/The Corpus-Based Approach*. Número especial de la revista *Meta* 43 (4). Montreal: Les Presses de L'Université de Montreal.
- Laviosa, S. 2002. *Corpus-based Translation Studies. Theory, Findings, Applications*. Ámsterdam y Nueva York: Rodopi.
- Levenston, E. 1971. "Over-Indulgence and Under-Representation: Aspects of Mother Tongue Interference". En G. Nickel (ed.) 1971. *Papers in Contrastive Linguistics*. Cambridge: Cambridge University Press. 115-121.
- Lindquist, P. y Giamburro, C. 2006. "The MRC Approach: Corpus-Based Techniques Applied to Interpreter Performance Analysis and Instruction". *FORUM - International Journal of Interpretation and Translation*. 4 (1). 103-138.
- Lüdeling, A. y Kytö, M. 2008-2009. *Corpus Linguistics: An International Handbook*. 2 vols. Berlín: Mouton de Gruyter.
- Mauranen, A. 2004. "Corpora, Universals and Interference". En A. Mauranen y P. Kujamäki (eds.). 2004. 65-82.
- Mauranen, A. y Kujamäki, P. (eds). 2004. *Translation Universals: Do They Exist?* Ámsterdam: John Benjamins.
- McEnery, T.; Xiao, R. y Tono, Y. (eds). 2006. *Corpus-Based Language Studies. An Advanced Resource Book*. Londres y New York: Routledge.
- McTear, M. F. M. 2004. *Spoken Dialogue Technology: Towards the Conversational User Interface*. Londres: Springer.

- Mitkov, R. (ed.). 2003. *The Oxford Handbook of Computational Linguistics*. Oxford y Nueva York: Oxford University Press.
- Olohan, M. 2003. "How Frequent are the Contractions? A Study of Contracted form in the Translational English Corpus". *Target*. 15. 59-89.
- Olohan, M. 2004. *Introducing Corpora in Translation Studies*. Abingdon, Oxfordshire y Nueva York: Routledge.
- Olohan, M. y Baker, M. 2000. "Reporting *That* in Translated English: Evidence for Subconscious Processes of Explicitation". *Across Languages and Cultures*. 1. 141-172.
- Ooi, V. B. Y. 1998. *Computer Corpus Lexicography*. Edimburgo: Edinburgh University Press.
- Oostdijk, N. 1991. *Corpus Linguistics and the Automatic Analysis of English*. Ámsterdam: Rodopi.
- Øverås, L. 1998. "In Search of the Third Code: an Investigation of Norms in Literary Translation". En S. Laviosa (ed.). 1998d. 571-588.
- Puurtilinen, T. 2004. "Explicitation of Clausal Relations. A Corpus-Based Analysis of Clause Connectives in Translated and Non-Translated Finnish Children's Literature". En A. Mauranen y P. Kujamäki (eds.). 2004. 165-176.
- Quah, C. K. 2005. *Translation and Technology*. Palgrave: MacMillan.
- Rabadán, R. y Fernández Nistal, P. 2002. *La traducción inglés-español: fundamentos, herramientas, aplicaciones*. León: Servicio de publicaciones de la Universidad de León.
- Russo, M.; Bendazzoli, C.; Sandrelli, A. 2006. "Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC (European Parliament Interpreting Corpus)". *FORUM - International Journal of Interpretation and Translation*. 4(1). 221-254.
- Sager, J. 1993. *Language Engineering and Translation: Consequences of Automation*. Ámsterdam y Filadelfia: John Benjamins.
- Saldanha, G. 2005. *Style of Translation: an Exploration of Stylistic Patterns in the Translations of Margaret Jull Costa and Peter Bush*. Tesis doctoral. Dublín: Universidad de Dublín.
- Saldanha, G. 2011/En prensa. "Style of Translation: the Use of Source Language Words in Translations by Margaret Jull Costa and Peter Bush". En Kruger, Wallmachy Munday (eds.). 2011/En prensa.
- Sargent, B. 2007. "Translation Management Systems and Subcategories". *Multilingual*. 86 (18-2). 83-86.
- Setton, R. 2002. "A Methodology for the Analysis of Interpretation Corpora", en G. Garzone y M. Viezzi (eds.) 2002. *Interpreting in the 21st Century: Challenges and opportunities: Selected papers from the 1st Forlì Conference on Interpreting Studies, 9-11 November 2000*. Ámsterdam y Filadelfia: John Benjamins. 29-45.
- Sinclair, J. M. y Renouf, A. 1988. "A Lexical Syllabus for Language Learning". En R. Carter y M. McCarthy (eds.). 1988. *Vocabulary and Language Teaching*. Harlow: Longman. 140-158.
- Somers, H. 2003b. "The Translator's Workstation". En H. Somers (ed.). 13-30.

- Somers, H. (ed.) 2003a. *Computers and Translation: A Translator's Guide*. Ámsterdam y Filadelfia: John Benjamins.
- Stewart, D. 2000. "Conventionality, Creativity and Translated Text: the Implications of Electronic Corpora in Translation". En M. Olohan (ed.). 2000. *Intercultural Faultiness. Research Models in Translation Studies I. Textual and Cognitive Aspects*. St. Jerome: Mánchester. 73-91.
- Teich, E. 2003. *Cross-Linguistic Variation in System and Text*. Berlín y Nueva York: Mouton de Gruyter.
- Tirkkonen-Condit, S. 2002. "Translationese — a Myth or an Empirical Fact?". *Target*. 14 (2). 207-220.
- Tirkkonen-Condit, S. 2004. "Unique Items – Over- or Under-Represented in Translated Language?". En A. Mauranen y P. Kujamäki (eds). 2004. 177-184.
- Torres del Rey, J. 2003. *Nuevas tecnologías y enseñanza de la traducción: límites y posibilidades de los modelos de aplicación tecnológica para la formación de traductores*. Tesis doctoral. Universidad de Salamanca.
- Torres del Rey, J. 2005. *La interfaz de la traducción. Formación de traductores y nuevas tecnologías*. Granada: Comares.
- Toury, G. 1995. *Descriptive Translation Studies and Beyond*. Ámsterdam: John Benjamins.
- UNE-EN 15038:2006 *Servicios de traducción. Requisitos para la prestación del servicio*.
- Varantola, K. 2003. "Translators and Disposable Corpora". En F. Zanettin, S. Bernardini y D. Stewart (eds.). 2003. 55-70.
- Willis, D. 1990. *The Lexical Syllabus: a New Approach to Language Teaching*. Londres: Collins.
- Willis, J. y Willis, D. 1989. *Collins COBUILD English Course*. Londres: Collins COBUILD.
- Zanettin, F. 2001. "Swimming in Words". En G. Aston (ed.). 2001c. 177-197.
- Zanettin, F.; Bernardini, S. y Stewart, D. (eds). 2003. *Corpora in Translator Education*. Mánchester: St. Jerome.