Creación de *corpus*

JOHN M. SINCLAIR*

Trad.: E. Lavín

Describo a continuación, en términos generales, las consideraciones más notables para la creación de un *corpus* de textos.

Un *corpus* general

La primera consideración es el objeto del trabajo, que debe ser o lo bastante general para proporcionar una buena selección de ejemplos de la lengua, para fines muy variados, que no es necesario enumerar, o bien debe ser específico, como una fase en el desarrollo de un producto, o en la persecución de un fin en investigación. En este artículo me voy a ocupar de la elaboración de *corpora* generales ya que es más difícil conseguir ayuda para su creación que para los específicos.

Ahora bien, el valor de un *corpus* general como lugar de referencia es muy grande, y es probable que crezca enormemente en los años venideros. ¿Quién puede advertir el valor de un diccionario antes de tenerlo a su disposición, de manejarlo, de comprenderlo, de usarlo? Todos somos conscientes ahora de que la documentación de una lengua en diccionarios, gramáticas, etc., es un paso esencial en su maduración. El *corpus*, como muestra de la lengua viva, asequible a través de ordenadores complejos, está abriendo caminos insospechados.

* Universidad de Birmingham.

Líneas generales en la creación de un *corpus*

La creación de un *corpus* es algo simple como proyecto. Hay que decidir un tamaño conveniente y las prioridades en la selección; después, mediante consulta de catálogos, etc., se recogen los textos que serán incluidos. Trataré de exponer a continuación cada uno de estos pasos, aunque haya dos cuestiones estrictamente prácticas que los oscurecen.

Forma electrónica

La primera es que si un ordenador maneja un *corpus* debe tener el material en forma electrónica, ya sea a partir de material impreso o conseguido directamente a través de procesamiento de textos mediante ordenadores (impresión, tratamiento de textos, correo electrónico, etc.). Hay tres métodos normales de entrada de textos en el momento actual.

a) Adaptación del material que ya esté en forma electrónica.
b) Conversión mediante reconocimiento óptico (lectura automatizada).
c) Conversión por medio del teclado.

La mayor parte de los proyectos harán uso de todos estos métodos, puesto que cada uno es apropiado para una clase diferente de material. Por ejemplo, el material escrito a mano tendrá que ser tecleado, y los textos de periódicos solamente se pueden incluir de forma económica si están disponibles en forma electrónica. El escáner es la mejor alternativa para la masa de libros que se imprimen siguiendo métodos convencionales. Ya que los escáners universales (capaces de leer cualquier carácter) son escasos y caros, y es muy importante que los directores de proyectos estimulen su empleo. Tendría que ser abandonada la práctica de los años sesenta y setenta, cuando se crearon los primeros *corpora* de forma cuidadosa mecanografiando dos veces el material impreso.

Autorizaciones

El otro problema de índole práctica es conseguir permisos para que el texto pueda ser puesto en forma electrónica y para que se puedan hacer citas del texto en artículos, informes y otras publicaciones. Esta parcela constituye un área sensible de la ley y, aunque la experiencia demuestra que existe casi siempre buena voluntad por parte de los editores, es importante el esfuerzo necesario para conservar un *corpus* extenso en buena salud legal. Es probable que si los responsables de los *copyrights* comprendieran con claridad las razones que hay para solicitar sus textos, y que existe protección contra la explotación y piratería, se podría evitar ese gran esfuerzo tan poco productivo. La ayuda del Consejo de Europa como autoridad cultural sería inestimable. Hasta que este problema no se resuelva internacionalmente, es probable que la situación empeore. Se puede aconsejar en algunos casos, pero los proyectos de *corpora* tendrían que diseñarse pensando en este problema como amenaza latente que se cierne sobre la empresa.

Diseño

Habiendo comprendido claramente esos problemas de índole práctica, podríamos centrarnos ahora en los criterios para la selección de textos. En lo esencial, debe hacerse con sentido común, pero existen unas cuantas reflexiones, producto de la experiencia, que valdría la pena considerar.

Lengua hablada y escrita

La decisión de mayor alcance quizá sea la de si el *corpus* tendrá textos escritos, o solamente transcripciones de la lengua hablada, o ambas cosas. Muchos *corpora* se mantienen alejados de los problemas de la lengua hablada —con honrosas excepciones—, lo cual no es muy afortunado en un *corpus* que de alguna manera pretenda reflejar un «estado de la lengua». Según mi

experiencia, no hay alternativa para la lengua hablada improvisada, y la decisión que adopté en 1961 de reunir un *corpus* de conversaciones es una de las más felices que he tenido. Ya entonces me aseguraron que estaba a la vuelta de la esquina una transcripción automática del lenguaje hablado. Aún no existe.

Casi lengua hablada

Si es imposible, al comienzo de un proyecto, recoger la lengua hablada, es entonces una equivocación reunir guiones de películas, textos de teatro, etc., como si de alguna forma esto compensara tal anomalía. En un *corpus* general tienen un valor muy limitado porque se les «considera» lengua escrita para imitar la lengua hablada en medios artificiales.

La creación de un *corpus* hablado no es tan sencilla como la de uno escrito, y requeriría un artículo de por sí. Puesto que la actividad de la mayoría de los *corpora* se dedica a la lengua escrita, me centraré en este problema en el resto de este trabajo.

Lengua formal y literaria

El material contendrá desde el estilo formal al informal, desde el literario al estándar. El formal será más fácil de conseguir que el informal, y el literario será más fácilmente identificable que el estándar. En un *corpus* general, sus diseñadores deberían compensar este hecho. Se envían por correo, se echan por debajo de la puerta, se amontonan en tiendas, oficinas, salas de espera, se adjuntan en paquetes y se acumulan en las bandejas de recepción de correo en nuestras oficinas cada vez con mayor profusión mensajes efímeros, y normalmente informales. Socialmente no tienen importancia alguna, pero proliferan y son propios de la prosa mundana. De igual forma, no es necesario adornar el *corpus* con los mejores fragmentos de la prosa literaria contemporánea. Una pequeña proporción bastará. La razón es simple y en absoluto maliciosa o antiliteraria.

Características

El uso principal de un *corpus* de dimensiones normales es el de identificar lo que es central y propio de la lengua. Esto es más valioso como lugar de referencia comparado con el trabajo de los escritores literarios, que puede ser estudiado con seriedad. Si el trabajo de escritores reconocidos como tales fuese de hecho el que dominara en el *corpus*, tendría poco o ningún valor desde ese punto de vista; aún más, puesto que es característico de la literatura la innovación, la mayor parte de su patrón distintivo se perdería porque no aparecería lo suficiente para que pudiese contar como central y típico.

De igual forma, en el periodismo, los escritores muy conocidos tienden a tener un estilo poco corriente, y los estilos más corrientes y vulgares para hacer reportajes serán más útiles en el *corpus*.

Ello es una fuente de prejuicios e incomprensión. Si hemos de tener una visión realista del modo como se usa la lengua, deberíamos recoger el lenguaje de la masa de escritores normales, y no el del genio insólito o del periodista de garra.

Criterios de diseño

Hay otros muchos criterios que pueden aplicarse —demasiados, de hecho, porque cada criterio se suma sustancialmente al número de muestras diferentes que tienen que conseguirse—. Para un debate sobre los criterios más señalados se puede consultar en Renouf (1983), y en el mismo artículo se encuentra una exposición de cómo se decidió el proyecto.

Como guía general, para un *corpus* general aconsejo que cualquier material especializado o se elimine o se almacene por separado como *corpus* auxiliar. Un *corpus* de consulta general no es una colección de material de diferentes áreas de especialización: técnicas, dialectales, de tema juvenil, etc. Es una colección de material que es en gran medida similar, pero que se recoge de varias fuentes de forma que la individualidad de una fuente quede oscurecida a menos que el investigador aísle un texto determinado.

¿Quién debe diseñar un *corpus*?

Vale la pena que alguna de las características de las fuentes se controlen, en un sentido lógico. La especificación de un *corpus* —los tipos y proporciones del material en el mismo— no es en manera alguna un trabajo propio de lingüistas, sino que es más propio de la sociología de la cultura. La postura del lingüista debe ser la disposición para describir y analizar cualquier fragmento de lengua que caiga en sus manos. Durante la infancia de la disciplina lingüística de los *corpora*, los lingüistas tenían también que hacer la selección de textos; cuanto mayor es el trabajo acumulado más posibilidades hay de que vengan ofrecimientos de ayuda. Ciertamente, el volumen de cualquier debate profano sobre *corpora* puede afectar a los criterios de selección de textos.

Períodos

La mayor parte de los *corpora* intentan abarcar un período determinado de tiempo, y usan la fecha más evidente, que es la del primer fragmento hablado o la de la primera publicación. Sin embargo, si intentamos tomar muestras del material que produce la sociedad, es entonces cuando se hacen más importantes otros factores. Un trabajo escrito necesita tiempo para que se conozca y puede que su influencia dure algún tiempo.

Tamaño

Las dimensiones de un *corpus* constituyen la preocupación fundamental para casi todos los investigadores en el momento de la conceptualización inicial, y en sus declaraciones públicas. A la larga, poco importan. El único consejo que daría es que un *corpus*, cuanto mayor sea, mejor, y siempre tiene que estar creciendo. Mi advertencia se basa en el patrón de frecuencia de aparición de las palabras en los textos, ya señalado por Zipf (1931). Fundamentalmente hay un enorme desequilibrio en la frecuencia del léxico. Un texto está lleno de palabras del tipo de *of, is, up* y *by;* tiene menos

del tipo de *like, take, any* y *most;* y aún menos de *words*, y menos aún de *text*.

Aproximadamente la mitad del vocabulario de un texto —incluso de uno bastante extenso— lo componen palabras que sólo han aparecido una vez en el mismo.

Para estudiar el comportamiento de las palabras en los textos, hemos de tener a nuestra disposición un gran número de apariciones léxicas. De nuevo, tenemos en contra a la estadística, puesto que si clasificamos las apariciones léxicas en términos de «uso» o «significado» encontraremos otra vez el mismo tipo de desequilibrio. Uno de los usos será característicamente el doble de frecuente que los otros; muchos otros aparecerán sólo una vez, y ello no es suficiente para fundamentar una relación descriptiva.

Ésta es la razón por la que un *corpus* necesita tener millones de palabras.

Tamaño de la muestra

La otra decisión necesaria desde el principio es contar con el tamaño apropiado para las muestras. En este punto hay opiniones muy diversas. Algunos *corpora* —como los de Brown y Lob que iniciaron este tipo de estudio— optan por un tamaño uniforme de muestro (unas 2.000 palabras) (Holland & Johansson, 1986). Esto tiene ventajas para la comparación, y tendrá valor si el resto de la organización se diseña atendiendo a factores estadísticos. Sin embargo, las mayores divisiones del *corpus* están hechas por géneros, identificados con criterios mayoritariamente intuitivos. También un *corpus* que no presta atención al tamaño y forma de los documentos de los que se alimenta está en peligro de aparecer como una colección de fragmentos en donde sólo son accesibles modelos hechos a pequeña escala.

Documentos completos

La alternativa es la de coleccionar documentos completos. Así no habrá que preocuparse sobre las diferencias señaladas que se

hayan observado en las distintas partes de un texto. A lo largo de un texto del tamaño de un libro no son muchas las características que se presentan uniformemente, y un *corpus* basado en documentos completos está más abierto a un mayor número de estudios lingüísticos que una colección de muestras pequeñas. No hay que preocuparse, tampoco, de la validez de las técnicas de muestreo. Aún más, si por alguna razón se desean tener muestras dispersas de unas 2.000 palabras, se puede conseguir fácilmente de una colección extensa de textos completos. Ésta es la otra razón de la defensa de una política de crecimiento continuo: de un *corpus* extenso se pueden sacar *corpora* de menor tamaño y más especializados, según las necesidades que vayan surgiendo.

El precio que hay que pagar por la inclusión de documentos completos es que al principio de su recogida, la cobertura no será tan buena y las peculiaridades de un estilo o tema concreto puede que se dejen ver a través de las generalidades. Ninguno de estos problemas es fundamental o de gran duración.

Criterios mínimos

De aquí que el consejo sobre creación de *corpus* sea acordar el menor número de criterios posibles que puedan justificarse en cada situación, para que el número de documentos diferentes sea lo más pequeño posible. Ante todo, háganse anotaciones muy detalladas del material para que los documentos se puedan identificar desde campos diferentes a aquellos por los que fueron seleccionados como integrantes del *corpus*. Después, inténtese encontrar un documento apropiado para cada combinación de criterios. Anótese su tamaño, y divídase en muestras si se desea.

Corpus provisional

Utilizando este sistema, tendría que existir un pequeño *corpus* general utilizable que pudiera contener entre cinco y diez millones de palabras. Si se desea hacer un análisis provisional, el *corpus* se

puede arreglar con objeto de que los criterios se equilibren. Este arreglo implica la adición de textos extra donde existan huecos, y dejar a un lado partes de textos muy extensos que den prominencia a un determinado estilo. Este tipo de *corpus* será el adecuado para el estudio de estructuras y significados de frecuencia regular de miles de palabras, pero no para una descripción fidedigna de la lengua en su conjunto.

Duplicado

Se podría entonces doblar el tamaño, observando los mismos criterios, y así tener una más clara y detallada visión. Pero aún quedan miles y miles de cabos sueltos —palabras y significados poco frecuentes, frases poco usuales pero características, subtipos de estilo, etc.—. Se querrá duplicar de nuevo, y es lo que estoy haciendo por el momento en inglés moderno. Para el estudio de las «colocaciones» y de los giros es necesario analizar montones de textos para aislar las estructuras que se repiten y para reducir la importancia de las que son transitorias.

Corpus monitor

Finalmente, será posible crear un nuevo tipo de *corpus*, uno que no tenga límites de extensión, porque, como la misma lengua, esté siempre desarrollándose. Cada vez entrará más material por fuentes de lectura automatizada, y será examinado con la finalidad de hacer anotaciones rutinarias. Después será guardado en un medio de rápido acceso para que pueda llevarse a cabo cualquier clase de uinvestigación. Gradualmente, a medida que el material entra, será desplazado a un almacenamiento (memoria) secundario hasta que sea borrado o archivado como material histórico (véase Clear, en preparación).

No necesitamos añorar el texto; vivimos en un tiempo de proliferación de textos. Aquellos que tienen gran valor intrínseco no son la preocupación esencial de este trabajo, y serán preserva-

dos cuidadosamente por otras instituciones. Su estudio se incrementará al compararlos con la visión global que proporcionará el *corpus* general.

Características de un *corpus* monitor

De este modo, alguna vez el *corpus* tendrá una gran y actualizada selección a nuestra disposición; tendrá una dimensión histórica y un amplio vocabulario a causa de su elaborado registro de datos. Un *corpus* así es lo que necesita toda lengua de rango internacional.

Procesado

La estrategia para guardar, procesar y recuperar información de este *corpus* será definida en pocas palabras.

En primer lugar, sería una gran ayuda estar de común acuerdo en toda Europa sobre prácticas normalizadas en la representación de textos en un ordenador. Por ejemplo:

a) Que exista información «bibliográfica» completa, quizá tanto en forma electrónica como en papel.
b) El texto de lengua actual debe ser separado de todos los restantes códigos por medio de una convención normalizada.
c) Que el texto de lengua sea codificado en un formato ampliamente reconocido, o que se den detalles para que sea reconvertido fácilmente.
d) Cualquier otro código distinto del de un texto corriente debe ser identificado y clasificado; por ejemplo, códigos tipográficos, sistemas de referencia, mantenimiento y conservación, marcas analíticas de lengua. Se facilitarán claves completas para códigos que no sean generales, y en la descripción habrá que dejar claro qué criterios son automáticos y cuáles son los mediatizados por agentes humanos.

Política de «texto limpio»

La norma más segura es la de conservar el texto tal como está, sin procesar y limpio de otros códigos que puedan añadirse en investigaciones concretas. Hay dos razones principales para hacerlo. En primer lugar, es posible que cada trabajo de investigación vea la lengua según prioridades diferentes. Su cuerpo analítico puede que sea valioso e interesante para el investigador siguiente, e incluso que se adapte a las nuevas necesidades; pero de ninguna manera debe estar tan normalizado que pueda convertirse en parte integral del *corpus*.

En segundo lugar, aunque los lingüistas evaden sin esfuerzo alguno abstracciones tales como «la palabra» (en el sentido de lema) y de ahí para arriba, no lo hacen todos de la misma manera, y tampoco diseñan normas precisas para lo abstracto. De aquí que los principios fundamentales de la lingüística, como la identificación de palabras, delimitación de divisiones morfológicas y de partes fundamentales de la oración, no se encuentren aún normalizadas en modo alguno. Cada análisis ayuda a los demás, pero no crea una plataforma sobre la cual los otros puedan asentarse directamente.

Provisiones básicas

Este enfoque conservador no implica que el director del *corpus* prescinda de ayuda alguna. La primera parte del análisis de un texto requiere gran cantidad de procesamiento básico (Sinclair, 1985), y esto se puede remediar con la provisión de lematizadores, identificadores y analizadores que sean sencillos y eficaces, lo que permite un trabajo de investigación más profundo mientras se trabaja sobre una parte concreta. Si estas herramientas primarias están claramente documentadas, puede que faciliten a veces atajos en la tarea. Hasta donde sea posible, estas ayudas deberían conseguirse *on line*, mejor que hechas y guardadas.

dos cuidadosamente por otras instituciones. Su estudio se incrementará al compararlos con la visión global que proporcionará el *corpus* general.

Características de un *corpus* monitor

De este modo, alguna vez el *corpus* tendrá una gran y actualizada selección a nuestra disposición; tendrá una dimensión histórica y un amplio vocabulario a causa de su elaborado registro de datos. Un *corpus* así es lo que necesita toda lengua de rango internacional.

Procesado

La estrategia para guardar, procesar y recuperar información de este *corpus* será definida en pocas palabras.

En primer lugar, sería una gran ayuda estar de común acuerdo en toda Europa sobre prácticas normalizadas en la representación de textos en un ordenador. Por ejemplo:

a) Que exista información «bibliográfica» completa, quizá tanto en forma electrónica como en papel.
b) El texto de lengua actual debe ser separado de todos los restantes códigos por medio de una convención normalizada.
c) Que el texto de lengua sea codificado en un formato ampliamente reconocido, o que se den detalles para que sea reconvertido fácilmente.
d) Cualquier otro código distinto del de un texto corriente debe ser identificado y clasificado; por ejemplo, códigos tipográficos, sistemas de referencia, mantenimiento y conservación, marcas analíticas de lengua. Se facilitarán claves completas para códigos que no sean generales, y en la descripción habrá que dejar claro qué criterios son automáticos y cuáles son los mediatizados por agentes humanos.

Política de «texto limpio»

La norma más segura es la de conservar el texto tal como está, sin procesar y limpio de otros códigos que puedan añadirse en investigaciones concretas. Hay dos razones principales para hacerlo. En primer lugar, es posible que cada trabajo de investigación vea la lengua según prioridades diferentes. Su cuerpo analítico puede que sea valioso e interesante para el investigador siguiente, e incluso que se adapte a las nuevas necesidades; pero de ninguna manera debe estar tan normalizado que pueda convertirse en parte integral del *corpus*.

En segundo lugar, aunque los lingüistas evaden sin esfuerzo alguno abstracciones tales como «la palabra» (en el sentido de lema) y de ahí para arriba, no lo hacen todos de la misma manera, y tampoco diseñan normas precisas para lo abstracto. De aquí que los principios fundamentales de la lingüística, como la identificación de palabras, delimitación de divisiones morfológicas y de partes fundamentales de la oración, no se encuentren aún normalizadas en modo alguno. Cada análisis ayuda a los demás, pero no crea una plataforma sobre la cual los otros puedan asentarse directamente.

Provisiones básicas

Este enfoque conservador no implica que el director del *corpus* prescinda de ayuda alguna. La primera parte del análisis de un texto requiere gran cantidad de procesamiento básico (Sinclair, 1985), y esto se puede remediar con la provisión de lematizadores, identificadores y analizadores que sean sencillos y eficaces, lo que permite un trabajo de investigación más profundo mientras se trabaja sobre una parte concreta. Si estas herramientas primarias están claramente documentadas, puede que faciliten a veces atajos en la tarea. Hasta donde sea posible, estas ayudas deberían conseguirse *on line*, mejor que hechas y guardadas.

Base de datos

La primera base de datos derivada de un *corpus* extenso debe ser algo que no pueda realizarse de forma automática; debe tener algo de intervención por parte del hombre como investigador. Casi con seguridad tendrá que estar ordenada léxicamente, y casi con seguridad será una de las modernas bases de datos relacionales. Para tratar en detalle de la estructura de estas bases de datos tendríamos que hacer otro artículo, y es urgente ya en el contexto europeo que las bases de datos léxicas en varias lenguas se hagan y sean compatibles unas con otras.

Mantenimiento

Una vez que un *corpus* cobra vida, necesita mantenimiento y renovaciones periódicas. Habrá siempre errores que corregir y mejoras que hacer, al igual que adaptaciones a los nuevos hardware y software y a los cambios de las exigencias de los usuarios.

También habrá que estar continuamente atento a los sistemas de recuperación de datos, y a los instrumentos de análisis y procesado. El *corpus* es un enorme almacén de información y puede realizar mejoras dentro de sí mismo. Por ejemplo, las clasificaciones intuitivas en variedades estilísticas están completamente respaldadas por hechos; pero los hechos pueden transformarse en evidencia interna en cuanto variedad diafásica, y sobre esta evidencia se podría reclasificar el *corpus* continuamente.

Bibliografía

Clear, J.: «Trawling the Language: Monitor Corpora», en *Proceedings International Congress European Association for Lexicography (EURALEX)*, Zurich, 1986.

Holland, K., y Johansson, S.: *Word Frequencies in British and American English*, Londres, Longman, 1986.

Renouf, A.: «Corpus development at Birmingham University», en Aarts, J., y Meijs, W. (eds.): *Corpus Linguistics*, Amsterdam, Rodopi, 1984.

Sinclair, J.: «Basic Text Processing», en Leech, G., y Candlin, C. (eds.): *Computers in English Language Teaching and Research*, Londres, Logman, 1986.