LINGÜÍSTICA CON CORPUS (I)[1]

MILKA VILLAYANDRE LLAMAZARES
Universidad de León

Resumen

En este artículo hacemos una introducción a la metodología lingüística basada en el trabajo con corpus. En primer lugar, repasamos los diferentes conceptos de corpus en relación con las corrientes teóricas más influyentes en los últimos siglos. En un segundo punto fijamos las características que definen el concepto actual de corpus y que lo diferencian de otras recopilaciones de textos. Por último, revisamos los principales tipos de corpus y comentamos algunos de los ejemplos más representativos, con especial énfasis en los corpus para el español y otras lenguas peninsulares.

Palabras clave: corpus, lingüística de corpus, empirismo.

Abstract

In this paper we focus on the corpus approach to language studies through a brief review of the concept of corpus from the past century to the present day. We also select some criteria to try to differentiate it from other compilations of texts. Finally, we consider the main types of corpora, paying especial attention to the Spanish resources.

Key words: corpus, corpus linguistics, empiricism.

0. Introducción

Los corpus son hoy en día uno de los recursos básicos para el estudio y la descripción de las lenguas. En particular, en el ámbito de la Lingüística Computacional, se han erigido como punto de partida imprescindible para la elaboración de léxicos y gramáticas, y representan una línea de investigación transversal en lo que al tratamiento con medios informáticos del lenguaje se refiere, al ser indispensables para el desarrollo de

[1] Este artículo tiene su origen en material elaborado para la asignatura *Lingüística Computacional II*.

aplicaciones basadas tanto en el texto como en el habla (*cf.* Moure y Llisterri, 1996).

1. Los corpus como metodología lingüística

El uso de corpus para el estudio de la lengua se considera una metodología empírica de trabajo, basada en el empleo de datos reales, de muestras de uso de la lengua. El conjunto de datos es lo que se denomina 'corpus' en un sentido general del término[2]. Pero ha sido el empleo de ordenadores para reunir, organizar y procesar esos datos el que ha dotado de modernidad a esta tarea, hasta el punto de propiciar el despegue de toda una forma de hacer lingüística, la llamada 'lingüística de corpus'. Aunque el término en sí mismo es de acuñación relativamente reciente, los trabajos basados en corpus siempre han contado con seguidores.

A continuación nos detendremos brevemente en algunos de los hitos que han contribuido al proceso de desarrollo y consolidación de la lingüística de corpus[3]. Prestaremos especial atención a la evolución del concepto de 'corpus'.

1.1. Precedentes

Es posible establecer un primer concepto de corpus, previo al ordenador y avalado por toda una tradición de trabajos lingüísticos.

Así, con anterioridad al siglo XIX, un corpus se definía por:

a) Ser un conjunto de textos escritos (datos).
b) Tener como finalidad el estudio de lenguas muertas (latín, sánscrito…).
c) Ser necesario para llevar a cabo los estudios lingüísticos: esos datos constituían el único acercamiento posible, pues esas lenguas ya no contaban con hablantes vivos.

Con el avance del siglo XIX y hasta mediados del XX se siguió empleando esta forma de trabajar basada en la recopilación de una gran cantidad de datos escritos (corpus) para:

i) Dar cuenta del proceso de adquisición del lenguaje infantil a través de la transcripción de las interacciones de los niños con sus padres.
ii) Establecer convenciones ortográficas.
iii) Obtener listas de vocabulario para la enseñanza de segundas lenguas.
iv) Realizar estudios comparativos de lenguas.

[2] Observemos la definición que proporciona el *DRAE* (2001): "Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación".
[3] McEnery y Wilson (1996) así como McEnery, Xiao y Tono (2006) proporcionan informaciones claras e interesantes al respecto, por lo que remitimos a su lectura para más detalles sobre los siguientes subapartados.

v) Elaborar gramáticas descriptivas.

1.2. Primera lingüística de corpus

Sin embargo, fue la lingüística estructural americana la que, durante la primera mitad del siglo XX, sentó las bases de la lingüística de corpus como metodología empírica basada en la observación de datos, aunque el término como tal ('lingüística de corpus') no aparecerá hasta más tarde, a principios de los 80. Consideraban estos investigadores el corpus como la 'única' herramienta válida para el estudio de las lenguas, ya que se pensaba que por sí mismo podía proporcionar los datos necesarios para una descripción exhaustiva de las mismas. Este nuevo concepto de corpus, el 'corpus estructuralista', se caracteriza por:

a') Ser un conjunto de muestras orales o transcripciones en papel (datos).
b') Tener por finalidad el estudio de lenguas vivas, pero no documentadas previamente por escrito (lenguas amerindias).
c') Ser necesario, ya que la recogida de datos orales era la única forma de acceder al conocimiento de esas lenguas.
d') Centrarse en los aspectos fonéticos y (morfo)fonológicos, niveles en los que es posible realizar un inventario de todos los elementos implicados, dada su naturaleza finita.
e') No atender a cuestiones de representatividad: debido a que el análisis tenía que efectuarse de forma manual y visual era imposible manejar un número elevado de datos, de ahí una de las principales críticas que recibirá la metodología, por su parcialidad a la hora de describir la realidad.

1.3. Críticas de Chomsky y Abercrombie

Precisamente la irrupción de la figura de Chomsky en el panorama lingüístico a finales de los 50 va a suponer un cambio radical de enfoque en los estudios sobre el lenguaje, ya que con él se impone el racionalismo como filosofía de fondo que debe guiar todas las investigaciones. El resultado será que el trabajo basado en corpus, una metodología empírica por excelencia, va a ser objeto, durante los años 60 y 70, de duras críticas que provocarán que haya una cierta discontinuidad entre los primeros trabajos en lingüística de corpus de los estructuralistas americanos y la lingüística de corpus actual.

Como consecuencia de estas críticas, se produjo un desprestigio general de la metodología basada en corpus (empirismo) a favor de una nueva ortodoxia en los estudios lingüísticos: un acercamiento basado en las intuiciones del lingüista (racionalismo).

Tabla 1. Empirismo vs. racionalismo

Empirismo	Racionalismo
Actuación	**Competencia**
Corpus	**Intuición**

Sin embargo, hay que destacar que en determinados campos el uso de muestras reales de la lengua era imprescindible, por lo que durante estas décadas se siguieron elaborando y utilizando corpus en:

i) Fonética: requiere datos externos, no juicios de valor.
ii) Adquisición de lenguas: los niños no han desarrollado su capacidad metalingüística, por lo que no pueden emitir juicios de valor sobre su lengua.
iii) Lingüística histórica: en muchos casos, no se podía recurrir a los hablantes.

Las críticas vertidas contra la lingüística de corpus fueron de índole teórica (Chomsky) y de índole práctica (Abercrombie).

1.3.1. Críticas teóricas (Chomsky)

Las críticas de Chomsky a la lingüística de corpus se basan en dos hechos fundamentales:

I. La apelación al recurso de la intuición, a la introspección del lingüista, como el único criterio válido para el estudio de la lengua.
II. El papel central otorgado a la sintaxis en las primeras versiones del modelo generativista.

Así, los corpus no se consideran instrumentos válidos, ya que desde la perspectiva del modelo teórico propugnado por Chomsky:

1) Los corpus dan cuenta de la actuación, que es la evidencia externa de la lengua, sujeta a variaciones o desviaciones de la norma de diverso tipo procedentes de limitaciones de la memoria, distracciones, errores, etc. Pero la labor del lingüista es reflejar la competencia: el conocimiento interiorizado de la lengua que posee un hablante-oyente ideal y que le permite discriminar las secuencias gramaticales de las agramaticales y que es ajena a las circunstancias materiales o de otro tipo que puedan afectar a la comunicación.

2) La parcialidad de los corpus: son incompletos, ya que no contienen todas las oraciones de una lengua, y son sesgados, ya que la inclusión de un elemento lingüístico vendrá determinada por su frecuencia de uso, es decir, los elementos más habituales estarán mejor reflejados en el corpus que aquellos más raros. El corpus es por definición cerrado, finito, tiene unos límites; por lo tanto, no puede dar cuenta de la naturaleza no finita, ilimitada de las lenguas. Estas se caracterizan por su infinita capacidad generativa: con un inventario limitado de

signos son capaces de 'generar' infinitas combinaciones de los mismos (capacidad creativa del lenguaje), especialmente en lo que a la sintaxis se refiere. Lo único finito, limitado y, por lo tanto, susceptible de estudio, son las reglas, pero no las combinaciones de signos obtenidas con esas reglas.

3) Por último, los corpus tampoco son la mejor forma de trabajar ni siquiera en términos de metodología, ya que el recurso a la competencia, a la intuición del hablante, nos ahorra tiempo frente a la búsqueda en un corpus: no necesitamos contrastar los datos, ya que con los conocimientos que tenemos aquellos son redundantes e innecesarios. Por otra parte, sólo la introspección nos permite determinar la gramaticalidad de un enunciado o resolver ambigüedades. En el corpus están los datos, pero la decisión última sobre su validez corre a cargo del lingüista y de su intuición.

Tabla 2. Primera lingüística de corpus vs. Chomsky

Primera lingüística de corpus	Chomsky
Se centra en la fonética y la fonología	Se centra en la sintaxis
La lengua se concibe como finita	La lengua se concibe como no finita
El corpus como única explicación	La intuición como única explicación
Los corpus son completos	Los corpus son parciales

1.3.2. Críticas prácticas (Abercrombie)

Aparte de las críticas teóricas de Chomsky, existían problemas prácticos en la primera lingüística de corpus. El procesamiento de datos era lento, propenso al error y caro, al tener que ser realizado por grupos de personas. Abercrombie (1965) tildó el acercamiento basado en corpus de 'pseudo-técnicas'.

Estas críticas prácticas eran correctas. La primera lingüística de corpus requería habilidades de procesamiento de datos que no estaban disponibles en ese tiempo. El hecho de que el trabajo tuviera que ser llevado a cabo por personas encarecía el proceso y reducía la fiabilidad de los análisis. Será la llegada del ordenador la que dé un nuevo impulso a la lingüística basada en corpus.

1.4. Segunda generación de lingüística de corpus

En este clima poco favorable de opinión de los años 60 y 70, se empezó a gestar, aunque al margen de la corriente lingüística dominante, lo que sería la segunda generación de trabajos en lingüística de corpus, marcada ahora por la presencia del ordenador, aunque algunos de los corpus que se recopilan durante este período no fueron diseñados para su informatización.

Fue en Estados Unidos donde se abordó la compilación del primer corpus informatizado sistemáticamente organizado. Desde entonces, los corpus electrónicos han llegado a erigirse en recursos imprescindibles para diversos

fines relacionados con la investigación lingüística. Las características más destacadas de los corpus de estas décadas son:

a'') La presencia del ordenador: sólo en los años 60 los ordenadores alcanzaron una potencia de procesamiento y una capacidad de almacenamiento suficientes para poder albergar grandes cantidades de texto, aunque en un principio no todos los proyectos para recopilar corpus se concebían pensando en su informatización. No obstante, el vínculo entre los corpus y los ordenadores ya había sido establecido a finales de los 40 por R. Bussa (McEnery, 2003: 452).
b'') Carácter representativo de los datos: la mayoría de los proyectos de elaboración de corpus pretendía recoger textos escritos que dieran cuenta del estado de la lengua en ese momento. Durante la década de los 50 (McEnery, 2003: 452), A. Juilland estableció los conceptos de marco de la muestra, representatividad y equilibrio, básicos en el concepto actual de corpus.
c'') Tendencia a desfavorecer los datos orales por las dificultades técnicas y de transcripción. Predominan los corpus de textos escritos, aunque con notables excepciones.
d'') Tamaño: un millón de palabras.

Algunos corpus destacados de este período:

1) En Inglaterra, R. Quirk (University College London) sentó en 1959 las bases para la elaboración del 'Survey of English Usage Corpus' (SEU)[4], corpus amplio y variado estilísticamente (un millón de palabras) que empezó a recopilarse en 1961 con la intención de constituirse en una descripción sistemática del inglés británico hablado y escrito –los textos comprenden el período 1955-1985–, proyecto que sería completado por S. Greenbaum. Aunque en la actualidad está informatizado, no se diseñó como corpus electrónico. Sin embargo, marcó las pautas de la futura lingüística de corpus.
2) En EE.UU., N. Francis y H. Kucera llevaron a cabo la recopilación del 'Brown University Corpus of American English' ('Brown Corpus')[5]: corpus de un millón de palabras procedentes de 500 muestras de prosa de

[4] Fue el primer proyecto llevado a cabo en Europa relacionado con la recopilación de un corpus para describir y analizar la lengua. Consta de 200 textos, cada uno con una extensión de 5000 palabras. Originariamente los textos fueron recogidos en fichas de papel y cintas de audio que, con posterioridad, se informatizaron y se anotaron de forma automática (asignación de la categoría gramatical de las palabras).

[5] El propósito del corpus, además de describir el inglés americano, era definir un estándar para el desarrollo de estudios comparativos basados en recopilaciones similares de la lengua inglesa, pero también de otras lenguas, efectuadas según los mismos parámetros de diseño. El análisis computacional de los textos desveló fenómenos bien conocidos hoy en día, como que las palabras más frecuentes son aquellas que carecen de contenido ('the', 'of') o que gran parte de las palabras con significado léxico aparece solo una vez en un corpus. Además de las aportaciones a la 'léxico-estadística', sirvió de base para la elaboración del primer diccionario basado en corpus, el 'American Heritage Dictionary', publicado en 1969.

diferentes estilos, de unas 2000 palabras cada una, obtenidas de publicaciones realizadas en Estados Unidos durante 1961. Trata de representar el inglés americano de esa fecha en su modalidad escrita. Es el primer corpus concebido de forma íntegra para ser informatizado.

3) El 'Lancaster-Oslo/Bergen Corpus' (LOB) es el resultado de los esfuerzos coordinados de G. Leech (Universidad de Lancaster), S. Johansson (Universidad de Oslo) y el 'Norwegian Computing Centre for the Humanities' en Bergen: se trata también de un corpus de un millón de palabras que recoge muestras de inglés británico escrito en 1961. Se elaboró siguiendo los mismos criterios de diseño que el 'Brown Corpus' con el objetivo de constituir su equivalente en inglés británico. Se compiló entre 1970-1978. Existe versión anotada.
4) J. Svartvik (Universidad de Lund) inició en 1975 la informatización de la parte correspondiente a textos orales no transcritos del SEU en el proyecto conocido como 'Survey of Spoken English' (SSE), que daría lugar al 'London-Lund Corpus of Spoken English' (LLC)[6]. Se trata de un recurso todavía inigualado para el estudio del inglés hablado. Contiene medio millón de palabras de inglés británico oral procedente de grabaciones realizadas entre 1953 y 1987.

Estos corpus son muy pequeños si los comparamos con los estándares de la actualidad, pero todavía se usan en la investigación debido a la utilidad de una estructura bien planificada y representativa, un rasgo que contrasta fuertemente con algunas de las colecciones a gran escala del momento.

1.5. Revisión de las críticas de Chomsky y Abercrombie

Así pues, la metodología de corpus continuó durante los 60 y los 70, pero como una metodología minoritaria, pese a la importancia que desde la perspectiva actual tienen algunos de los proyectos emprendidos en la época.

En el resurgir de la lingüística de corpus que ocurrió en la década de los 80 tuvieron especial importancia diversos autores, entre los que sobresale G. Leech (1992), que rebatieron las críticas teóricas y prácticas que se habían formulado contra la primera lingüística de corpus. Si bien dichas críticas eran parcialmente válidas en su momento, según la opinión de Leech, la mayoría de las desventajas de los corpus se habían exagerado o no eran ciertas, sobre todo gracias a la evolución de los ordenadores. Algunos de los argumentos de G. Leech a favor del uso de corpus son:

i) El corpus como metodología científica: desde el punto de vista del método científico, el corpus ofrece una serie de ventajas frente a la intuición[7], ya que está sujeto a verificación, lo que descarta el recurso a

[6] Las grabaciones se transcribieron ortográficamente y se anotó la información relativa a la prosodia (unidades y movimientos tonales, pausas, acentos…) y a los rasgos paralingüísticos.
[7] Rebate la tercera crítica de Chomsky.

ejemplos inventados por los lingüistas de forma interesada. Además, en el caso de datos cuantitativos, como la frecuencia, la intuición no es un recurso válido: nuestra percepción de la frecuencia es totalmente subjetiva.

ii) La gramaticalidad de los enunciados de un corpus: la mayoría de enunciados de un corpus es gramatical, por lo que los corpus reflejan la competencia[8]. Según Chomsky, los corpus, como muestras de uso de la lengua (actuación), no eran más que un pobre reflejo de la competencia. Sin embargo, los trabajos de Labov (1969) mostraron el alto porcentaje de secuencias gramaticales en un corpus.

iii) La utilidad de los datos cuantitativos: los corpus son una fuente inigualable para la extracción de este tipo de datos[9]. Si el corpus está bien diseñado, los datos relativos a la frecuencia de uso serán representativos de la lengua en su totalidad.

iv) Con el uso del ordenador, el procesamiento de los datos de un corpus no es un conjunto de 'pseudo-técnicas'[10]. Los ordenadores son capaces de procesar gran cantidad de datos a un coste reducido, de forma mucho más rápida que las personas y sin cometer errores.

Tabla 3. Lingüística de corpus vs. Generativismo

Lingüística de corpus	Generativismo
• Datos	• Juicios del hablante
• Externos	• Internos
• Públicos	• Privados
• Observables	• No observables
• Verificables	• No verificables
• Naturales	• Artificiales
• Noción exacta de frecuencia	• Noción vaga de frecuencia

1.6. Renacer de la lingüística de corpus

Superadas las críticas –teóricas y prácticas– y con las nuevas ventajas y posibilidades que ofrecían los ordenadores, los corpus electrónicos se convierten desde la década de los 80 en un recurso indispensable para el estudio del lenguaje, para probar hipótesis lingüísticas y para construir sistemas de procesamiento del lenguaje natural. Solo entonces se generaliza el término, especialmente a partir de 1984, año en que J. Aarts y W. Meijs editaron el volumen titulado 'Corpus Linguistics I: Recent Developments in the Use of Computer Corpora', y se empieza a hablar de 'lingüística de corpus' en el sentido actual. Algunos hechos que favorecieron este renacer de la lingüística de corpus como metodología de trabajo en lingüística fueron:

[8] Rebate la primera crítica de Chomsky.
[9] Rebate la segunda crítica de Chomsky.
[10] Rebate la crítica de Abercrombie.

1) El auge de las áreas aplicadas de la lingüística en general y de la lingüística computacional en particular, lo que ha puesto en evidencia la necesidad de contar con datos de uso de la lengua, con datos procedentes de la actuación, tanto de hablantes nativos como no nativos. Por una parte, los corpus reflejan la variedad de la lengua y, por otra, pueden recoger estructuras nuevas o que no se ajustan a las descripciones teóricas –a las que el lingüista no podría haber accedido desde su competencia– y que, sin embargo, requieren una explicación. Además, en el caso de hablantes no nativos, los corpus son una muestra excelente de evidencias de uso de la lengua.
2) El eclecticismo: el uso de corpus no se concibe como incompatible con el recurso a los juicios del lingüista; por sí solos ni los corpus (postura de los estructuralistas americanos) ni los juicios o intuiciones del hablante-oyente ideal (postura de Chomsky) son suficientes para explicar los fenómenos lingüísticos. En la actualidad, se reconoce que los corpus, al suplir datos textuales de primera mano, no se pueden analizar válidamente sin la intuición y la facultad interpretativa del analista, que usa conocimientos de la lengua (como hablante nativo o no nativo competente) y conocimientos acerca del lenguaje (como lingüista).
3) La mayor disponibilidad de corpus electrónicos, sobre todo gracias a las posibilidades que ofrece Internet para la obtención de textos en dicho formato.
4) El desarrollo de nuevas tecnologías para la informatización de textos de forma más rápida, como el OCR o reconocimiento óptico de caracteres, el dictado automático, etc.
5) La utilidad de los datos cuantitativos en el estudio de determinados aspectos del lenguaje.
6) En Lingüística Computacional en particular, el desarrollo de productos comerciales que se empieza a producir en los 80 pone de manifiesto que los formalismos gramaticales del momento[11], que tan elegantes resultaban desde una perspectiva puramente teórica, no eran útiles para tratar los textos reales producidos por los hablantes:

   – Necesidad de contar con vocabularios o diccionarios más extensos para ampliar la cobertura de los sistemas: sistemas capaces de trabajar con cualquier tipo de texto y no solo con sublenguajes o lenguajes limitados a dominios muy restringidos (p. ej. el de los partes meteorológicos).
   – Necesidad de manejar frecuencias, estadísticas y cálculos de probabilidades para manipular cantidades cada vez más grandes de

[11] Es cuando surge la familia de gramáticas de unificación, inaugurada por la gramática de cláusula definida de Pereira y Warren (1980), a la que siguieron la gramática de estructura de frase generalizada, la gramática léxico-funcional, la gramática de unificación funcional, etc. *Vid.* Moreno Sandoval (2001) para una descripción detallada de estas gramáticas.

texto y para desarrollar sistemas de extracción automática de reglas, así como desambiguadores estocásticos[12].

Como consecuencia de estos hechos, los grandes corpus textuales se erigen en uno de los recursos fundamentales de la llamada 'ingeniería lingüística' o, más recientemente, 'tecnologías del lenguaje', áreas de la Lingüística Computacional en los que son necesarios para desarrollar sistemas prácticos; por otra parte, sin ellos tampoco se concibe hoy en día el desarrollo de gramáticas y de léxicos computacionales. Sobre estas premisas nace la lingüística de corpus tal y como se entiende en la actualidad, definida como "el área de la lingüística especializada en el aprovechamiento de los corpora (*sic*)" (Abaitua, 2002: 62), 'the study of language on the basis of text corpora' (Aijmer y Altenberg, 1991: 1) o 'the use of large collections of text available in machine-readable form' (Svartvik, 1992: 7). Estos corpus se van a caracterizan por:

a''') Formato electrónico: conjunto de textos informatizados.
b''') Tamaño en aumento progresivo, hasta superar los cien millones de palabras, aunque existen corpus de tamaños inferiores.
c''') Carácter abierto: corpus no cerrados, sino en continua actualización (corpus monitor).
d''') Vertiente comercial: los corpus no se limitan a centros de investigación, sino que muchos proyectos son desarrollados por consorcios comerciales, principalmente editoriales.
e''') Se amplía el repertorio de lenguas que disponen de corpus y también se elaboran corpus multilingües.
f''') Automatización de las diferentes tareas de procesamiento de los textos de un corpus gracias a avances técnicos que permiten realizar, de forma automática o semiautomática, la asignación de categoría gramatical a cada una de las palabras del corpus, la desambiguación, la extracción de concordancias o ejemplos en contexto, el alineamiento de las grabaciones de audio con su correspondiente transcripción, etc.
g''') Estímulo para el desarrollo de nuevos modelos y campos de investigación en lingüística, así como para la realización de estudios sobre los más variados aspectos lingüísticos, desde los gramaticales hasta los discursivos, pasando por los históricos, los psicolingüísticos o los culturales.

Algunos ejemplos de corpus que responden a estas características son el 'Bank of English' y los bancos de datos de la Real Academia Española, CREA y CORDE:

[12] Estos programas toman decisiones de forma automática en casos de ambigüedad categorial, semántica, sintáctica, etc. La decisión depende del peso estadístico que cada una de las opciones tenga. Esta información en ocasiones se combina con conocimiento lingüístico.

1) 'The Bank of English', corpus de más de 524 millones de palabras de inglés moderno tanto escrito como oral de diferentes procedencias. El proyecto, conocido en la actualidad como Proyecto COBUILD[13], se desarrolla en la Universidad de Birmingham bajo la dirección de J. Sinclair en colaboración con la editorial Collins COBUILD. El corpus se lanzó en 1991, pero COBUILD ya llevaba desde 1980 recopilando textos electrónicos para elaborar sus diccionarios. Está integrado en la 'Collins Word Web', una base de datos de más de dos billones y medio de palabras, a la que cada mes se le añaden otros treinta y cinco millones. Se trata del recurso de este tipo más grande existente en el mundo[14].
2) 'Corpus de Referencia del Español Contemporáneo' (CREA)[15], banco de datos del español contemporáneo (desde 1975 a la actualidad) elaborado por la Real Academia Española. Dispone de ciento sesenta millones de palabras (abril de 2005) procedentes de textos escritos y orales.
3) 'Corpus Diacrónico del Español' (CORDE)[16], banco de datos del español elaborado por la Real Academia Española. Dispone de doscientos cincuenta millones de palabras (abril de 2005), procedentes de textos escritos desde los orígenes del idioma hasta 1975.

---

[13] El corpus comprende textos de diferentes variedades de inglés: británico, americano, canadiense y australiano. Igual que en el anterior, se han introducido marcas para indicar la categoría gramatical; por otra parte, se han analizado sintácticamente unos doscientos millones de palabras. Otra característica reseñable es que constantemente se introducen nuevos materiales para mantener el corpus lo más actualizado posible. Además del corpus mismo, el equipo de lexicógrafos y lingüistas con que cuenta el proyecto ha desarrollado herramientas adicionales para analizar el corpus y extraer todo tipo de información: patrones de combinación de palabras o colocaciones, frecuencias de aparición, ejemplos de uso, etc. La meta es examinar estos datos para conseguir diccionarios y materiales de referencia sólidos.

[14] Otro recurso importante para el inglés es el 'British National Corpus' (BNC), un corpus de unos cien millones de palabras de inglés británico contemporáneo, tanto escrito como hablado. El proyecto depende de un consorcio académico-industrial liderado por la editorial Oxford University Press junto a otras editoriales especializadas en diccionarios, la Universidad de Lancaster, la Universidad de Oxford y la British Library. Se inició en 1991 y se finalizó en 1994, por lo que es de naturaleza cerrada. Con el objetivo general de representar el inglés británico de finales del siglo XX para desarrollar materiales de referencia y llevar a cabo investigaciones lingüísticas, está formado por un 90% de textos escritos y un 10% de textos orales. El corpus ha sido anotado de forma automática, de tal manera que cada palabra lleva una etiqueta indicativa de su categoría gramatical.

[15] El tamaño del corpus alcanzará los ciento setenta millones de palabras en cuanto terminen de introducirse los textos correspondientes al período 2000-2004. Sigue el mismo modelo que el BNC, en el sentido de que el 90% de los textos son escritos y el 10% restante, orales. Considera criterios geográficos, temáticos, cronológicos y de medio de publicación a la hora de seleccionar los textos. Como el 'Bank of English', es un corpus monitor: periódicamente se introducen nuevos textos para cumplir con el objetivo del corpus, ser representativo del español contemporáneo. Se trata del recurso de este tipo más importante para el español, por lo que, además de servir como fuente de datos reales para las obras académicas, sus aplicaciones tanto para la investigación como para el diseño de productos comerciales son numerosas.

[16] Se trata de un corpus complementario del CREA. Juntos suman un total de cuatrocientos diez millones de palabras para el español. Su principal aplicación está relacionada con su uso para los estudios diacrónicos, en especial para proveer material al relanzado proyecto del 'Diccionario histórico' de la RAE.

Quizá no tan importantes en tamaño, pero sí en cuanto al objetivo que persiguen, son algunos proyectos relacionados con otras lenguas peninsulares:

4) 'Corpus de Referencia do Galego Actual' (CORGA), dirigido por el académico Guillermo Rojo y su equipo en la Universidad de Santiago de Compostela. Con sus diecisiete millones y medio de palabras, prevé alcanzar los veinticinco. Como los anteriores, pretende ser representativo de una lengua, la gallega, motivo por el cual recoge textos de diferente procedencia temática, geográfica y cronológica. En este caso, estamos ante un proyecto cerrado.
5) 'Corpus Textual Informatitzat de la Llengua Catalana' (CTILC): este corpus de cincuenta y dos millones de palabras procedentes de textos escritos en catalán entre 1832 y 1988, se inscribe dentro de las actividades lexicográficas del 'Instituto d'Estudis Catalans'. Terminado en 1997, se recopiló inicialmente para elaborar el 'Diccionari del Català Contemporani' (DCC), aunque con el tiempo se han ampliado sus usos.
6) 'Corpus estadístico del euskera del siglo XX', con más de cuatro millones y medio de palabras procedentes de textos escritos en vasco durante el siglo XX. El proyecto se desarrolló entre 1987 y 1999 en el Centro Vasco de Terminología y Lexicografía (UZEI) con el fin de reflejar el uso de la lengua vasca durante el período comprendido.

2. El concepto de corpus

Como se aprecia en el apartado anterior, en la actualidad el concepto de corpus ha cambiado mucho con respecto al que manejaban los primeros lingüistas que lo empleaban como recurso para sus investigaciones.

Hoy en día se considera que los corpus deben cumplir los siguientes requisitos:

1) Formato electrónico: un corpus, para ser una herramienta útil al lingüista, debe estar informatizado, es decir, los textos de que consta tienen que estar en formato electrónico (corpus informatizado o automatizado). El hecho de que para los primeros corpus no se pudiera disponer de ordenadores motivó la crítica de las pseudo-técnicas: el procesamiento de los datos debía efectuarse de forma manual, con los errores y problemas que eso ocasionaba. Sin embargo, el empleo del ordenador permite automatizar tareas tales como:
    a. Búsqueda de información: un corpus informatizado permite localizar de forma rápida una palabra, secuencia de palabras o incluso una categoría o esquema gramatical en décimas de segundo.
    b. Recuperación de información: un corpus informatizado permite obtener todos los casos de una palabra, secuencia de palabras, etc. registrados en el corpus, normalmente con su contexto inmediato anterior y posterior (lo que se conoce como 'concordancia').

   c. Cómputo de la frecuencia de aparición de una palabra, secuencia de palabras, etc.
   d. Clasificación de los datos contenidos en el corpus según diferentes criterios: orden alfabético, frecuencia de aparición, autor, procedencia geográfica, tema, medio de publicación, etc.
2) Autenticidad de los datos: los textos recogidos en el corpus han de ser muestras reales de uso de la lengua objeto de estudio. A partir de ellas se construyen de forma empírica las teorías que tratan de explicar el funcionamiento de la lengua o las aplicaciones computacionales.
3) Criterios de selección: los textos que forman parte del corpus deben haber sido elegidos de acuerdo con unos determinados criterios –lingüísticos y/o extralingüísticos– condicionados siempre por la finalidad concreta que persiga el corpus[17].
4) Representatividad: la selección de los textos, además de a unos criterios adecuados, debe responder a parámetros estadísticos que garanticen que los textos 'representan' la variedad de lengua objeto de estudio ('muestra representativa'). Esta variedad puede referirse a la obra de un autor determinado, a un período de tiempo, a un género, etc. Cuando lo que nos interesa es la lengua en su conjunto, la opción de reunir en un corpus todas las muestras de esta se hace impracticable, a diferencia, p. ej., de lo que ocurre si queremos recoger todas las obras de Cervantes, que son un universo cerrado. La única solución posible, entonces, es tomar una muestra más pequeña de esa lengua, que refleje, a pequeña escala, el funcionamiento del todo que es la lengua. Como Chomsky criticó con acierto, los corpus corren el riesgo de ser sesgados. Para subsanar este problema se recurre a la selección según criterios estadísticos de textos de diversos géneros, tipologías, temas, medios de publicación, etc.
5) Tamaño: por lo general, los corpus constan de un tamaño finito, que se suele medir en millones de palabras o de formas y que se fija antes de empezar la recogida de los textos (p. ej. un millón de palabras); una vez alcanzado ese número, se da por terminada la recopilación del corpus, que no es más que el primer paso de todo el proceso[18]. Sin embargo, también existen corpus abiertos o monitor, como el del proyecto COBUILD dirigido por J. Sinclair en la Universidad de Birmingham, de especial interés para la lexicografía, o el propio CREA de la Real Academia Española. En el pasado se pensaba que el tamaño era muy importante: mientras mayor fuera el corpus, más posibilidades tenía de reflejar el funcionamiento real de la lengua en todas sus variedades, pero en la actualidad se priman los criterios de diseño, es decir, el tamaño solo es importante en la medida en que así lo exija la finalidad del corpus[19].

[17] Precisamente el uso de unos criterios previos diferencia un corpus de otras recopilaciones de textos tales como archivos, colecciones o bibliotecas electrónicas.

[18] Una vez recogidos los textos, estos se codifican, anotan y explotan de múltiples formas, por lo que la recopilación en sí no es más que una primera fase necesaria en todo proyecto de corpus.

[19] Lógicamente, un corpus que pretenda ser representativo de una lengua en toda su variedad (español, inglés, francés…) no podrá conformarse con unos pocos millones de palabras,

A continuación se recogen algunas definiciones que ilustran estas características[20]:

> A *collection of texts* assumed to be *representative* of a given language, dialect, or other subset of a language, to be used for *linguistic analysis* (Francis, 1982: 7 *apud* Francis, 1992: 17).
>
> A *collection of pieces of language* that are *selected* and *ordered* according to *explicit linguistic criteria* in order to be used as a *sample* of the language (Sinclair, 1994: 14).
>
> A *finite-sized* body of *machine-readable texts sampled* in order to be maximally *representative* of the language variety under consideration (McEnery & Wilson, 1996: 24).
>
> Un corpus és una *mostra* d'una llengua que habitualment s'ha construït a partir d'una *selecció* de textos realitzada segons uns determinats *criteris* i amb un determinat *objectiu* (Martí y Castellón, 2000: 151).
>
> The term *corpus* should properly only be applied to a *well-organized collection of data*, collected within the boundaries of a *sampling frame* designed to allow the *exploration* of a certain linguistic feature (or set of features) via the collected data" (McEnery, 2003: 449).
>
> Un corpus es un *conjunto de textos* de lenguaje *natural* e *irrestricto*, almacenados en un *formato electrónico homogéneo*, y *seleccionados* y *ordenados*, de acuerdo con *criterios explícitos*, para ser utilizados como *modelo* de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el *análisis lingüístico* (Santalla, 2005: 45-46).
>
> Rather, the term *corpus* as used in modern linguistics can best be defined as a *collection of sampled texts*, written or spoken, in *machine-readable form* which may be annotated with various forms of linguistic information (McEnery, Xiao y Tono, 2006: 4).

3. Tipos de corpus

Establecidos la metodología y el concepto de corpus, en este apartado pasamos a comentar algunos de los principales tipos de corpus. Autores como J. Sinclair (1996) o J. Torruella y J. Llisterri (1999) han propuesto sendas clasificaciones en función de una serie de criterios, aunque en la práctica no siempre está clara ni se hace explícita la tipología de un corpus.

En general, los principales parámetros para establecer tipologías de corpus se centran en:

i) La modalidad de la lengua
ii) El número de lenguas a que pertenecen los textos
iii) Los límites del corpus

---

mientras que un corpus cuyo objetivo sea describir un sublenguaje (jurídico, informático….) puede permitirse un tamaño más reducido.

[20] Se han destacado en cursiva las características relevantes.

iv) El carácter general o especializado de los textos
v) El período temporal que abarcan los textos
vi) El tamaño de los textos
vii) El tratamiento aplicado al corpus.

Con frecuencia, estos criterios vienen determinados por la finalidad u objetivo que se persigue con el corpus: el estudio de la obra de un autor (Cervantes) o de la producción literaria de una época determinada (el Barroco), la descripción de una lengua en general (el español contemporáneo) o de una variedad, sublenguaje o aspecto lingüístico concreto (p. ej. la norma culta en Madrid, textos técnicos, léxico jurídico, etc.), la obtención de un determinado producto comercial (un diccionario, una aplicación telefónica relacionada con las tecnologías del habla, etc.).

Algunos de los principales tipos de corpus son:

i) Según la modalidad de la lengua, se distinguen tres tipos de corpus: corpus escritos, corpus orales y corpus mixtos.

-Los corpus escritos están formados únicamente por muestras procedentes de la modalidad escrita de la lengua. Véase, p. ej., el CTILC.

-Los corpus orales, por su parte, únicamente recogen muestras de lengua hablada, que pueden ser transcripciones ortográficas de grabaciones, utilizadas sobre todo en lingüística de corpus, o grabaciones acompañadas de transcripciones ortográficas y/o fonéticas, más usadas en lingüística y tecnologías del habla. P. ej. el 'Arquivo Sonoro de Galicia' (ASG), que contiene grabaciones de diferentes variedades del gallego así como su transcripción ortográfica.

En ocasiones, se reserva el término 'corpus oral' para los corpus orientados a la descripción fonética de la lengua o para el desarrollo de sistemas de reconocimiento o síntesis en el ámbito de las tecnologías del habla. Los corpus de este tipo se graban bajo condiciones muy controladas y suelen consistir en segmentos aislados, unidades de síntesis (para convertir texto a habla), dígitos (para los sistemas de reconocimiento del habla), frases aisladas, textos leídos o grabaciones y transcripciones de diálogos naturales entre personas, o entre personas y simulaciones de sistemas informáticos, que se emplean para desarrollar servicios automáticos a través del teléfono (venta de entradas, consulta de horarios de transportes públicos, servicios bancarios, etc.). En general, se diseñan con mucho cuidado para recoger el fenómeno objeto de estudio y tienen un tamaño reducido al no utilizar un número elevado de hablantes. En el caso de los corpus orales para fonética, suelen consistir en bases de datos de sonidos, realizadas en laboratorios de fonética. Por lo que respecta a los corpus orales para las tecnologías del habla, abundan los proyectos nacionales y europeos, así como los de empresas de telecomunicaciones (Telefónica I + D) e Informática (IBM). Véase, por ejemplo, el proyecto SpeechDat.

Los corpus desarrollados en el marco de la lingüística de corpus se denominan, desde esta perspectiva más restringida, 'corpus de lengua oral': en este caso, las grabaciones se realizan en entornos naturales y se favorecen las muestras espontáneas, no planificadas (diálogos, conversaciones, discursos, muestras procedentes de medios de comunicación…). El objetivo no es tanto el análisis de las características de tipo fonético y prosódico, sino contar con una transcripción ortográfica de la lengua hablada para efectuar diferentes análisis lingüísticos sobre el texto transcrito, no sobre la grabación: estudios sociolingüísticos (p. ej. el proyecto PRESEA para el estudio sociolingüístico del español de España y de América), discursivos ('Corpus de conversación coloquial' del grupo Val.Es.Co.), etc.

-Los corpus mixtos combinan ambas modalidades, aunque siempre favoreciendo la lengua escrita, ya que su obtención es menos costosa que la de la lengua oral que, además, siempre requiere un proceso posterior de transcripción de las grabaciones. Corpus como el BNC o el CREA pertenecen a este tipo: el 90% de sus textos son escritos y el 10%, orales.

ii) Según el número de lenguas, los corpus se clasifican fundamentalmente en monolingües y bilingües o multilingües.

-Los corpus monolingües están compuestos por textos en una sola lengua. Se recopilan con el objetivo de dar cuenta de una lengua o variedad lingüística en general (o de un subconjunto de la misma). Es el caso del CREA, del CORGA, etc.

-Los corpus bilingües o multilingües están formados por textos en dos (bilingües) o más lenguas (multilingües) sin que, en principio, sean traducciones unos de otros y sin compartir criterios de selección.

Cuando el corpus consiste en una selección de textos en más de una lengua, parecidos en cuanto a sus características y que comparten criterios de selección, se habla de 'corpus comparables' o 'paired texts'. Se utilizan sobre todo para comparar variedades de la lengua en estudios contrastivos. El más conocido es 'The International Corpus of English' (ICE), un corpus en el que desde 1990 se están recopilando materiales escritos y orales posteriores a 1989 pertenecientes a diferentes variedades de inglés. En la actualidad, están en marcha 19 proyectos en otros tantos países (desde Australia hasta Estados Unidos, pasando Jamaica, Nueva Zelanda o Pakistán, la última incorporación). Cada corpus consta de un millón de palabras y todos siguen el mismo esquema de diseño y de anotación.

Cuando el corpus contiene textos en más de una lengua pero, a diferencia de los anteriores, se trata de los mismos textos y sus traducciones o equivalentes en una o más lenguas, se usa el término 'corpus paralelos' o 'bi-texts'. Son especialmente útiles en los estudios de traducción y en entornos bilingües o multilingües, como la ONU, la OTAN, la UE o Canadá. Como ejemplo, puede consultarse el 'Corpus Paralelo CLUVI' de la Universidad de Vigo.

Si, además, para facilitar su explotación, los textos de un corpus paralelo están dispuestos unos al lado de otros por párrafos o frases, de tal forma que

sea más fácil extraer las equivalencias de traducción (aquellos elementos que son traducciones mutuas), entonces se habla de 'corpus alineados'. Aunque no siempre es un proceso simple, el alineamiento de oraciones y palabras se puede conseguir automática o semiautomáticamente con un alto grado de exactitud. Se utilizan, sobre todo, como entrenamiento para sistemas de traducción automática basados en estadísticas. El CLUVI también ilustra perfectamente este tipo de corpus.

iii) Según los límites establecidos, los corpus se clasifican en: corpus cerrados y corpus abiertos o monitor.
-Los corpus cerrados constan de un número finito de palabras, que se establece de forma previa a la recopilación del corpus. Una vez alcanzado ese número o límite, el corpus se da por finalizado, como ocurre en varios de los corpus ya mencionados (p. ej. el BNC). Este tipo de corpus son útiles cuando interesa estudiar fenómenos estáticos o estados de lengua.
-Los corpus abiertos o corpus monitor, por el contrario, son corpus dinámicos, que se mantienen en constante crecimiento, normalmente mediante la introducción periódica de nuevas cantidades de textos según unas proporciones previamente definidas. Es lo que ocurre con el CREA o el 'Bank of English'. Son un material excelente para los estudios diacrónicos, para observar tendencias de uso, cambios de significado, frecuencias de distribución, etc. No obstante, no están exentos de críticas frente al modelo predominante de corpus, basado en una concepción estática (tamaño finito) y más preocupado por ser equilibrado en cuanto a sus muestras. En cambio, el modelo del corpus monitor suele centrarse en alcanzar un tamaño considerable y prefiere incluir textos enteros en vez de simples muestras.

iv) Según la especificidad de los textos, los corpus pueden ser generales o especializados.
-Los corpus generales pretenden reflejar la lengua o variedad lingüística de la forma más equilibrada posible; cuantos más tipos de textos, modalidades (textos orales, textos escritos), géneros y materias, mejor.
-Los corpus especializados recogen textos que puedan aportar datos para la descripción de un tipo particular de lengua ('sublenguaje'). P. ej., un corpus que slo recoge textos poéticos o jurídicos. Un corpus de este tipo es el 'Corpus textual especializado plurilingüe", desarrollado por el Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra. Consta de textos en catalán, castellano, inglés, francés y alemán sobre economía, derecho, medio ambiente, medicina e informática, con la meta de estudiar cómo funciona la lengua en cada una de esas áreas y extraer información útil para elaborar diccionarios, tesauros, etc.

v) En función del período temporal que abarcan los textos, las principales tipologías de corpus que encontramos son:
-Corpus diacrónicos o históricos: incluyen textos de diferentes etapas temporales sucesivas con el fin de poder observar evoluciones de la lengua

en un período largo, lo que los diferencia de los corpus monitor, que no abarcan períodos temporales tan amplios. P. ej. el CORDE.
-Corpus sincrónicos: su finalidad es permitir el estudio de una o más variedades lingüísticas en un momento determinado del tiempo (año, período…), pero sin prestar atención a su evolución. Son mucho más frecuentes que los corpus diacrónicos.

vi) Según el tamaño de los textos, los tipos más interesantes son:
-Corpus de referencia: aquellos formados por fragmentos de textos, habituales en los corpus que quieren proporcionar una información lo más completa posible sobre una lengua y tienen que incluir textos de diferentes géneros, temáticas, etc.
-Corpus textuales: aquellos que incluyen textos enteros, sin fragmentar. Más habituales cuando el objeto del corpus es un sublenguaje o lenguaje de especialidad.

vii) Según el proceso al que se someta el corpus, se distingue entre:
-Corpus simples, en bruto, no anotados o no codificados: consisten en textos guardados sin formato alguno y sin añadir ningún tipo de información adicional, como pueden ser códigos o anotaciones. Estos corpus, aunque útiles, son muy limitados en cuanto a las posibilidades de extraer información que ofrecen.
-Corpus codificados o anotados: están formados por textos a los que se les han añadido, de forma manual o automática, determinadas informaciones. Estas pueden referirse a la estructura de los textos: etiquetas especiales para indicar el autor, el título, los capítulos, etc. ('codificación'); o, lo que es más interesante, a aspectos puramente lingüísticos, como la categoría gramatical, la estructura sintáctica, etc. ('anotación'). La explicitación de estos datos enriquece el corpus y aumenta considerablemente las posibilidades de explotación que ofrece. El tipo de anotación más frecuente, con diferencia, consiste en añadir una marca a cada palabra del corpus indicando su categoría y características morfosintácticas. P. ej., si se trata de un nombre, de un verbo, etc. Cuando, además, los textos han sido analizados sintácticamente de manera completa, el corpus recibe el nombre de 'treebank'. Cada vez son más habituales este tipo de corpus. Destaca la Base de Datos Sintácticos del Español Actual (BDS) o, más recientemente, el corpus CESS-ECE.

Por supuesto, son posibles más tipologías, pero nos hemos limitado a mencionar aquellas que son más habituales y que están más claramente delimitadas.

4. Conclusión

Los corpus son un recurso hoy en día inigualable para cualquier estudio lingüístico en general y de lingüística computacional en particular. Sus

principales ventajas derivan del hecho de consistir en muestras reales de la lengua, de aportar objetividad y de ofrecer la posibilidad de verificar teorías fácilmente. Con la incorporación de ordenadores cada vez con mayor capacidad de almacenamiento y procesamiento, el acceso a los datos es rápido y fiable, así como su manipulación, extracción y procesamiento de los mismos. Por otra parte, permite obtener estadísticas y datos cuantitativos que de otra forma resultarían muy costosos o imposibles, dado el tamaño que están alcanzando algunos corpus. Por todo ello, encontramos estudios descriptivos de las lenguas apoyados en corpus sobre cualquiera de los niveles lingüísticos (fonético-fonológico, gramatical, semántico, pragmático…); también destaca su aplicación como fuente de datos para la enseñanza de lenguas o la elaboración de materiales didácticos, diccionarios, gramáticas, productos relacionados con la traducción automática o con las tecnologías del habla, etc.

BIBLIOGRAFÍA

Aarts, J. y Meijs, W. (eds.) (1984): *Corpus Linguistics*, Amsterdam, Rodopi.

Abaitua, J. (2002): "Tratamiento de corpora bilingües", en M. A. Martí y J. Llisterri (eds.): *Tratamiento del lenguaje natural: tecnología de la lengua oral y escrita*, Soria/Barcelona, Fundación Duques de Soria/Edicions de la Universitat de Barcelona, 61-90.

Abercrombie, D. (1965): *Studies in Phonetics and Linguistics*, London, Oxford University Press.

Aijmer, K. y Altenberg, B. (eds.) (1991): *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Londres, Longman.

Biber, D., Conrad, S. y Reppen, R. (1998): *Corpus Linguistics. Investigating Language Structure and Use*, Cambridge, C.U.P.

Francis, W. N. (1992): "Language Corpora B.C.", en J. Svartvik (ed.): *Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin/New York, Mouton de Gruyter, 17-32.

Labov, W. (1969): "The logic of non-standard English", *Georgetown Monographs on Language and Linguistics*, 22.

Lavid, J. (2005): *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*, Madrid, Cátedra.

Leech, G. (1992): "Corpora and theories of linguistic performance", en J. Svartvik (ed.): *Directions in Linguistics: Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin, Mouton de Gruyter, 105-122.

Martí Antonín, M. A. y Castellón Masalles, I. (2000): *Lingüística computacional*, Barcelona, Universitat de Barcelona.

McEnery, T. (2003): "Corpus Linguistics", en R. Mitkov (ed.), *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press, 448-463.

McEnery, T. y Wilson, A. (1996): *Corpus Linguistics*, Edinburgh, Edinburgh University Press. Suplemento web disponible en la dirección: http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

McEnery, T., Xiao, R. y Tono, Y. (2006): *Corpus-Based Language Studies. An advanced resource book*, London/New York, Routledge.

Moreno Sandoval, A. (2001): *Gramáticas de unificación y rasgos*, Madrid, A. Machado Libros.

Moure, T. y Llisterri, J. (1996): "Lenguaje y nuevas tecnologías: el campo de la lingüística computacional", en M. Fernández Pérez (coord.): *Avances en Lingüística aplicada*, Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico, 147-227. Disponible electrónicamente en: http://liceu.uab.es/~joaquim/publicacions/llisterri_moure_96.html

Real Academia Española (2001): *Diccionario de la lengua española*, Madrid, Espasa.

Santalla del Río, M.ª P. (2005): "La elaboración de corpus lingüísticos", en M. Cal, P. Núñez e I. M. Palacios (eds.): *Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas*, Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 45-63.

Sinclair, J. (1994): *EAGLES*, Document EAG-CWG-IR-2.

Sinclair, J. (1996): *EAGLES Preliminary recommendations on Corpus Typology*. Documento electrónico en: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Svartvik, J. (1992): "Corpus linguistics comes of age", en J. Svartvik (ed.): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (Stockholm, 4-8 August 1991)*, Berlin/New York:, Mouton de Gruyter, 7-13.

Torruella, J. y Llisterri, J. (1999): "Diseño de corpus textuales y orales", en J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (eds.): *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona, Milenio/Universidad Autónoma de Barcelona, Dpto. de Filología Española, 45-77. Disponible electrónicamente en: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

CORPUS MENCIONADOS

Arquivo Sonoro de Galicia (ASG)
<http://www.consellodacultura.org/arquivos/asg/anosafala.php>
The Bank of English
<http://www.titania.bham.ac.uk/>
Base de Datos Sintácticos del Español Actual (BDS)

<http://www.bds.usc.es/>
The British National Corpus (BNC)
<http://www.natcorp.ox.ac.uk/>
Brown University Corpus of American English (*Brown Corpus*)
<http://icame.uib.no/brown/bcm.html (manual)>
CESS-ECE, Corpus etiquetado sintáctico-semánticamente (español, catalán, euskera)
<http://clic.ub.edu/cessece/>
CLUVI, Corpus Lingüístico da Universidade de Vigo
<http://sli.uvigo.es/CLUVI/info.html>
Corpus de conversación coloquial del grupo Val.Es.Co.
<http://www.uv.es/~valesco/valesco_5.html>
Corpus Diacrónico del Español (CORDE)
<http://www.rae.es/>
Corpus estadístico del euskera del siglo XX
<http://www.uzei.com/antbuspre.asp?nombre=1901&cod=1901&sesion=1>
Corpus de Referencia del Español Contemporáneo (CREA)
<http://www.rae.es/>
Corpus de Referencia do Galego Actual (CORGA)
<http://corpus.cirp.es/corga/>
Corpus textual especializado plurilingüe, IULA
< http://www.iula.upf.edu/corpus/corpuses.htm>
Corpus Textual Informatitzat de la Llengua Catalana (CTILC)
<http://www.iec.cat/gc/ViewPage.action?siteNodeId=690&languageId=1&conten
tId=3284; http://ctilc.iec.cat/>
The International Corpus of English (ICE)
<http://www.ucl.ac.uk/english-usage/projects/ice.htm>
<http://www.ucl.ac.uk/english-usage/ice/index.htm>
Lancaster-Oslo/Bergen Corpus (LOB)
<http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM> (manual)
<http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM> (anotaciones)
PRESEEA, Proyecto para el Estudio Sociolingüístico del Español de España y de América
<http://www.linguas.net/Default.aspx?alias=www.linguas.net/portalpreseea>
Proyecto COBUILD
<http://www.collins.co.uk/books.aspx?group=140>
SpeechDat
<http://www.speechdat.org/>
Survey of English Usage (SEU)
<http://www.ucl.ac.uk/english-usage>
Survey of Spoken English (SSE) y *London-Lund Corpus of Spoken English (LLC)*
<http://icame.uib.no/london-lund/>
<http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM>
Tesouro Medieval Informatizado da Lingua Galega (TMILG)
<http://ilg.usc.es/tmilg/>