Capítulo 1

El concepto de corpus lingüístico

Ya desde los estudios tempranos en adquisición del lenguaje, con el fin de hacer un análisis cualitativo de la información, el investigador ha tenido que recolectar una serie de datos lingüísticos. Se tiene registrado entre los primeros trabajos, a fines del siglo XIX, la obtención de un corpus para hacer análisis cuantitativo de la frecuencia y secuencia de las letras en alemán. Tanto en los análisis cualitativos como en los cuantitativos se requiere trabajar con muestras que brinden datos reales de diferentes textos, ya sea orales o escritos. Con cualquiera de sus nomenclaturas, los corpus lingüísticos o los corpus de textos constituyen, de algún modo, un modelo de la realidad lingüística que se quiere observar.

Cualquiera que sea el tamaño o extensión del corpus, el procesamiento de estos textos fue, hasta antes del uso de la informática en la década de los cincuenta, de manera manual. Los investigadores tenían que hacer fichas de papel, de manera que tareas simples, como la búsqueda de ocurrencia de palabras, resultaba ser un trabajo muy laborioso que se realizaba mediante lectura directa en cada una de las fichas.

1.1. Definición de corpus lingüístico

Definir el concepto de corpus en la actualidad no es tan simple como parece. Desde el punto de vista etimológico, corpus proviene del latín y significa cuerpo, donde el cuerpo es texto. Antes de entrar en la definición, conviene hacer mención al uso del término mismo. Es curioso que proviniendo del latín, la Real Academia de la Lengua acepte el nombre de manera indistinta para el singular y para el plural, en tanto que en el inglés se diferencia corpora para el nombre plural. Si bien algunos colegas de habla hispana diferencian corpus de corpora, en este libro se adoptará la postura general para el español, no obstante se conservará el término corpora cuando venga así referenciado de algún otro autor.

Ahora bien, de manera amplia se puede definir que un corpus lingüístico consiste en un conjunto de textos de materiales escritos y/o hablados, debidamente recopilados para realizar ciertos análisis lingüísticos. Una definición de esta naturaleza, sin embargo, requiere realizar algunas precisiones.

Conjunto de textos

Los textos que conforman un corpus deben ser representativos y se compilan según criterios lingüísticos que les permiten ser analizados.

Retomando la definición general, en primera instancia se habla de un conjunto de textos, de manera que puede constituirse por uno o varios libros, una revista o simplemente por un artículo de periódico; puede tratarse de un texto científico, de uno literario o incluso de los mensajes enviados entre dos personas mediante computadoras en línea; se puede hacer un compendio sincrónico o diacrónico de la lengua, referirse a la obra entera de un autor, a solo una de sus obras, o bien al habla de un niño en la etapa temprana de adquisición de su lengua materna; es decir, los corpus pueden formarse con cualquier tipo de texto. A menudo se usan textos individuales para muchos tipos de análisis literario y lingüístico, como ocurre con el análisis estilístico de un poema o con el análisis de la conversación de una muestra de charla de televisión. Con este último caso se aborda otro punto interesante de la definición, pues los textos también pueden proceder de material hablado, aunque en este libro se dedicará especial atención a los corpus de procedencia escrita.

Un corpus puede estar formado por cualquier tipo de texto, es decir, por uno o varios libros, una revista o simplemente por un artículo de periódico. Asimismo, los textos que lo conforman pueden pertenecer a cualquier género textual, esto es, pueden ser literarios, científicos o de lenguaje coloquial.

Debidamente recopilados

Un segundo punto importante para la definición de corpus lingüístico se relaciona con la recopilación de los textos, pues el concepto de corpus como tal puede resultar ambiguo. Una biblioteca, sea cual sea su forma, su tamaño o su tipo, sea digital o de material impreso, no constituye un corpus como tal, independientemente de que se trate de textos escritos bien diferenciados por estratos o marcas culturales, geográficas, dialectales, temáticas o de cualquier otra naturaleza. No obstante, una selección cuidadosa de documentos de esta biblioteca, escogidos con criterios bien delimitados para su posterior análisis, como se verá más adelante, y con los textos capturados adecuadamente, constituirá el primer paso en la construcción de un corpus.

Para análisis lingüísticos

El tercer aspecto a señalar en la definición de corpus se refiere al objetivo, esto es, a la realización de análisis lingüísticos. Los análisis que pueden realizarse en un corpus son de dos tipos: cualitativos y cuantitativos. En los primeros se estudian las características de una lengua o algún fenómeno ocurrido en ésta. En los segundos, todo lo competente a cifras numéricas, frecuencias de aparición, etc. del fenómeno que se estudie en el corpus que se tenga. Por ejemplo, si se quisiera hacer un estudio cualitativo de la correlación grafía-sonido del fonema /s/ en el español del siglo XVI, tendría que describirse en qué contextos se empleaba determinada grafía; en tanto que un estudio cuantitativo indicaría la frecuencia de aparición de cada tipo de grafía.

Tanto en los análisis cualitativos como en los cuantitativos está muy difundido el término corpus para describir el material sobre el cual se realizan las investigaciones: es un conjunto de datos reales y aceptables, debidamente ordenado, codificado y organizado, de diferentes textos recopilados, pertenecientes a un código lingüístico determinado, oral o escrito. Con cualquiera de sus nomenclaturas, los corpus lingüísticos o los corpus de textos constituyen, de algún modo, un modelo de la realidad lingüística que se quiere observar, mas no la realidad misma. Es mediante criterios lingüísticos pertinentes que un corpus no sólo contiene el texto en sí mismo, sino que además proporciona información que facilita su análisis.

Con el innegable valor que constituyen los corpus para las investigaciones lingüísticas, cabe extender su empleo más allá del alcance de las mismas. Como se verá ampliamente en el último capítulo, su uso es tanto para la lingüística teórica como para la aplicada, pero también para las investigaciones y desarrollos en las llamadas tecnologías del lenguaje. Hoy por hoy podemos decir que prácticamente no hay estudios de dichas tecnologías sin el uso de corpus. Cada vez el uso es más extendido y es posible encontrar diversos corpus para distintas áreas en particular. En tanto que antes era un concepto que se limitaba al ámbito de la ciencia del lenguaje, ahora abarca otras disciplinas y se aplica en múltiples tareas.

Ya después de varios años, hoy existen corpus muy potentes, algunos de los cuales pueden ser consultados gratuitamente para múltiples investigaciones, desarrollos y aplicaciones:

En lingüística. El uso de corpus permite realizar estudios que abarcan los distintos niveles de análisis de la lengua: fonético, fonológico, morfológico, sintáctico, semántico y pragmático.

En lingüística aplicada. Los corpus se emplean en áreas como enseñanza de lenguas, análisis del discurso, patologías del lenguaje, lexicografía, terminología, tra-

ducción y lingüística forense, por sólo mencionar algunas.

En tecnologías del lenguaje. Dado que las tecnologías del lenguaje procesan voz y texto, el uso de corpus se extiende para crear sistemas de diálogo, generadores de documentos, recuperadores y extractores de información, traductores y resumidores automáticos, entre muchos otros.

1.2. La lingüística de corpus

A pesar de que el concepto corpus no tiene que ver directamente con la informática, en la actualidad está estrechamente relacionado con ella. Los progresos computacionales y tecnológicos han permitido manejar los corpus con programas informáticos apropiados, lo que proporciona un excelente material para el trabajo de investigación. Por ello, es pertinente hablar de corpus informatizados, es decir, un conjunto de textos elegidos y anotados con ciertas normas y criterios para el análisis lingüístico, de forma que se sirven de la tecnología y de las herramientas computacionales para generar resultados más exactos.

Tomado en cuenta la labor de la lingüística basada en textos, conviene especificar que un corpus es, sobre todo, una colección de textos en soporte informático. Esta colección, si bien llega a ser muy extensa, incluso de varios millones de palabras, puede ser también de dimensiones mínimas. Por extensión, se ha llamado lingüística de corpus a la parte de la lingüística en la que se estudian con medios informáticos de diferentes tipos grandes masas de datos, inabordables de otro modo, para obtener de su análisis, por ejemplo, las características lingüísticas de una lengua en un cierto momento de su historia, de cierto tipo de textos, de un conjunto de autores o un autor determinado, etc.

Con el nacimiento de la informática es posible el análisis de textos de manera eficaz, ya que permite llevar a cabo cálculos complejos en cuestión de segundos y sin los errores naturales de cualquier persona. A pesar de las ventajas, hoy en día existe cierta renuencia al uso de las computadoras, una antipatía que se debe más al miedo infundado a la tecnología que a considerarlas falibles o usurpadoras de la labor humana. Por el contrario, conviene tomar en cuenta que la posibilidad de falla del humano es mayor que la computadora cuando se trata de labores sistemáticas y de alta precisión. Sólo con el fin de mostrar los equívocos que pueden ocurrir durante el procesamiento manual de ciertas tareas, en el cuadro siguiente se encuentra un texto corto que aparece en Internet en el que se pide al lector contar el número de veces que aparece la letra “F”.

FINISHED FILES ARE THE RE-
SULT OF YEARS OF SCIENTIF-
IC STUDY COMBINED WITH THE
EXPERIENCE OF YEARS

Del ejercicio del cuadro anterior, es fácil comprobar con diferentes personas que el total de “F” va de tres a seis. Esto se debe a que el ser humano muchas veces pasa por alto las preposiciones o los artículos, en este caso las tres ocurrencias de “OF”. Este error, trasladado a análisis lingüísticos más extensos, les restaría por supuesto confiabilidad y sentido.

1.3. Características que debe cumplir un corpus

Existen corpus que se construyen en función de requerimientos específicos para ciertos proyectos particulares como puede ser el análisis estilístico de la obra de Octavio Paz o el alineamiento de las actas de una reunión bilingüe para encontrar la equivalencia de los nombres propios; por el contrario, en otras investigaciones se buscaría recolectar un estrato mayor de lengua para la realización de diversos estudios lingüísticos a partir de métodos empíricos, como sería el caso del desarrollo de un sistema de diálogo en un dominio muy concreto, en donde intervienen estudios fonéticos y acústicos, sintácticos, semánticos y pragmáticos. Algunos estudios requieren de anotaciones muy precisas, mientras que otros abarcan un panorama más amplio de códigos lingüísticos. Cualquiera que sea el estudio, los corpus comparten características dominantes que les permiten ser concebidos como tales y ser diferenciados de otras colecciones de textos que no lo son.

El concepto de corpus lingüístico ha sido discutido desde varias perspectivas, lo que ha suscitado controversias y diferencias en donde se ponderan unos criterios frente a otros, a veces sin llegar a un acuerdo. En el extremo clásico de la lingüística empírica, todo registro sobre hechos observables puede constituir un corpus, incluyendo las transcripciones en un cuaderno sobre el habla espontánea de una niña de cuatro años, en donde se anotan con círculos y cuadros los hechos que se quieren destacar. Siguiendo la misma línea, se puede argumentar que cualquier colección de textos lo es, ya sea desde aquellos en los que se quiere extraer cierta información, como sería la recopilación de la terminología de un texto de especialidad, o de aquellos que van a ser objeto de estudio, como sería una serie de oficios y memorandos a los que se les quiere identificar la estructura para elaborar modelos de uso en una compañía, por ejemplo. Como tal, estas son muestras de lo que normalmente se ha llamado corpus, es decir, material de registro para la realización de un trabajo en particular. Si bien parece haber acuerdo con esta

perspectiva, todavía el concepto está lejos de lo que se denomina corpus lingüístico, y no propiamente porque no sean textos para estudios lingüísticos, sino por los criterios con los que están construidos.

Contención de datos reales

Los datos proporcionados en los textos que conforman un corpus deben mostrar a pequeña escala cómo funciona una lengua natural. A pesar de las diferencias que se puedan suscitar entre los estudios empíricos y los intuitivos, es necesario reconocer que un corpus constituye un reflejo de la realidad y que la lengua, sea oral o escrita, puede ser modelada mediante los corpus, ya que en ellos se tienen pruebas fehacientes y constatables de actos de habla que se convierten en fuentes invaluables de datos para la investigación lingüística empírica y para las aplicaciones de la lingüística aplicada. Se puede ser tan específicos en el estudio que es válido hacer un tratado completo de la obra de un autor o incluso trabajar con sólo una de sus obras. Sin embargo, en la lingüística de corpus se pretende algo más que un análisis tan específico, muchas veces se busca que el material recopilado sirva como una referencia representativa para una variedad de lenguaje en particular. Aunque un solo acto de habla puede argumentarse como verdadero, no es suficiente para generalizarlo y extenderlo a otros hablantes. La cuantificación en la lingüística de corpus es más significativa que la cuantificación en las otras formas de lingüística empírica porque las conclusiones son válidas para una mayor población, antes que para la muestra sola, con lo que puede constituirse una hipótesis en torno a la organización y funcionamiento de una lengua natural en un determinado contexto. De esta manera se abre un abanico de posibilidades en las aplicaciones. Por ejemplo, en la enseñanza de lenguas que basa su instrucción en los modelos verdaderos de uso, es posible ahora dotar a los estudiantes con las herramientas que necesitan para producir y comprender la lengua apropiadamente en cualquier tipo de contextos.

Representatividad

A fin de poder asegurar que las conclusiones que se obtengan de un estudio puedan ser generalizadas, el primer aspecto a considerar en el diseño de un corpus lingüístico es la obtención de muestras representativas de las lenguas o lengua en cuestión, debido a que un corpus siempre es una muestra de lengua y no pretende ser la totalidad de ella. Para ello, el primer paso es definir la población que se desea estudiar, lo que tiene que estar acorde con los objetivos del proyecto. En tanto en el objetivo se defina una población específica será posible considerarla así, pero para hacer generalizaciones y presentar el estado de una lengua, debemos ser cuidadosos en escoger la población para que sea representativa, esto es, para que la variabilidad de la muestra represente lo más aproximado posible a la población en los aspectos que se desean estudiar. En investigaciones de adquisición del lenguaje materno para niños de tres años de edad

en zonas rurales, la población estará entonces específicamente restringida y habrá que buscar informantes de esta edad que cumplan con estos requerimientos, y de ninguna manera conformarse con escoger tres niños que vivan en zonas rurales. Asimismo, en casos como el análisis de frases nominales en el Siglo XVI no se puede uno restringir únicamente a la obra de Bernal Díaz del Castillo, no obstante sean las crónicas más completas sobre la conquista de México, pues el habla de un autor no es reflejo fiel del resto de sus contemporáneos. Por tanto, hay que tener cuidado en los resultados que se exponen y evitar hacer generalizaciones a partir de aspectos particulares; por ejemplo, concluir que grabaciones del habla de los estudiantes en Ciudad Universitaria, por más que exista diversidad de estudiantado, es reflejo del habla de la ciudad de México, es una conclusión desmesurada.

Definir la representatividad de un corpus no es una tarea sencilla, será necesario tener una cierta organización o estructura jerárquica de la población estudiada, de manera que sea posible identificar los puntos de interés. Esto involucra la ponderación de diversos factores, generalmente asociados a la delimitación de nuestro objeto de estudio y de los alcances del proyecto. Por ejemplo, en un estudio concreto de la anotación de las grafías del sistema de sibilantes castellanas presentes en documentos coloniales del siglo XVI, en donde el objetivo es establecer normas ortográficas de acuerdo con el origen dialectal del escribano, será pertinente restringir la población de acuerdo con parámetros como la zona geográfica que delimita los dialectos de los escribanos, las etapas que comprende el periodo de la Colonia y el tipo de documentos a buscar.

Variedad

La división del corpus con base en los estratos de la población en marcas geográficas, culturales, dialectales, étnicas, temporales, históricas, o cualesquiera que sean para los fines concretos de estudio, serán de mucho provecho para organizar el corpus, pero también servirán como criterios de clasificación y búsqueda de la información. Los criterios de organización y estratificación de los corpus dependen de cada proyecto, su utilidad rebasa las fronteras de los estudios específicos para los cuales fueron creados y es posible la apertura a otros estudios. La siguiente lista servirá para dar una idea de la diversidad de criterios utilizados en diferentes aplicaciones.

Localidad geográfica. La localidad geográfica de procedencia de los textos o de los informantes, puede dividirse por países (sin dejar de tomar en cuenta que, por ejemplo, en Estados Unidos se encuentra una población muy alta de mexicanos), ciudades (Ayutla, Tlahuitontepec, Totontepec; Ciudad de México, Managua, Caracas) o regiones (chinantecos, mixes, otomíes; Sierra Gorda, Altiplano, Península Ibérica). En este rubro hay que evitar la interferencia geográfica que puede condu-

cir a datos erróneos, en donde, por ejemplo, una persona nacida en Buenos Aires radicada por largo tiempo en México muy posiblemente siga teniendo un acento rioplatense.

Información personal. Los datos personales del informante pueden ser decisivos para algunos estudios léxicos y pragmáticos, por lo que habrá que diferenciar, entre otros datos, el género,la edad o grupo etario, el estrato sociocultural, el nivel de estudios y la preferencia sexual. En este último caso es anecdótico señalar de primera fuente que un traductor germano que ya había proporcionado un par de textos para un corpus, se retractó y negó todos los derechos de usarlos cuando se le entregó el cuestionario donde se preguntaba la preferencia sexual. Si bien este es un caso aislado, de cualquier manera hay que advertir que la información proporcionada será confidencial y para fines estadísticos.

Tópico. El tópico, entendido como el tema o asunto del que se habla, es un criterio subjetivo que siempre será cuestionable por aquellos que ven el trabajo desde fuera. Sin embargo, es necesario tomar en cuenta que los criterios con los que se haya definido el tópico obedecen a los objetivos de un estudio en particular. La tipificación conviene ser apoyada por expertos en la materia, quienes no sólo dominan la tipología del área, sino que además conocen las fuentes más representativas. La división puede ser muy general (ciencias, sociales, humanidades) o bien llegar a nivel de detalle (transportes terrestres, estructuras de concreto, instrumentación sísmica), según se requiera.

Tipo de texto. El tipo de texto, ya sea oral o escrito, es una característica que marca diferencias léxicas, semánticas y discursivas. En textos escritos se podrá hacer una división entre ficción y no ficción (texto científico, biografías, ensayos, periódicos), en tanto en otros estudios interesará llegar a niveles específicos de precisión y diferenciar entre una carta formal, un memorando o una nota manuscrita. En texto oral convendrá distinguir entre habla espontánea y la que no lo es (discursos, lecturas, habla de laboratorio).

Fuente. La fuente explícita del texto es un dato que siempre debe existir y proporcionarse de manera obligatoria en un corpus. Desde el punto de vista de la organización del material será relevante el tipo de fuente (libro, revista, Internet, manuscrito) o la forma de obtención de un “texto oral” (programas de radio y televisión, entrevistas, grabaciones telefónicas).

Tiempo. Tanto para estudios diacrónicos como para estudios sincrónicos es indispensable ubicar los textos en el tiempo. Es necesario registrar el año (normalmente referido a la fecha de publicación o de grabación), o bien el periodo (década, 1925-1945) o la época del texto (Siglo de Oro, Generación del 27, Romanticismo).

Otros. Existen otros diversos criterios utilizados en la categorización de los textos que conforman un corpus, pero dependen más de los objetivos para los que van a ser empleados. Por ejemplo, puede ser de utilidad diferenciar entre el autor, el editor o el traductor de la obra, o bien especificar la nacionalidad, el país de nacimiento y el de residencia.

Equilibrio

La representatividad del corpus requiere variedad de información, aunque en un corpus multipropósito en el que se pretende abarcar un estado general de lengua con el fin de servir a diferentes tipos de investigación es todavía insuficiente. Hay otro factor importante para tener una mejor representatividad y que puede generar mejores resultados, la cual consiste en seleccionar los textos para obtener una muestra equilibrada. El equilibrio significa que el material contenido para cada uno de los rubros sea relativamente proporcional, evitando ser tendencioso a una parte únicamente. Sin embargo, hay que considerar que la adquisición del material muchas veces está supeditado a la disponibilidad o facilidad de acceso al mismo, de manera que es comprensible que el Corpus de la Real Academia de la Lengua Española, al ser realizado en España, esté conformado en un cincuenta por ciento de textos provenientes de dicho país; o en el caso de un corpus especializado, por ejemplo, es esperable que contenga más textos de un área que de otra. Equilibrar un corpus cuando se tienen varios rubros de organización (países, años, tipos de texto y tópicos) requiere hacer un balance de todos y cada uno de los rubros, lo cual se convierte en una tarea muy complicada. Cuando se tiene el apoyo de una base de datos para el registro del material, en donde se diferencian diversos criterios, entonces es posible obtener tablas y gráficas en las cuales pueda verse la distribución de los datos según los criterios y con ello mantener un mejor equilibrio en el corpus, solicitando los textos faltantes.

Selectividad

Los textos de un corpus deben filtrarse de acuerdo con los propósitos del estudio que se quiera realizar, ya que no es posible recopilar todo lo escrito y/o hablado de una lengua. Los criterios para la selección de los textos serán siempre controversiales y cuestionables. Una recopilación de textos literarios buscará tener sin duda a escritores renombrados de la talla de García Márquez, Borges y Paz, en el caso del español, esto sin perder otros exitosos y populares como Corín Tellado, así como una selección variada de autores menores. Las preguntas irán en función de cuántos, cuáles, y qué tanto de cada uno de los textos son pertinentes para la muestra, con el riesgo de encontrar inconformidades respecto a los criterios de selección. Existen algunas propuestas sobre las medidas que deben usarse a fin de tener mayor representatividad de la lengua en la selección de

los textos y su variabilidad, pero todas ellas han respondido a las necesidades específicas del proyecto y su presupuesto. Un criterio común que se ha utilizado para corpus de referencia es la selección aleatoria de muestras, pero también se ha discutido si las características lingüísticas se encuentran distribuidas uniformemente en los textos. Ante tal diversidad de criterios, la mejor opción es hacer una reflexión profunda de lo que se pretende obtener, una delimitación de la población que se va estudiar, una estimación de costos y tiempos, un sondeo de las posibilidades de obtener el material y el apoyo de un heterogéneo equipo de trabajo familiarizado con el área, tanto de lingüística como de computación y estadística.

Tamaño finito

En la literatura es común referirse al tamaño del corpus, esto es, al número de palabras o registros que contiene, considerando palabra a la secuencia de caracteres separados por espacios vacíos, sean éstos números, letras, símbolos o combinación de los anteriores. Dicha aclaración es importante puesto que el término “palabra” es muy controversial y discutido dentro del campo de la lingüística en general, por lo que es conveniente precisar lo que se entiende por “palabra” dentro de la perspectiva de la lingüística de corpus. Existen proyectos más ambiciosos que otros, hay corpus que tienen como objetivo proporcionar una amplia representación de la lengua y normalmente están subvencionados por grandes instituciones. En algunos casos se habla de cientos de millones de palabras, como el de la Real Academia de la Lengua con más de 400 millones, o bien el del Collins Cobuild que tiene más de 520 millones, mientras existen otros corpus de menor tamaño que se conforman con alrededor de 100 millones cada uno, como sucede con el del español de Mark Davis o los corpus nacionales de Gran Bretaña o de Estados Unidos. Ante estas cantidades, erróneamente se llega a considerar que un corpus es rico cuanto más palabras contiene, destacando de esta manera los aspectos cuantitativos a los cualitativos, cuando en realidad la riqueza de un buen corpus no reside tanto en su tamaño cuanto en la variedad y representatividad del mismo. De hecho, un corpus pequeño puede estar mejor diseñado y proporcionar mejores resultados que un gran corpus que considera textos enteros y poco representativos. Como ejemplo, el corpus que se recopiló en El Colegio de México para obtener un conocimiento riguroso del uso del vocabulario en el que se basara la redacción del Diccionario del español usual en México, tan solo consta de dos millones de registros, pero es lo suficientemente rico al contener una extensa y variada recopilación de muestras, mil textos de dos mil palabras gráficas cada uno, rigurosamente seleccionadas de todo tipo de textos hablados y escritos, provenientes de todas las regiones del país, de toda clase de hablantes y de una amplia variedad de géneros.

En este sentido, resulta absurdo tratar de construir un corpus exhaustivo que cubra todos los aspectos de la lengua de manera que se consiga, por ejemplo, todo el léxico,

pues para asegurarlo se requeriría considerar todo texto, oral y escrito, que sea producido por cada individuo, labor que llevaría incontables vidas. Por el contrario, el lingüista busca la representatividad en los textos, esto es, intenta recoger una muestra de la lengua que se estudia para seleccionar los ejemplos cercanos a la realidad lingüística con el fin de analizarlos de manera pertinente. El lingüista tiene que ser selectivo con el material que escoge, tiene que ser muy cuidadosos durante el diseño para que, de acuerdo con el presupuesto y los alcances, se pueda obtener el material esperado. Si bien en un corpus más grande se podrán encontrar ocurrencias que no se localizan en corpus pequeños, vale la pena meditar si estas ocurrencias son representativas o son simplemente datos aleatorios que escapan de la norma lingüística y no son pertinentes para el trabajo de investigación.

Ahora bien, para dar una idea de lo que significa el tamaño del corpus y evitar dar una impresión falsa de los números, vale la pena advertir lo que significa un millón de palabras. Como se dice en la introducción del British National Corpus, si se quisiera leer todo el corpus de cerca de 100 millones de palabras, a razón constante de 150 palabras por minuto durante 8 horas al día, se podría terminar su lectura después de 4 años ininterrumpidos. Esta cantidad resulta impresionante y uno puede imaginarse el trabajo que costará obtener los textos. Sin embargo, si se cuentan las palabras que hay en un escrito, sea un libro, el artículo de una revista, una tesis de doctorado o el ejemplar de un periódico, se puede observar que en sólo unos cuantos textos llegamos a tener el millón de registros. Para el número 38 de la revista Estudios de Lingüística Aplicada se tiene un total de 52,605 palabras en las 150 páginas de contenido, y si consideramos esta cantidad como un promedio general de cada número, entonces se tiene un total de prácticamente dos millones de palabras en los primeros 38 números de la revista.

Cabe advertir, sin embargo, que el millón de palabras como unidad de medida no es la única que existe, ya que ésta es más utilizada cuando se tienen textos grandes. Para corpus de referencia resulta más relevante mencionar el número de fragmentos de textos (llámense muestras, registros, textos, oraciones, cláusulas, etc.) y la longitud media de los mismos (normalmente el número de palabras, ya sea el promedio o el rango). En corpus orales se da mucho peso al tiempo, por lo que es común referirse al número de horas de grabación o al número de grabaciones por el tiempo aproximado de duración de cada una.

Con todo, se define que una característica de los corpus es que su tamaño debe ser finito y aunque hablemos de pocas o muchas palabras, se trata de una muestra de lengua delimitada. La ventaja del tamaño finito es que los corpus no son estáticos, pueden agregarse nuevos textos para obtener una muestra mayor del tema que se trata. Decidir el tamaño del corpus es una cuestión difícil que involucra varios aspectos. El costo y el tiempo son dos factores relacionados para tomar una decisión, pues entre mayores sean éstos, más grande podrá ser el corpus, ya que se requiere de horas/hombre para realizar

todas las tareas de procesamiento del mismo, tareas que van desde la solicitud del material y su recopilación, hasta la digitalización, limpieza y anotación. Asimismo, hay que prever si se puede disponer del material que va a constituir el corpus, porque muchas veces es difícil conseguirlo y los proyectos se quedan truncados. A fin de conseguir variedad de textos, es necesario recurrir a distintas fuentes de información, lo cual dificulta aún más la tarea.

1.4. Corpus informatizado

La tendencia actual en la lingüística de corpus se orienta a informatizar los textos, ya que al estar en soporte electrónico pueden ser recuperados y manejados electrónicamente. Incluso, varios corpus conocidos que en su momento fueron desarrollados en forma impresa, actualmente se han pasado a un formato electrónico. Hoy en día, los estudios de lingüística de corpus ya dan por hecho que se habla de textos informatizados.

El tener textos en un soporte electrónico tampoco es suficiente para construir un corpus: se pueden convertir todos los catálogos de una biblioteca en lo que se llama biblioteca digital, pero no por ello constituir un corpus. Aunque puede ser un recurso sumamente productivo para búsquedas de información lingüística o incluso para el tratamiento de datos numéricos, no debe considerarse como un corpus lingüístico hasta no pasar por un filtro metodológico pertinente. Conviene recordar que en todo corpus se debe tener una clasificación de las fichas o de los textos, de manera que sea recuperable únicamente la información que interesa. En este sentido, una biblioteca digital tendrá una buena clasificación de los documentos, pues es trabajo que compete a la bibliotecología. Sin embargo, una vez que se obtengan los documentos recuperados, el lingüista se interesará por analizar los textos, hacer anotaciones, obtener concordancias, hacer conteos estadísticos y otras labores para las que no es suficiente acceder y manipular las imágenes del documento; en este momento es cuando se alejan las bibliotecas digitales del sentido de corpus.

Asimismo, tampoco hay que dar por sentado que la WEB o Telaraña Mundial constituye un corpus porque en él podemos buscar patrones lingüísticos determinados y obtener con ello los diferentes usos léxicos de una palabra, o bien porque proporciona material suficiente para encontrar las flexiones y formas derivativas del español. Los archivos digitales de la correspondencia de una empresa, una base de datos terminológica bien estructurada y con opción a múltiples búsquedas, o la transcripción de los segmentos de noticias en radio y televisión, todas estas colecciones electrónicas tampoco constituyen un corpus lingüístico como tal, aunque bien son el sustento para integrarlo. Es necesario dar a la colección de textos un tratamiento informático que permita la recuperación y análisis de los documentos con fines lingüísticos, pero sobre todo con base en ciertas

normas y estándares.

Cuando los textos están electrónicamente respaldados puede ser de dos maneras diferentes: uno como imágenes (como si fueran fotos) y otro como textos (por transcripciones o por un proceso de digitalización y reconocimiento óptico de caracteres). Para ver las diferencias entre ambos puede tomarse el caso de un usuario que va a una biblioteca tecnologizada y quiere consultar los periódicos de ciertas fechas o incluso un manuscrito antiguo. Será normal entonces que el bibliotecario, en lugar de traerle los documentos, lo lleve a una sala de cómputo en donde pueda verlos y pasar entre sus páginas a la velocidad que quiera y, en algunos casos, al mismo tiempo que otros lectores podrán estar consultando los mismos. Podrá ver las imágenes insertas en el periódico, los diferentes tipos de letra, los mismos encabezados con su formato correspondiente, pero todo esto no son más que imágenes, una fotografía tal cual en donde las letras no son más que la unión de puntos. Por el contrario, cuando un documento se encuentra en lo que se conoce como formato texto, las letras dejan de ser imágenes para convertirse en letras que a su vez forman palabras, de manera que se pueden manipular los textos de otro modo o buscar las ocurrencias de una palabra e, incluso, de todas las palabras que contienen una letra o una secuencia de letras. Con archivos en formato texto podremos hacer cambios y reemplazar una cadena de letras por otra, cambiar los tipos y formatos de letra, añadir texto o insertar anotaciones; asimismo, podremos traer una lista de las ocurrencias de una palabra junto con los caracteres a su alrededor. Todo esto sería imposible de llevar a cabo en archivos con formato imagen.

En resumidas cuentas, un corpus tiene que estar en soporte electrónico, debidamente clasificado para la recuperación de la información que se necesita y debe ser manejable para poder hacer análisis lingüístico.

Ventajas

Las ventajas de tener el corpus en formato electrónico sobre la forma impresa en fichas de papel son varias, a saber:

La información se puede manipular de manera más sencilla. Debido a que los datos contenidos en un corpus son accesibles dada su naturaleza digital, se puede manipular más fácilmente. Es posible guardar la información, duplicarla y respaldarla; cambiar los nombres y la ubicación de los archivos; transferir, eliminar y añadir datos; realizar las tareas cotidianas de cortar, copiar y pegar, tanto dentro de un archivo como entre archivos; ordenar, clasificar, reubicar textos; en fin, múltiples tareas para las que se requiere conocimiento básico del uso de los programas de cómputo, pues hoy en día existen varias herramientas y programas que facilitan estas tareas.

La velocidad de procesamiento es mayor. La velocidad de procesamiento y de recuperación de información es insuperable mediante la computadora. No tiene punto de comparación con la manera manual y tradicional de los lingüistas que implica tener una muy buena organización y clasificación de las fichas en papel. Generalmente el resultado de trabajar manualmente es que las fichas se vuelven muy personales y difícilmente pueden ser empleadas por otros investigadores.

La precisión de los resultados es exacta. Hay algunas tareas sistemáticas que difícilmente podrían hacerse sin apoyo de los programas de cómputo, sin contar que de manera manual son más frecuentes los errores y por tanto los resultados pueden ser inexactos. Tan solo considérese el conteo de la letra F en un pequeño texto, ejemplo mencionado al principio de este capítulo. No sólo el conteo de letras o de palabras es más productivo en soporte electrónico, su ordenamiento, agrupación, clasificación, combinación y recuperación en el contexto de aparición se vuelve labor más sencilla en un soporte electrónico manipulable por la computadora. Las tareas son múltiples y sólo algunas se describirán más adelante.

La actualización del corpus se posibilita. Al igual que las fichas en papel, se puede actualizar la información. Sin embargo, con el formato electrónico se pueden realizar cambios en todos los registros a la vez, en tanto en papel hay que hacerlo ficha por ficha. Con unos cuantos comandos se puede pedir a la computadora que reemplace una palabra por otra o que incluya una etiqueta a cierta palabra, independientemente del número de veces que ocurra dicha palabra.

El costo de acceso disminuye. La mayoría de los corpus textuales son gratis o de bajo costo, además siempre existe la posibilidad de obtener textos a través de Internet, como se verá más adelante en este capítulo. Sin embargo, cabe hacer notar que los corpus orales pueden llegar a ser muy costosos (por ejemplo, un académico actualmente tendría que pagar arriba de 150 mil euros para obtener registros telefónicos grabados), lo cual se convierte en una desventaja desde el punto de vista del investigador que quiere adquirir los textos, pero como una ventaja si se piensa en la lingüística aplicada y orientada a la comercialización de sus productos.

La posibilidad de compartir la información. Se tienen mecanismos para facilitar el acceso y compartir el material. La información puede estar disponible vía Internet, de manera que puede ser consultada por medios electrónicos desde cualquier lugar del mundo por múltiples usuarios a la vez, aunque evidentemente existen algunas restricciones de uso para garantizar los derechos de autor y evitar el empleo malintencionado del material.

La transferencia de datos. Es posible y más fácil transportar grandes cantidades de texto para leerse en prácticamente cualquier computadora, gracias a las capacidades de almacenamiento que se puede tener en un disco compacto (CD) o en

una memoria portátil USB, a la vez que también existen protocolos para transferir información a través de Internet.

Desventajas

Sin embargo, conviene ser justos y tomar en cuenta que a pesar de todas las bondades que representa contar con corpus en formato electrónico, también existen algunas desventajas:

Se requiere de programas especializados. Si bien resulta relativamente sencillo tener los textos en soporte electrónico, el objetivo final consiste en el análisis de los mismos. Existen varios programas especializados para el manejo de textos, algunos disponibles en forma gratuita y los más comerciales a precios muy accesibles. El uso de estos programas implica sujetarse a las limitaciones que cada uno de ellos impone. Por ello, otra alternativa sería crear programas a la medida para el manejo de los textos a conveniencia de cada proyecto. Esto implicaría, por supuesto, tiempo y dinero para el equipo de trabajo que haga el diseño y realice los programas adecuados a los requerimientos particulares.

Es necesario digitalizar los textos. En algunos casos se requiere digitalizar los textos, esto es, pasarlos al formato electrónico. Ya sea que se transcriban los textos manualmente o se utilice un digitalizador y un programa de reconocimiento óptico de caracteres, de cualquier forma se está sujeto a errores. Pasar los textos a formato digital es un proceso largo y caro, pues requiere una mayor inversión en horas/hombre de lo que sería el diseño y la elaboración de un programa para manejar los textos.

El equipo de cómputo tiene requerimientos particulares. Es necesario adquirir el equipo de cómputo adecuado. Conviene tener una planeación adecuada para comprar el equipo conforme a los requerimientos del proyecto. Es indispensable considerar el tipo y las dimensiones del corpus en cuestión (sea oral o escrito) y los análisis que se realizarán a partir de éste. Para un corpus pequeño de un solo investigador puede ser suficiente una computadora personal, pero para proyectos más complejos habrá que pensar en un servidor con estaciones de trabajo, impresoras, digitalizadores de alta resolución y equipo de respaldo.

Se precisa una inversión económica. Aunque es evidente, pero no por ello resulta obvio, se requiere de la computadora en todo momento. Esto implica que el lingüista debe de alguna manera enfrentarse directa o indirectamente con los equipos de cómputo, por lo menos para referir a los expertos sus necesidades y requerimientos. Si bien puede sólo ver los resultados en forma impresa y nunca acercarse a una computadora, al menos debe tener la perspicacia para ver lo que se puede obtener con un corpus lingüístico.

El equipo empleado se debe actualizar. Al cabo del tiempo y con el avance de la tecnología habrá que ir actualizando todo, desde el equipo de cómputo y los mismos programas, con relativa frecuencia, hasta los formatos, códigos y etiquetas de nuestro corpus, más ocasionalmente. El equipo de cómputo se vuelve obsoleto con una velocidad inusitada, a tal grado que de la fecha en que se hace un plan de compras hasta que se obtiene el presupuesto y se compra el equipo, ya hay en el mercado equipo más sofisticado y de menor precio. Las compañías continuamente desarrollan mejorías en sus programas y ofrecen versiones con aditamentos novedosos y prácticos en tanto surgen nuevos programas a precios más competitivos. Con los avances científicos también se irá viendo la posibilidad de incorporar códigos y etiquetas especiales a los textos para ser reconocidos por los nuevos programas, lo que puede implicar cambios internos en los corpus.

Siempre existe la posibilidad de fallas técnicas. Con todo y que se busca que los sistemas sean más robustos, es indudable que la tecnología llega a fallar, razón por la que será necesario siempre tener un respaldo y estar conscientes de las limitaciones. Solo así se puede garantizar que el trabajo de varios años no desaparezca de un día para otro y sea irrecuperable. Siempre existe la posibilidad de fallas técnicas, con todo y que se busca que los sistemas sean más robustos, y el mejor recurso es estar preparados ante las posibles eventualidades. Por ejemplo, contar con respaldos para el resguardo de la información y para la infraestructura de cómputo. Solo así se puede garantizar que el trabajo de varios años no desaparezca de un día para otro y sea irrecuperable.

Mencionar que un corpus debe ser manejable por la computadora implica que se tengan archivos textuales, no como imágenes. Conviene advertir que existe una gran variedad de formatos de texto, los cuales son manejados por los editores y procesadores de texto. Los editores de texto permiten manejar los archivos que constan de texto plano, esto es, sin formato alguno y sin imágenes, salvo los saltos de línea, la tabulación horizontal y los retornos de carro. Por el contrario, los procesadores de texto pueden incluir diferentes estilos de letra, márgenes, imágenes, comentarios, hiperenlaces y muchos otros elementos. Los programas de procesamiento de texto utilizan su propio formato con el cual codifican los distintos elementos. Actualmente los programas procuran que sus formatos sean compatibles, de manera que un archivo con cierto formato pueda ser leído por otro procesador de texto. Sin embargo, a pesar de que los datos y el contenido se convierten en este proceso, se llegan a cambiar o perder algunas características.

Cuando construimos un corpus es común pensar en conservarlo y compartirlo. Así como las fichas en papel se hacen amarillentas a lo largo de los años y para garantizar su perdurabilidad se adquieren tarjetas de cierta calidad, también es necesario asegurar que los textos electrónicos sigan siendo consultados a pesar de las innovaciones tecnológicas y equipo más sofisticado, de los nuevos procesadores de palabras, programas de

cómputo y sistemas operativos. Si se quisiera que las fichas en papel tengan un almacén central, habría que uniformizar el tamaño y formato de las mismas; de igual manera, los textos informatizados deberán seguir un formato común que pueda ser accesible para cualquiera, aun para uno mismo que los consulta en diferentes máquinas.

1.5. Derechos de autor en corpus

Los derechos de autor son tan importantes como todas las consideraciones anteriormente mencionadas para la adecuada construcción de un corpus, sobre todo cuando se trata de textos que van a ser publicados y difundidos para que el público pueda acceder fácilmente a ellos. Debido a que un corpus recopila textos (orales o escritos) de testimonios personales o privados, o bien de publicaciones ya hechas, hay que considerar las reglas pertinentes para su utilización o divulgación, esto con el fin de proteger los derechos de propiedad intelectual que corresponde a los informantes o a los autores de los textos que conforman el corpus.

Se ha discutido en varios foros si en todos los casos hay que solicitar permiso para usar parte o la totalidad de un texto. Es bien sabido el concepto y las consecuencias del plagio, razón por la que se debe citar el extracto de una obra, normalmente dando referencia a la fuente. Sin embargo, ahora que se aplican con mayor rigor leyes internacionales para respetar los derechos de autor, de manera que se considera ilegal hacer fotocopias de libros y revistas, surgen dudas naturales acerca de la validez de usar un extracto de texto y mostrarlo como ejemplo a los alumnos en una clase impartida con una red de cómputo. De similar manera, uno llega a cuestionarse la posibilidad de extraer un fragmento que servirá únicamente para los fines de un trabajo académico, en donde sólo aparecerán los resultados de la investigación, pero nunca como tal los fragmentos extraídos.

El principio básico de la propiedad intelectual es respetar y proteger la obra de los autores, de manera que cualquier forma de uso, en este caso desde la recopilación del material hasta la distribución del mismo, si se va a hacer del dominio público la explotación de los textos, requiere del consentimiento del titular de los derechos de la obra. Es de particular relevancia para un corpus, en donde justo la materia prima la constituye la obra de los autores, que se cumpla con las reglas establecidas para el acceso a la información y con los derechos de autor para evitar el mal uso.

Aunque el tema es muy controvertido y podría llevar mucho tiempo discutirlo, conviene mencionar algunas normas que aplican en la recopilación de textos, orales o escritos, para la construcción de un corpus, sea cual sea su fin:

Tener el consentimiento del poseedor de los derechos. No se deben publicar los documentos que lo forman total ni parcialmente, por ningún medio, incluyendo el electrónico, sin consentimiento del propietario del derecho de autor. En algunos casos hay que identificar quién es el poseedor de los derechos: si el autor, el editor, el publicista o el traductor, sea una persona física o moral, razón por la cual es conveniente mantener registros en bases de datos como la ejemplificada en la figura 1.1. En esta figura se muestra un ejemplo de una de las fichas del Translational English Corpus (TEC), cuya mira central está orientada a la recopilación de textos traducidos al inglés y, por tanto, involucra al menos tres datos necesarios: el idioma del texto original, el autor y el traductor. De estos dos últimos se registran datos que no se muestran en esta ficha, pero aquí se puede observar otra información usada para cuestiones administrativas, tales como el publicista, el poseedor de los derechos de autor y las fechas en que se solicitó y se obtuvo permiso para utilizar el material. La labor puede ser titánica y requerir un encargado de tiempo completo para dedicarse a cuestiones administrativas y de relaciones públicas.

Obtener cartas de autorización. Las cartas deberán ser para el uso y publicación de los textos del corpus, pues de ninguna manera debe dejarse a convenios verbales, pues no son válidos y son revocables. Las normas indican que la autorización para usar la obra debe ser previa, explícita y por escrito por el titular de los derechos. Hacerlo previamente puede ahorrar mucho trabajo, sobre todo cuando hablamos de libros que hay que digitalizar, limpiar y procesar para adecuarlos a los formatos del corpus.

Explicitar el uso. El acuerdo explícito asentado en la autorización será el tipo de uso que se le va a dar a la obra. Por ejemplo, si se va a hacer uso de los textos para una investigación particular, con acceso restringido, o si se va a permitir el acceso de los textos a todo el público. Asimismo, si sólo se permitirá el acceso a parte de los documentos y a su totalidad. O bien, si el acceso será en línea con lo que permita la interfaz de consulta o se podrán descargar los textos.

Citar las fuentes. Independiente a los permisos para explotar una obra, deben darse referencias explícitas a las fuentes de cada uno de los textos, señalando los datos bibliográficos pertinentes, incluso para los textos obtenidos de Internet que, aunque sean del dominio público, también deben ser referenciados y, en los casos que corresponda, solicitarse derechos para el uso de los textos. Según el tipo de resultados que se pretenda desplegar, a nivel de concordancias se indicará la fuente del fragmento, en tanto para cuestiones generales, como el conteo de palabras, se puede tener un apartado con el listado de todas las fuentes del corpus.

Existen excepciones o límites a las normas para las que no se requiere solicitar autorización del uso de la obra. Una de ellas es pertinente cuando se trata de trabajos

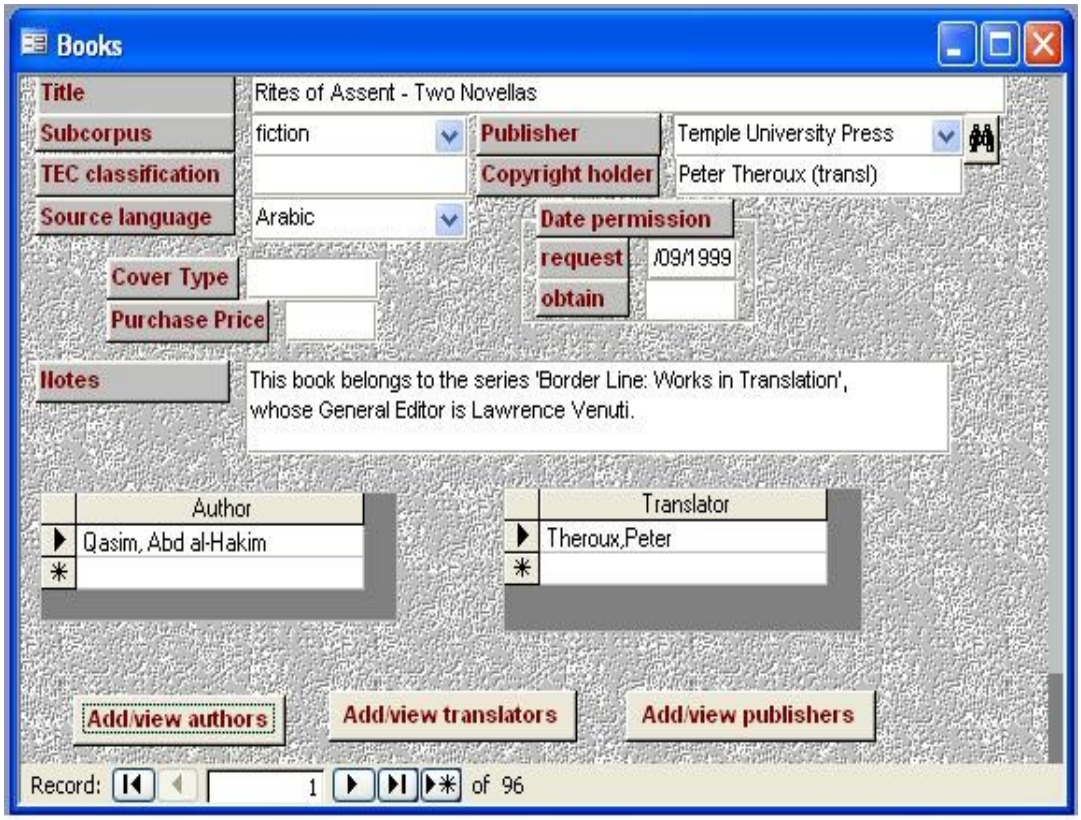

Figura 1.1: Ejemplo de la base de datos del Translational English Corpus.

de investigación o docencia sin fines de lucro, siempre y cuando no se permita el acceso más allá de fragmentos de textos y se indique claramente la procedencia de los mismos. Ante este rubro de uso honrado se han manejado varios corpus, sobretodo cuando se trata de corpus de referencia en donde se recopilan partes pero no la totalidad de múltiples obras. Otra excepción muy útil para corpus históricos es que la posesión de los derechos se conserva hasta 70 años después del fallecimiento del autor.

Para cualquiera que sea el fin de usar corpus, se busca tener datos reales. Se intenta que los hablantes, en el caso de corpus de procedencia oral, no modifiquen su estilo de habla y que, por tanto, no estén conscientes de que son observados, con el fin de que los datos obtenidos del corpus se aproximen a la lengua real y sean una fuente que proporcione información más acertada acerca del uso lingüístico. Por eso, en corpus orales, es preferible solicitar permiso al hablante una vez que ha terminado la entrevista

o grabación, a fin de no perder espontaneidad en los actos del habla. Lo mismo es aplicable en textos escritos como correos electrónicos y pláticas electrónicas o chats, por ejemplo. Cuando se trata grabaciones personales de esta naturaleza, en muchas ocasiones el autor busca proteger su privacidad, por lo que podrá solicitar que no se proporcione su nombre; además, para los fines lingüísticos lo importante será anotar los datos del informante: edad, sexo, escolaridad, etc.

Ahora bien, en cuanto al tema de derechos de autor cabe señalar que una vez construido un corpus también éste está protegido como propiedad intelectual, por lo que es importante reconocer el trabajo del equipo (diseñadores, programadores, capturistas, digitalizadores, etc.) y agradecer al patrocinador. A fin de justificar con el patrocinador el monto del dinero invertido se requiere obtener datos sobre los tipos de usuarios y tener estadísticas de uso. Se puede solicitar un registro a los usuarios y pedir que confirmen estar de acuerdo con los derechos de autor y establecer cláusulas a conveniencia; por ejemplo, si se quiere utilizar un fragmento de texto obtenido en una consulta, se debe dar la referencia al corpus, más que a la fuente misma correspondiente al texto.

1.6. Referencias

Lecturas sugeridas

Biber, Douglas (1993). ”Representativeness in Corpus Design“. Literary and Linguistic Computing 8 (4), pp. 243-257.

McEnery, Tony y Andrew Wilson (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press. (Véanse capítulo 2 y sección 3.3).

MacMullen, John (2003). Requirements Definition and Design Criteria for Test Corpora in Information Science. SILS Technical Report 2003-03, School of Information and Library Science, University of North Carolina at Chapel Hill. (Véanse secciones 4.1 y 4.2).

Torruela, Joan y Joaquim Llisterri (1999). ”Diseño de corpus textuales y orales“. En J.M. Blecua et al. (Eds.), Filología e informática: Nuevas tecnologías en los estudios filológicos. Barcelona: Editorial Milenio-Universidad Autónoma de Barcelona, pp.45-77. (Véase sección 2).

Villayandre Llamazares, Milka (2008). Lingüística con corpus (I). Estudios Humanísticos. Filología 30, pp. 329-349.