9. Selección del corpus

9.1. Criterios de selección del corpus

Desde los comienzos de la lingüística de corpus se ha especulado sobre los criterios que hay que tener en cuenta a la hora de recopilar un corpus. Renouf (1984: 10) cita los aspectos considerados al elaborar el corpus en el que se basaría el diccionario COBUILD: fecha de publicación, inclusión de textos escritos de amplia difusión en diversos medios, variación dialectal, género y tema. Estos criterios se han ido refinando posteriormente para atender también a criterios como sexo, edad y nacionalidad del autor del texto, factualidad del texto, participantes en la comunicación, etc. (Atkins *et al*. 1992, Clear 1993, Sinclair 1991), aunque referidos principalmente a córpora de la lengua general.

En relación a los córpora de la lengua general han surgido debates en torno a si en la selección del corpus debe primar la cantidad o la calidad, entendida como *equilibrio* en la composición del corpus (Church y Mercer 1993: 17-18). Este debate se ha orientado hacia el predominio de la calidad, gracias al iluminador artículo de Biber (1993), en el que se expone que el principal criterio es la *representatividad*. Esta está relacionada con el grado en el que la muestra incluye todo el espectro de variabilidad de una población. Dicho espectro abarca los tipos textuales de una lengua y la distribución de sus recursos lingüísticos. Por ello, para Biber, el muestreo debería seguir criterios estadísticos. Sin embargo, como apunta Kennedy (1998: 74), la población suele ser tan amplia y los tipos textuales tan difícilmente delimitables, que un muestreo aleatorio es poco factible.

Asimismo, aunque se tenga en mente el criterio de la representatividad, el investigador debe tener presente que los resultados obtenidos pueden ser significativos desde dos puntos de vista: con referencia a un corpus concreto o con referencia a la lengua como conjunto (Hanks 2000: 313). Esto último es bastante inusual. Somos conscientes de que, dada la profusión de textos sobre medicina tanto en soporte papel como electrónico, difícilmente podemos recopilar un corpus representativo. Por ello, creemos que hay que añadir otros criterios que reflejen que estamos ante un corpus especializado, que va a ser analizado con herramientas informáticas y que pretendemos que sea útil desde el punto de vista de la traducción y de la terminología. Entre estos

criterios podemos citar la función que va a tener el corpus (el estudio de la cohesión en textos biomédicos), los usuarios (el investigador y el público que visite el servidor *web* del sistema bilingüe de información y recursos oncológicos *Oncoterm*), la inclusión de textos escritos, la disponibilidad de textos en soporte electrónico, la reutilizabilidad y la necesidad de colocar los textos recopilados en contexto (Hockey y Walker 1993: 239-240), en concreto, en el marco de referencia que hemos llamado *EVENTO ONCOLÓGICO*. Añadimos a estos criterios dos más propuestos por Tercedor (1999: 117): la variedad comunicativa en cuanto a intenciones comunicativas y tipo de lector, y la necesidad de incluir tanto textos originales como traducidos.

9.2. Limitaciones del corpus utilizado

La función del corpus es la de reflejar usos del lenguaje biomédico para sistematizar y computar los recursos cohesivos. Aún cuando se disponga de un corpus en soporte electrónico, este objetivo requiere trazar relaciones cohesivas no entre las unidades léxicas y terminológicas de todo el corpus, sino en cada texto como unidad. Esta premisa impone que se tome como unidad de muestreo (*sampling unit*) el texto completo, en el que se objetivarán las instacias de cohesión. Al analizar la cohesión dentro de cada texto, cualquier inferencia que se haga requiere la comparación de los datos de cada texto individual con el resto, una perspectiva nada usual dentro de la lingüística de corpus.

Por otra parte, la cohesión de un texto es extremadamente sensible a las variaciones de contenido y de audiencia, por lo que para comparar diferencias cohesivas ha sido necesario encontrar textos sobre exactamente el mismo tema. Como el factor que más influye en la cohesión es el conocimiento compartido del emisor y el receptor, ha sido preciso trabajar con parejas de textos sobre un tema muy restringido de la oncología. Por tanto, para el análisis extensivo, se ha decidido seleccionar exclusivamente textos sobre el tratamiento del cáncer de pulmón, lo cual ha hecho que se descarten muchos de los textos seleccionados inicialmente. En el momento de comenzar la investigación, muy pocos textos cumplían este requisito.

También ha supuesto una limitación la utilización de un generador de cadenas léxicas en fase experimental del que sólo se han podido obtener resultados gracias a la

buena voluntad de su creador, Jeremy Ellman. Para no abusar de esa buena voluntad, los textos que se enviaron para ser analizados han tenido una extensión razonable, en concreto, 106 922 palabras.

Por último, la disponibilidad de textos traducidos sobre el cáncer ha sido menor de lo que pensábamos en un principio. Hoy en día el mercado profesional de traducciones del inglés al español es cada vez más reducido porque se supone que los especialistas leen las publicaciones internacionales en inglés. Lo habitual es que estos contraten al traductor para que redacte en inglés el resultado de sus investigaciones o haga de revisor de la traducción. La proporción mayor de textos traducidos la constituye la versión española de sitios *web* sobre oncología de Estados Unidos.

A pesar de estas limitaciones, reconocemos la validez de aplicar al estudio de la cohesión los datos auténticos que aporta un corpus formado por textos completos, en lugar de servirnos de oraciones aisladas o párrafos, como se ha hecho en muchos estudios previos sobre cohesión.

9.3. Composición del corpus

En esta tesis se ha trabajado tanto con un corpus comparable de textos originales sobre cáncer de pulmón en inglés y en español como con un corpus paralelo sobre el mismo tema. De esta forma se pueden delimitar las diferencias con respecto a la cohesión que responden al proceso traductor en sí. En la composición del corpus, los textos destinados al análisis extensivo forman una unidad aparte.

El análisis extensivo se centra en un corpus de textos sobre cáncer de pulmón que cuentan en su forma final con un total de 106 922 palabras, aunque antes de modificarlos para el análisis contaban con más[113]. Han sido extraídos de *CancerNet*[114], la página de Internet del *National Cancer Institute,* que es la principal organización norteamericana contra el cáncer. Esta página contiene información sobre el cáncer en forma de folleto explicativo en las áreas del tratamiento, prevención, detección precoz,

[113] El corpus contaba con 153097 palabras inicialmente. Sin embargo, para utilizar *Hesperus* hemos tenido que suprimir las referencias bibliográficas de los textos para oncólogos porque distorsionaban los datos. También distorsionaban los índices y la información sobre el PDQ que aparecía en todos y cada uno de los textos, por lo que hemos tenido que eliminarlos de los textos.

[114] La dirección es http://cancernet.nci.nih.gov

cuidado médico, datos estadísticos y oncología experimental (ensayos clínicos)(véase Apéndice II, *Textos extraídos del PDQ*).

En concreto, hemos trabajado con una sección de resúmenes llamada *Physician Data Query* (*PDQ*) por una serie de motivos. En primer lugar, el *National Cancer Institute* proporciona a un número muy elevado de visitantes una cantidad ingente de información, que se actualiza constantemente con rigurosidad, y cuenta con el prestigio que confiere el hecho de que es referencia obligada para muchos oncólogos. En segundo lugar, los textos de *CancerNet* tienen el interés de estar redactados con miras a dos tipos de lectores con un conocimiento distinto del tema: médicos especialistas y pacientes o familiares de pacientes. Por eso, esta página tiene dos secciones: *CancerNet Health Professionals* y *CancerNet Patients.* Teniendo en cuenta que la cohesión de un texto es extremadamente sensible a las variaciones de contenido y de audiencia, es un privilegio disponer de textos sobre exactamente el mismo tema en los que las variaciones con respecto a la cohesión se deben sobre todo a cambios en el conocimiento de los lectores de los mismos. Por otra parte, estos textos aparecen en la misma página *web* traducidos al castellano, lo cual nos permitirá comparar cambios con respecto a la cohesión dependientes de las lenguas con las que se trabaja y que son resultado de la traducción. La separación temporal entre los textos originales y traducidos oscila entre un mes y tres meses.

Se someterán 36 textos del *PDQ* al programa *Hesperus* para localizar las principales cadenas cohesivas de los mismos. De estos, 18 están destinados a especialistas (*Cancernet Health Professionals*) y otros tantos a pacientes y a familiares de pacientes (*Cancernet Patients).* Como diferencias más significativas entre unos y otros podemos destacar que los primeros presentan una mayor extensión y complejidad léxica, las cuales ilustran una mayor complejidad conceptual. De ahí que en su forma original contaban con bibliografía sobre cáncer de pulmón. Ambas secciones comparten dos de los apartados en los que se agrupa la información sobre cáncer de pulmón: uno, centrado en el tratamiento (*Cancer Treatment Information Summaries*) y otro, sobre cuidados médicos (*Supportive Care and Advocacy Issues).* Los textos sobre tratamiento están más centrados en el cáncer de pulmón, mientras que los textos sobre cuidados

médicos tratan sobre trastornos y patologías asociadas, que no se restringen exclusivamente a este tipo de tumor.

Por otra parte, los ejemplos que se presentan a lo largo de la tesis para ilustrar las funciones léxicas, para la clasificación de categorías de cohesión léxica, para el análisis intensivo y en el capítulo sobre cohesión y traducción han sido extraídos de las fuentes que se especifican en la tabla 38. En total computan 1 397 515 palabras, 1241780 correspondientes al corpus comparable y 155 735 de los textos paralelos que no están incluidos en la cifra anterior (en realidad, el corpus paralelo cuenta con 417 822). Dado que los recursos en soporte electrónico son más numerosos en lengua inglesa, la parte en inglés (893068) ocupa una proporción mayor del corpus bilingüe en comparación con el español (509 447).

En la selección de las fuentes, se ha intentado cubrir el espectro definido por Lévy-Leblond (1996) y Tercedor (1999: 172). Es decir, distinguimos entre *intercambio especializado*, donde se incluyen las publicaciones especializadas, las páginas *web* destinadas a profesionales de la salud y los manuales; *intercambio público*, que aglutina las publicaciones de divulgación semiespecializada y general y los folletos de salud pública en Internet; e *intercambio familiar*, donde se han incluido los correos electrónicos de la lista de distribución *LUNG-ONC*, en la que participan pacientes de cáncer de pulmón y sus familiares (LUNG-ONC@LISTSERV.ACOR.ORG).

TABLA 38: Composición del corpus

CORPUS COMPARABLE		
	INGLÉS (821 456 palabras)	**ESPAÑOL** (420 324 palabras)
Intercambio especializado	Publicaciones especializadas: *British Medical Journal* *Cancer* *CANCERLIT* *C-A. A Cancer Journal for Clinicians* *Lancet* *Medline* *The New England Journal of Medicine*	Publicaciones especializadas: *Archivos Bronconeumológicos* *Asociación Española de Endoscopia Respiratoria* *Medicina Clínica* *Neoplasia* *Revisiones en cáncer*
	Páginas *web* destinadas a profesionales de la salud: *CancerBacup* *Physician Data Query* (*PDQ*) *PSL* *Virtual Hospital*	Páginas *web* destinadas a profesionales de la salud: *Atheneum / Jano, Medicina y Humanidades.* *Diario Médico* *Sociedad Iberoamericana de Información Científica*
	Manuales: *Cancer: Principles and Practice of Oncology*	Manuales: *Medicina Interna* *Oncología Médica. Guía de Oncología Médica*
Intercambio público	Publicaciones de divulgación especializada: Monográficos *Scientific American* (monográfico)	
	Publicaciones de divulgación general: Monográficos *TIME* Noticias sobre oncología: *Daily Telegraph* *Oncolink (CNN Health and Food)* *Reuters Health* Publicadas en *Bacup* y la *web* del *NCI*	Publicaciones de divulgación general: Monográficos *Blanco y negro* *El Semanal* *Muy interesante* *QUO* Noticias sobre oncología: *El Mundo (CD y versión on-line)* *Ideal (Suplemento Campus)*
	Folletos de Salud Pública en Internet para pacientes y familiares: *Alcase* *American Cancer Society* *Cancer Bacup* *American Lung Asociation* *CancerHelp UK* *Imperial Cancer Research Fund* *PDQ* *P. Connelly Lung Cancer Page* *Oncolink* *Virtual Hospital*	Folletos de Salud Pública en Internet para pacientes y familiares: *Ciencia y Cáncer* *Instituto Madrileño de Oncología* *PDQ* *Sarenet*

Int. familiar	Correos electrónicos de la lista de distribución *LUNG-ONC*	
CORPUS PARALELO		
	INGLÉS (71 612 palabras)	**ESPAÑOL** (84 123 palabras)
Intercambio especializado	Publicaciones especializadas *Medicina Clínica (abstract)* *Neoplasia (abstract)*	Publicaciones especializadas *Medicina Clínica (abstract)* *Neoplasia (abstract)*
	Páginas *web* destinadas a profesionales de la salud: *American Cancer Society* *American Lung Asociation* *PDQ*	Páginas *web* destinadas a profesionales de la salud: *American Cancer Society* *American Lung Asociation* *PDQ*
	Manuales: *Cancer: Principles and Practice of Oncology* Merck Manual	Manuales: *Cancer: Principios y Práctica de Oncología* *Manual Merck*
Intercambio publico	Publicaciones de divulgación especializada: Monográficos *Scientific American*	Publicaciones de divulgación especializada: Monográficos Investigación y Ciencia
	Enciclopedias *Encarta*	Enciclopedia *Encarta*
	Folletos de Salud Pública en Internet para pacientes y familiares: *American Cancer Society* *American Lung Asociation* *PDQ*	Folletos de Salud Pública en Internet para pacientes y familiares: *American Cancer Society* *American Lung Asociation* *PDQ*

9.4. Tratamiento del corpus: eliminación de palabras forma y lematización

Para extraer las *palabras plenas* (*content words*) que crean cohesión, serán de utilidad las listas de frecuencia correspondientes al corpus de 36 textos sacados del *PDQ* (en adelante, *Corpus PDQ*). Para ello, es necesario eliminar de estas listas las palabras forma y aglutinar las diversas variantes morfológicas derivadas de la misma raíz (lematización). Con este fin se ha creado una *stoplist* de 256 palabras, es decir, una lista con las *palabras forma* que son muy frecuentes tanto en nuestro corpus como en córpora de la lengua general. Teniendo en cuenta que nuestros textos están escritos en inglés norteamericano, la hemos elaborado a partir de las 150 palabras más frecuentes del *Corpus Brown* (véase Apéndice VIII). De estas se han seleccionado las palabras

forma y se ha comprobado que estas coinciden con las del *BNC* excepto en la frecuencia, que les confiere un orden distinto. También se han añadido las *palabras forma* que encabezaban la lista de frecuencia del *Corpus PDQ*.

Por otra parte, como en esta tesis es fundamental el estudio de la representación conceptual, se llevará a cabo un proceso de *lematización*. Se ha utilizado como base la lista de lemas elaborada por Yasumasa Someya[115], a la que se han añadido algunos lexemas con su flexión. De todas formas, hemos preferido la lematización manual para evitar errores ocasionados al lematizar.

[115] La autora de esta lista, cuyo correo electrónico es ysomeya@gol.com, la ha puesto a disposición de cualquier visitante de la página *web* de Mike Scott en http://www.liv.ac.uk/~ms2928/index.htm.