5. La lingüística de corpus al servicio de la traducción y la terminología

5.1. Introducción

La metodología denominada *lingüística de corpus* ha tenido y sigue teniendo una importante repercusión en el estudio del lenguaje[50] y en el desarrollo de programas informáticos para el procesamiento del lenguaje natural[51] (Butler 1992; McEnery y Wilson 1996: 87-145). Podemos encontrar una explicación de esto en el hecho de que, desde sus comienzos, esta metodología ha tenido la pretensión empiricista de estudiar el funcionamiento real de la lengua mediante el estudio de ingentes cantidades de texto producidos por los hablantes de una lengua. Asimismo, la importancia económica que reportan los córpora informatizados a los grandes editores de diccionarios y a las industrias de la lengua[52] pueden explicar este vigor. Otros factores que también han propiciado el desarrollo de córpora informatizados han sido la creciente capacidad de los nuevos ordenadores y una mayor accesibilidad a textos en soporte electrónico. Esta se ha logrado gracias a la comunicación por Internet y a instituciones que se dedican a recopilar córpora reutilizables como ACL (Association for Computational Linguistics), ECI (European Corpus Initiative), LDC (Linguistic Data Consortium), ICAME (International Computer Archive of Modern and Medieval English), ACL/DCI (Association for Computational Linguistics Data Collection Initiative) o ELRA (European Language Resources Association).

De todas formas, antes de proseguir, es preciso aclarar qué se entiende por *corpus*, ya que este término se suele confundir con otros relacionados como *archivo*, *colección* o *textoteca*. Por otra parte, en los estudios de traducción, el término *corpus* designa un conjunto muy limitado de textos en soporte papel del que difícilmente se

[50] Algunas de las disciplinas que se pueden beneficiar de esta metodología son la semántica, lexicología y lexicografía, fonética, gramática, pragmática y análisis del discurso, sociolingüística, estilística, lingüística diacrónica, didáctica de las lenguas, psicolingüística, estudios culturales y traducción.

[51] En este sentido, se han desarrollado infinidad de aplicaciones para el procesamiento del lenguaje y del habla que han sido aprovechadas en lexicografía monolingüe y bilingüe, en el análisis y etiquetaje sintáctico automático de textos, en la traducción automática y en sistemas de gran utilidad para el traductor como los bancos terminológicos y las memorias de traducción.

[52] El País, jueves 2 de abril de 1998.

pueden extraer conclusiones generalizables. A continuación presentamos una definición que puede aclarar el sentido del término en el ámbito de la lingüística de corpus (McEnery y Wilson 1996: 24).

> So a corpus in modern linguistics, in contrast to being simply any body of text, might more accurately be described as a finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. (McEnery y Wilson 1996: 24)

A la idea de tamaño limitado, formato electrónico, muestreo y representatividad, McEnery y Wilson añaden que todo corpus debe constituir una referencia estándar a la lengua que pretende representar. Para que el corpus constituya una referencia reutilizable (Hockey y Walker 1993: 235), ciertas organizaciones como TEI (Text Encoding Initiative), en el pasado, y EAGLES (Expert Advisory Group on Language Engineering Standards), en la actualidad, se esfuerzan por elaborar normas de estandarización. Otros autores como Leech y Fligelstone (1992: 116) también inciden en que los córpora suelen ser representativos de una determinada lengua o tipo textual y añaden que son recopilados con un claro objetivo en mente.

Teniendo en cuenta que una revisión de los avances alcanzados por la lingüística de corpus constituiría de por sí una tesis, en este apartado se destacarán los aspectos de la lingüística de corpus relacionados con la traducción y las aplicaciones de esta metodología a la traducción. En Biber *et al.* (1998), Kennedy (1998), Pérez Guerra (1998) y en los archivos de la lista de distribución *Corpora* [http://www.hd.uib.no] se pueden consultar algunas de las referencias bibliográficas, publicaciones periódicas, listas de distribución, organizaciones y direcciones de Internet más relevantes.

5.2. Evolución de la lingüística de corpus

La utilización de córpora en el estudio del lenguaje entronca, por una parte, con la tradición anglosajona iniciada por Firth (1935, 1957a, 1957b) y continuada principalmente por sus discípulos, Halliday y Sinclair, y por otra parte, con la labor de estructuralistas norteamericanos como C.C. Fries, H. Ku¥era y W.N. Francis (*apud* Stubbs 1996: 22-50). La primera influencia preconiza que el significado de una palabra

depende de su coocurrencia con otras palabras[53] y vislumbra lo que será una concepción contextual del significado léxico (Cruse 1986; Pustejovski 1995; Wanner 1996). La segunda se enmarca en el auge que los métodos empíricos y estadísticos experimentaron en la década de los cincuenta.

Estos planteamientos cristalizaron en la creación del corpus *Brown* de inglés americano (Francis y Ku¥era 1982), al que siguieron otros como el corpus *London-Lund* o el corpus *LOB (Lund-Oslo-Bergen).* En la década de los setenta, la investigación en traducción automática y lingüística computacional fue impulsada en Canadá y Europa Occidental, en las que se aspiraba a una política multicultural. La demanda de traducciones de documentación legal, administrativa, técnica y científica fue la que hizo que se construyeran los primeros córpora bilingües, que contenían textos en una lengua junto con su traducción a otra. Un buen ejemplo lo constituye el *Canadian Hansard*, en el que se transcriben en inglés y en francés las reuniones del parlamento canadiense de mediados de los setenta (Brown *et al.* 1993). El millón de palabras que lo conforman fue utilizado en traducción automática.

Mientras se elaboraban los primeros córpora bilingües y multilingües en soporte electrónico, se produjo un importante esfuerzo de compilación de diccionarios en inglés, que ya se había convertido en *lingua franca.* Este interés sería el motor del primer proyecto lexicográfico basado en corpus: la elaboración del diccionario *Collins COBUILD* (1987), auspiciado por la editorial HarperCollins y la Universidad de Birmingham. Este diccionario se fundamentaba en un corpus de 20 millones de palabras llamado *Birmingham corpus,* que se convirtió en un instrumento capaz de aportar para cada palabra no sólo información semántica sino también información sobre su combinatoria, su sintaxis más típica y sus patrones estructurales y colocacionales. Al mismo tiempo, este corpus, que ha seguido creciendo hasta alcanzar 329 millones de palabras en julio de 1998 y que ahora se denomina *Bank of English*[54], sentó la base para

[53] Esto queda magistralmente expresado en la famosa cita de Firth (1957): "You shall know a word by the company it keeps". Esta cita también vislumbra lo que se denominará *ecología del lenguaje* (Walker 1991), es decir, la tendencia a relacionar los textos con un contexto que sirva de marco de referencia.

[54] El servicio *CollinsDirect* [http://titania.cobuild.collins.co.uk/direct_info.html] hace accesibles concordancias y colocaciones extraídas a partir de 50 millones de palabras de este corpus, que contiene textos en inglés británico principalmente, aunque también norteamericano (9 millones), tanto orales como escritos.

ingentes córpora monolingües posteriores como el *British National Corpus (BNC)*, el *International Corpus of English (ICE)* o el *Survey of English Usage (SEU).*

También en la década de los ochenta, proliferaron córpora especializados destinados a sistemas de traducción automática de sublenguajes (McKeown 1992) y a la didáctica del inglés. En esta última categoría, entraría un corpus científico-técnico desarrollado en China por Yang (1985) o el *Guangzhou Petroleum English Corpus* destinado a la industria petrolífera (Zhu 1985).

La década de los noventa va a suponer un avance significativo en cuestiones como el diseño del corpus, que intenta reflejar las necesidades del usuario, la recopilación, la anotación y tratamiento informático de los textos, la accesibilidad, la estandarización y la evaluación[55] de resultados.

Al igual que ocurriera en los cincuenta, se ha producido un renacer del interés por métodos empíricos y estadísticos, en cierto modo motivado por la existencia de ordenadores mucho más potentes. Junto con esto, las relaciones políticas, comerciales y científicas de ámbito internacional, que demandan más que nunca la edición de documentos en más de una lengua, y la necesidad de aprender idiomas suponen un empuje para los diccionarios monolingües y, por ende, para los córpora bilingües. Y así, la década de los noventa asiste a un aumento del interés por córpora bilingües y multilingües por parte de investigadores en traducción automática (Brown *et al.* 1993)[56], lexicografía bilingüe (Klavans y Tzoukermann 1990), didáctica de lenguas extranjeras (Barlow 1994, 1996) y en terminología (Budin y Wright 1997; Cabré, 1997). Estos córpora, llamados también *córpora paralelos* (véase 5.3.), pueden proporcionar información documentada sobre cómo textos escritos en una lengua se pueden trasladar a otra de acuerdo con una serie de factores contextuales como el estilo, registro,

[55] Este tema fue el hilo conductor del congreso internacional que *ELRA (European Language Resources Association)* celebró en Granada en 1998 con el título: *First International Conference on Language Resources and Evaluation.*

[56] A finales de los ochenta y principios de los noventa, la traducción automática se ha relanzado de nuevo, una vez asumidas sus limitaciones y la necesidad de restringir sus objetivos (Teubert 1996: 242). El resultado ha sido la creación de sistemas tanto de traducción automática como de traducción humana asistida por ordenador de gran eficacia. Entre los principales sistemas de traducción automática, podemos citar Systran, Logos, Eurotra y Globalink Translation Systems (GTS). En cuanto a los programas de traducción asistida, los más conocidos son IBM Translation Manager 2, Star Transit, Trados Translator's Workbench y Déjà Vu.

dominio, etc. La elaboración de estos córpora se ve beneficiada también por la experiencia adquirida en la creación de córpora monolingües y por la posibilidad de edición en CD-Rom.

Barlow (1996: 49) menciona una serie de proyectos relacionados con la elaboración de córpora bilingües y multilingües. Entre los primeros, se encuentran *Intersect* (francés-inglés), *Contragram*, el *Canadian Hansard Corpus* y un corpus paralelo inglés-noruego recopilado en la Universidad de Oslo. En cuanto a los córpora multilingües, Barlow hace mención al corpus *Aarhus* de derecho contractual (danés-inglés-francés), al corpus *Triptic* (inglés-francés-neerlandés) y a una serie de proyectos europeos como Lingua, Multext y Multex-East[57].

Hasta aquí, la mayoría de los córpora mencionados han sido concebidos para representar la lengua general[58]. Sin embargo, la utilización del inglés como *lingua franca* en los dominios de especialidad hace que estos córpora no suplan las necesidades de sus usuarios. Surge entonces una acuciante necesidad de crear córpora con fines específicos o córpora de sublenguajes que, sin ser tan extensos como los de la lengua general, puedan facilitar la comunicación en inglés entre expertos y la traducción especializada de o hacia el inglés. En este sentido, Gledhill (1996: 109) afirma que la mayor parte de la investigación en lingüística de corpus ha estado orientada hacia la elaboración de córpora generales representativos y sólo recientemente se ha prestado atención a variedades especializadas del inglés.

En consecuencia, en la actualidad se están desarrollando numerosos córpora monolingües y multilingües especializados y también proyectos dentro del sector de la ingeniería lingüística[59]. En estos córpora y proyectos están implicados investigadores de

---

[57] Para más información sobre estos corpus se puede visitar la página sobre textos paralelos elaborada por este investigador [http://www.ruf.rice.edu/~barlow/para.html].

[58] A este tipo de córpora el Observatorio Español de Industrias de la Lengua [http://cervaxp.cervantes.es/oeil/oeilin2.html] lo denomina *corpus de la lengua general con fines generales*.

[59] Dentro del Programa de Aplicaciones Telemáticas de la Comisión Europea hay abundantes proyectos en marcha como MABLe (*Multilingual Authoring of Business Letters*), AVENTINUS (*Advanced Information System for Multinational Drug Enforcement*), MultiMeteo (*Multilingual production of weather forecasts*) o ARISE (*Automatic Railway Information Systems for Europe*). Para obtener más información sobre los mismos se puede consultar el folleto *Language Engineering: Progress and Prospects*, editado por el equipo *LINGLINK* de la empresa *Anite Systems* [http://www.anite-systems.lu].

áreas como la lingüística computacional, la traducción automática, el inglés para fines específicos y la terminografía. En la siguiente tabla presentamos algunos ejemplos de córpora especializados:

TABLA 12: Ejemplos de córpora monolingües y multilingües especializados[60]

	NOMBRE	**LENGUA (S)**	**DOMINIO DE ESPECIALIDAD**
MONOLINGÜES	Cranfield collection	Inglés	Aerodinámica
	RAT (University of Reading Academic Text) corpus	Inglés	Artículos experimentales y tesis doctorales de diferentes dominios
	IBM corpus	Inglés	Informática (sublenguaje de los manuales de IBM)
	Medlars collection	Inglés	Medicina
	PSC (Pharmaceutical Sciences Corpus)	Inglés	Medicina y Farmacología aplicado a la oncología
	Medicor	Inglés americano	Medicina
	Corpus Textual del Español Periodístico	Español	Textos periodísticos
	LAN (Micro Focus SA)	Español	Técnica
	LEJES	Español	Derecho (textos académicos)
MULTILINGÜES	Danish-German-Spanish Corpus of Biotechnology	Danés-alemán-español	Biotecnología
	TEST corpus	Inglés-italiano	Debates parlamentarios
	CRATER (Corpus Resources and Terminology Extraction)	Francés-Inglés-Español	Telecomunicaciones
	Corpus textual plurilingüe especializado	Catalán-español-inglés-francés-alemán	Economía, derecho, medio ambiente, medicina e informática

Sin menoscabar la calidad y utilidad de los multilingües, estimamos que estos córpora no dan cuenta de la variedad textual que se produce en estos dominios de especialidad. Siendo esto así, queda mermada la enorme potencialidad que brindan los córpora en el estudio de los distintos géneros (Carne 1996: 135). Asimismo, el hecho de que contienen una importante proporción de textos traducidos hace que las colocaciones

[60] Para un estudio exhaustivo remito al ya citado Observatorio Español de Industrias de la lengua, a los mensajes de la lista de distribución *CORPORA*, de la que se puede obtener información en http://hd.uib.no, y a la bibliografía presentada en 5.1.

y convenciones retóricas asociadas a los tipos textuales de una determinada lengua aparezcan distorsionados por la influencia del texto origen. Por este motivo, creemos conveniente que estos córpora multilingües tengan dos componentes: un corpus de *textos paralelos* y otro de *textos comparables*.

En este proceso se han acuñado principalmente dos términos para designar los córpora multilingües: *corpus paralelo* y *corpus comparable*.

5.3. Córpora paralelos y córpora comparables

En la literatura de la lingüística de corpus, a la hora de referirse a los córpora multilingües aparecen los términos *corpus paralelo* y *corpus comparable*. Aunque no existe un consenso en el uso de estos términos, en la actualidad, su significado más extendido es el que presentamos a continuación.

Un *corpus paralelo* es aquel que presenta el mismo texto en más de una lengua, es decir, un texto y su traducción a una o más lenguas (Mc Enery 1994: 312; 1996: 58). Según Barlow, estos textos normalmente aparecen alineados para facilitar la búsqueda de equivalencias de traducción. Por ejemplo, el *Canadian Hansard corpus* y *Crater*, un corpus paralelo con textos en francés, inglés y español del dominio de especialidad de las telecomunicaciones.

En cuanto a los *córpora comparables* (Peters *et al.* 1996: 69), son un conjunto de textos en más de una lengua que, sin ser traducciones, por coincidir en el tema, motivación situacional y función comunicativa, proporcionan una excelente base para la comparación de dos o más lenguas. La mayoría de estos córpora representan un determinado sublenguaje. En cierto modo, estos son un conjunto de pequeñas colecciones de córpora monolingües que han sido recopilados siguiendo criterios muy parecidos. Como ejemplo podemos mencionar el corpus *Aarhus*, que contiene textos originales del subdominio especializado del derecho contractual en danés, inglés y francés. Algunos de los investigadores pioneros en la utilización de estos tipos de corpus como Hartmann (1996)[61], Laffling (1992) y Zanettin[62] (1994, 1996) los han

[61] Para evitar la ambigüedad, Hartmann denomina *bitextos* a los textos en el que uno es el texto origen y el otro, el texto término.

denominado córpora paralelos, siguiendo la terminología más usual en los estudios de traducción. En esta última disciplina, los textos paralelos son textos lingüísticamente independientes aunque funcionalmente equivalentes resultantes de una situación comunicativa idéntica o muy similar.

Los córpora paralelos fueron los primeros en aparecer. De hecho, hasta hace poco, la mayoría de los córpora multilingües contenían textos originales y su traducción, sobre todo en Europa, donde todas las lenguas oficiales de la Unión Europea gozan de un estatus igual. Hartmann (1980) fue uno de los primeros en hablar de textos paralelos, un término con el que incluía tanto los textos resultantes de una traducción como los que, sin ser traducciones, eran funcionalmente similares. En 1996, Hartmann distingue entre cuatro tipos de textos paralelos y los presenta en la tabla que presentamos abajo.

TABLA 13: Tipos de textos paralelos

	Approach	Parallel text type	Text corpus	
Metaphrase	Translation Text comparison	Bitext	'aligned', e.g. Canadian Hansard	'sampled' multilingual, e.g. MULTEXT
	Crosscultural Text adaptation	Twinned text	'normalised', e.g. ISSCO	
	Contrastive Text typology	Paired text	'domain-specific' e.g. corpus by Göpferich (1995)	
Paraphrase	Intralingual Text typology	Intertext	'register-diversified', e.g. BNC 'region-specific', e.g. ICE 'period-specific', e.g. Helsinki	

Teubert (1996: 245) también distingue entre distintos tipos de córpora paralelos. En primer lugar, aquellos que incluyen textos escritos en una lengua A y su traducción a una lengua B (y C...). Otro tipo lo forman textos que tienen una igual proporción de textos originales en lengua A y B y sus respectivas traducciones. Por último, Teubert

[62] En concreto, Zanettin utiliza el término *textos globalmente paralelos* y los define así: "collections of texts which are not translations, but rather texts which share a wide range of features, to serve as the basis for comparison and then for contrastive analysis between languages". (Zanettin, 1996:101).

habla de córpora paralelos compuestos exclusivamente de las traducciones a las lenguas A, B y C, de textos escritos inicialmente en una lengua Z.

Por otra parte, los córpora paralelos han sido criticados sobre la base de que no presentan el uso real del lenguaje natural sino textos influenciados por el texto origen, lo cual hace que no quede representada la lengua término en su forma más genuina (Hartmann 1994). Estos córpora son útiles en tanto que presentan ejemplos de equivalencias de traducción y de cómo el traductor resuelve problemas traductológicos, lo cual puede aportar datos relevantes sobre el proceso de la traducción.

Por eso, se empezaron a elaborar córpora bilingües/multilingües comparables. Sobre los criterios que garantizan la comparabilidad de los textos, Corazzari y Picchi (1994) consideran que es necesario que exista un vocabulario común. En consecuencia, proponen una combinación de criterios externos como dominio, tema y periodo temporal, y criterios internos, como la selección de textos que contengan una serie de palabras clave extraídas del vocabulario especializado de un sublenguaje. De ahí que, normalmente, los textos comparables contengan textos pertenecientes a lenguajes restringidos.

En cuanto a la conveniencia de utilizar uno u otro tipo de corpus en la traducción de textos, estimamos que ambos son complementarios y necesarios para llegar a comprender el proceso y el producto de la traducción[63]; y en esto coincidimos con Tercedor (1999) y Teubert (1996). En consecuencia, en nuestra investigación los datos emanarán de dos repositorios: por una parte, un corpus comparable de textos sobre cáncer de pulmón compuesto por una proporción similar de tipos textuales para el inglés y el español; y por la otra, un corpus paralelo formado por textos en inglés sobre el mismo tema y su traducción al español.

5.4. Aplicaciones de la lingüística de corpus a la traducción

Siguiendo la tendencia a la interdisciplinariedad de la que se alimenta la traducción, numerosos investigadores en el área de los estudios de traducción han defendido la

[63] No obstante en la traducción de términos culturales, resulta más apropiada la utilización de corpus comparables, ya que "es este tipo de corpus el que puede ayudar al traductor a reflejar en la lengua término no sólo el significado proposicional del texto sino también su valor semiótico dentro del contexto más amplio de la ideología y de la cosmovisión de una sociedad" (López Rodríguez 1998).

utilización de córpora informatizados (Baker *et al.* 1993, 2000; Barlow 1996; Peters y Picchi 1997; Hartmann 1994; López Rodríguez 1998; Muñoz Martín y Sánchez Trigo 1994; Teubert 1996; Zanettin 1996).

Baker *et al.* (1993) busca un punto de encuentro entre los partidarios de estudios descriptivos dentro de la traducción y las propuestas de la lingüística de corpus de extraer inferencias sobre el funcionamiento de la lengua basadas en datos reales. De hecho, Baker sostiene que las técnicas estadísticas y la metodología puestas en funcionamiento por Sinclair (1991) pueden contribuir al paso de estudios prescriptivos a descriptivos. De todas formas, Baker no quiere limitar esta metodología al estudio de la lengua origen y la lengua término, sino que pretende que se profundice en el conocimiento del proceso de la traducción mediante la búsqueda de tendencias en la traducción ya sean de carácter universal o establecidas por una norma (Toury 1980), el análisis de los distintos borradores de una traducción y la especulación sobre cuál es la unidad de traducción. Baker (2000) compara diversos rasgos lingüísticos tal y como se presentan en algunos tipos textuales del *BNC* y los compara con un corpus de traducciones hacia el inglés que contiene tipos textuales parecidos.

Michael Barlow, uno de los principales especialistas en córpora paralelos[64], también señala la utilidad de los córpora informatizados en la traducción:

> By looking at multiple instances of the translation of a word or structure, the idiosyncratic influences will be overpowered by the consistent and meaningful regularities. The accumulation of motivated translation choices revealed by a concordance program allow the general patterns to be perceived. Hence generalisations emerge from the aggregation of large numbers of individual instances". (Barlow 1996: 50)

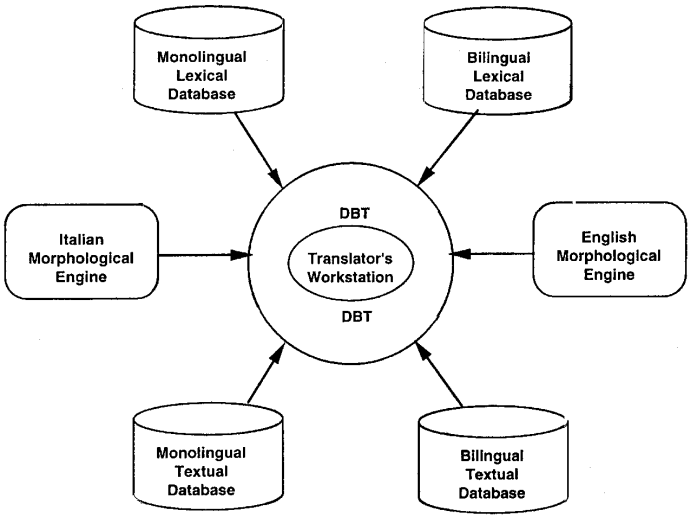
También menciona algunas de las áreas de dificultad dentro de la traducción que pueden esclarecerse con la utilización de córpora bilingües alineados: la fraseología, la polisemia, la organización del discurso y las estructuras retóricas (ibid: 52-53).

---

[64] De hecho, Barlow ha creado una página web llamada *Parallel Corpora* [http://www.ruf.rice.edu/~barlow/para.html] y un programa informático que crea concordancias multilingües y facilita la búsqueda de equivalencias de traducción.

Peters y Picchi (1997: 247) también reconocen lo provechosos que resultan avances producidos en lingüística de corpus y en lexicografía computacional como los léxicos electrónicos y córpora de referencia tanto para los profesionales de la traducción como para los académicos en el campo de los estudios de traducción. Los primeros necesitan encontrar equivalentes de traducción con mucha rapidez. Los segundos se plantean posibles realizaciones de un concepto en distintas lenguas al mismo tiempo que consideran una serie de factores como el uso, estilo, registro o dominio. Sus esfuerzos han cristalizado en una *estación de trabajo* de traducción humana asistida por ordenador para las lenguas italiana e inglesa desarrollada en el Instituto de Lingüística Computacional en Pisa. Esta consta de dos bases de datos (BD) léxicas y dos BD textuales, tanto monolingües como bilingües y dos repositorios con información morfológica sobre el italiano y el inglés. Un motor de búsqueda permite encontrar equivalentes, colocaciones, frases hechas e información morfológica, sintáctica y semántica a partir de los diccionarios y el corpus paralelo y comparable que sustentan el sistema.

DIAGRAMA 12: Prototipo de estación de trabajo para la traducción humana asistida por ordenador. (Peters y Picchi 1997: 270)



La traducción automática y la traducción asistida por ordenador también han encontrado un apoyo en la utilización de córpora. Laffling (1991, 1992), que Peters y

Picchi consideran uno de los primeros en defender la utilización de córpora paralelos, ha empleado un corpus de textos comparables en la construcción del lexicón de transferencia de un sistema de traducción automática, aunque él los denomina textos paralelos. Según Laffling, estos textos facilitan la traducción porque proporcionan al ordenador una base de conocimiento parecido al repertorio lingüístico del que dispone el traductor. También proporcionan información sobre las colocaciones más típicas en determinados tipos textuales y permiten la identificación de contextos equivalentes desde el punto de vista léxico que van a facilitar la traducción.

Por último, no sólo la teoría y la práctica de la traducción se han abonado con las conclusiones derivadas del estudio de córpora comparables. Algunos profesores de traducción (Muñoz Martín y Sánchez Trigo 1994; Zanettin 1994, 1996) han utilizado córpora comparables elaborados a partir de textos periodísticos en dos idiomas en la formación de futuros traductores. Según Zanettin (1996: 100), los textos periodísticos "proporcionan instancias de uso de la lengua actual en una gran variedad de tipos textuales y ofrecen información valiosa sobre la cultura y la sociedad". Este autor ha ensamblado los textos en una base de datos textual de consulta para los alumnos de la Escuela de Traductores e Intérpretes de Bolonia.

5.4.1. La utilización de córpora en la traducción biomédica

La ventaja de utilizar una metodología de corpus en la traducción biomédica reside en que estos contienen textos que proporcionan información que los diccionarios bilingües no contienen. Desde que Tomaszczyk (1989: 180) observara que los diccionarios especializados no son adecuados para la traducción científica y técnica, son muchas las voces que han destacado las limitaciones de estos diccionarios, cuanto más si son bilingües (Faber 1998b, Tercedor 1999, Williams 1996). En efecto, estos suelen presentar equivalentes de traducción y omiten información fundamental en el nivel oracional y supraoracional como los términos y unidades fraseológicas más apropiados en determinados registros y géneros, información gramatical, estilística y retórica.

Por este motivo, parece obvia la necesidad de otras fuentes de información. Aparte de la consulta a especialistas, Williams (1996: 296) sostiene que los textos paralelos—en nuestra terminología, textos comparables—son los que aportan una

información más fidedigna al traductor. De ahí que sostenga que un método combinado de diccionarios especializados y, al menos, tres textos paralelos pueden satisfacer el 85-90% de las necesidades del traductor. Siendo esto así, si a la hora de traducir el traductor dispusiera, no de tres textos, sino de un conjunto de textos tanto paralelos como comparables en soporte electrónico, su labor se agilizaría significativamente.

No obstante, han sido escasos los córpora biomédicos compilados hasta el presente y se han realizado únicamente en una lengua. Entre estos podemos mencionar una serie de córpora biomédicos en inglés:

a) El corpus Medlars, que consta de 1 033 *abstracts*, por lo cual sólo se puede considerar representativo de este tipo textual
b) El *PSC (Pharmaceutical Sciences Corpus)*, que cuenta con 150 artículos experimentales sobre cáncer publicados en veinte revistas médicas y farmacéuticas entre 1990 y 1993.
c) El Corpus *Medicor* de textos médicos en inglés norteamericano actual, recopilado en la Universidad de Helsinki[65]. Incluye los siguientes tipos textuales: artículos experimentales, muestras de libros de texto, artículos de divulgación, editoriales de revistas especializadas, manuales para profesionales, folletos informativos / guías (*popular guidebook*).

También podemos mencionar algunas fuentes que pueden proporcionar un acopio de textos biomédicos en soporte electrónico. Una de ellas es el *BNC*, que, por su gran extensión, contiene una importante representación de textos biomédicos, etiquetados y listos para cualquier análisis. Un vistazo a la base de datos bibliográfica de este corpus puede facilitar la selección de textos de este tipo[66]. Así, el *BNC* incluye

[65] Su compiladora, Minna Vihla, ha elaborado este corpus con el objetivo de estudiar la modalidad en textos médicos. Por este motivo, ha primado la representatividad en cuanto a tipos textuales incluidos sobre la cantidad, 397 311 palabras.

[66] Esta idea fue sugerida en la lista de distribución *Corpora* por David Lee (Departamento de Lingüística de la Universidad de Lancaster) el 23 de octubre de 1998 como respuesta a una pregunta sobre la existencia de córpora de textos científicos.

bajo las categorías *ciencia aplicada* y *ciencia pura,* textos pertenecientes a las siguientes publicaciones periódicas:

- *Journal of Gastroenterology* (713 164 palabras)
- *The Lancet* (135 850)
- *British Medical Journal* (449 961)

Asimismo, muchas revistas médicas ponen a disposición de sus suscriptores una versión electrónica[67] y hay numerosos sitios sobre medicina en Internet, lo cual facilitará cada vez más la elaboración de córpora biomédicos comparables y paralelos.

Siempre que el análisis sea el adecuado, estos córpora informatizados pueden corroborar o corregir los resultados obtenidos en estudios previos basados en reducidos córpora textuales médicos y completar la literatura sobre los patrones léxicos, colocacionales y retóricos presentes en textos médicos en inglés y en otras lenguas. De estos estudios previos, nos gustaría destacar el trabajo en terminología realizado por Thomas y Hawes (1994), que analiza los complementos del verbo, y las investigaciones en el análisis del discurso de artículos experimentales médicos realizados por Adams-Smith (1984), Busch-Lauer (1995), Kretzenbacher (1990), Myers (1990, 1992), Nwogu (1991, 1997) y Salager-Meyer (1989, 1994). Ya revisamos algunos de estos estudios en 3.3.

En este área también podemos situar la excelente labor de Gledhill (1996, 1997), compilador del anteriormente citado *PSC* (*Pharmaceutical Sciences Corpus)*. Tanto el diseño del corpus como su extensión, 500 000 palabras, hacen que se pueda considerar representativo de este subdominio. Su objetivo es el de detectar elementos que son especialmente significativos en las secciones retóricas de artículos experimentales[68] y representar las variaciones de la fraseología propia en cuanto a sus elementos gramaticales a lo largo de estas secciones.

Resulta difícil citar investigaciones similares que se enmarquen en los estudios de traducción, a no ser las de Gledhill (1996, 1997), Tercedor (1999), García de

[67] Existe un exhaustivo listado de publicaciones médicas en soporte electrónico disponible en www.gen.emory.edu/MEDWEB/medweb.html.
[68] Es decir, título, resumen, introducción, métodos, resultados y discusión.

Quesada (tesis doctoral en curso), López Rodríguez (2000) y Pérez Hernández (2000), que giran en torno al subdominio de la oncología. Tercedor estudia las variaciones en la fraseología de tipo léxico en inglés y español correspondientes a los registros más usuales en este subdominio.

Como punto final a este breve repaso de cómo la lingüística de corpus puede ser un valioso instrumento en los estudios de traducción, se hará una aclaración. Si bien esta metodología empírica y cuantitativa ha supuesto una superación de la lingüística basada exclusivamente en la introspección y la intuición en favor de una lingüística basada en datos auténticos (Biber y Finegan 1991), no obstante, no se debería considerar la lingüística de corpus como una panacea. Los estudios lingüísticos, tanto los basados como los guiados por los datos extraídos de un corpus, presentan algunos puntos débiles como su dependencia de la labor de abstracción por parte del lingüista, no exenta de subjetivismo, y el hecho de que la calidad del estudio depende en gran medida de la calidad y amplitud del corpus (Márquez 1998: 157). Estas limitaciones son aún mayores cuando el objeto de estudio, la cohesión, es un fenómeno que se manifiesta a lo largo de todo un texto y tiene un alcance intraoracional e interoracional.