Los corpus textuales del español

Guillermo Rojo

1. INTRODUCCIÓN

Un *corpus* es un conjunto de (fragmentos de) textos naturales, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en su totalidad, y reunidos con el propósito de facilitar su estudio científico (cf. Rojo 2014). Esta definición muestra explícitamente que los textos deben ser naturales (no artificiales ni creados expresamente para su incorporación al corpus), han de estar en formato electrónico porque esa es la única forma de que podamos recuperar la información que precisamos, tienen que ser representativos de la variedad de la que proceden y, por último, deben permitir su estudio científico (no exclusivamente lingüístico), lo cual suele implicar la adición de información gramatical, léxica y pragmática a la simple secuencia de formas gráficas que constituyen el texto en el sentido más habitual de la palabra.

Lo anterior significa que, aunque ese rasgo no figure explícitamente en las definiciones habituales, un corpus presupone la existencia de un determinado diseño en su configuración. Esto es, un corpus se construye con la intención de que reúna unos determinados requisitos en cuanto a la procedencia de los textos, la época a la que pertenecen, el tipo al que corresponden, etc. La razón de ello es muy clara: lo que interesa obtener de un corpus es el perfil general de una expresión o estructura en un momento determinado de la historia de la lengua correspondiente, pero también –y en muchos casos, sobre todo– las diferentes características con que la entidad o fenómeno estudiados se presentan en los distintos tipos y subtipos de textos incluidos en el corpus. Un corpus, pues, se opone a lo que en términos tradicionales en lingüística de corpus (LC) se considera un *archivo*, es decir, un conjunto heterogéneo de textos integrados en un recurso único como consecuencia de factores que no se vinculan al deseo de lograr una determinada composición general. Un archivo es una reunión de textos independientes entre sí que son habitualmente analizados de forma individual. Como ha señalado Tognini Bonelli (2010: 18-20), un texto se lee de forma continua, línea a línea, mientras que un corpus no se lee de corrido, sino que se estudia, mediante el análisis de las concordancias, el conjunto de casos del fenómeno en cuestión contenidos en él. Con sus propias palabras, un texto “is an instance of *parole* while the patterns shown up by corpus evidence yield insights into *langue*”.

Como es bien sabido, la historia de la LC es todavía corta, pero se ha desarrollado a gran velocidad. En efecto, dejando a un lado los antecedentes (cf. Rojo 2015), su evolución está fuertemente

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.

determinada por la que han experimentado las computadoras y, en general, los recursos electrónicos de todo tipo vinculados a la forma de lograr versiones electrónicas de los textos, almacenarlos de modo que se pueda extraer la información pertinente, añadir información, abrirlos a la consulta general, etc. Todo ello hace que en cincuenta años se haya pasado del *Brown Corpus*, constituido por un millón de formas, con posibilidad de ofrecer solo resultados totales y consultable únicamente en la computadora en la que residía, a cualquiera de los corpus que manejamos en la actualidad, formados por cientos o incluso miles de millones de formas, previamente etiquetadas y lematizadas de forma automática, que admiten la recuperación selectiva de la información y pueden ser consultados, a través de internet, desde cualquier parte del mundo. En una línea paralela y congruente con la anterior, la tipología de los corpus se ha diversificado enormemente, en respuesta a las posibilidades que esta forma de acceder al análisis de los fenómenos lingüísticos brinda a cultivadores de diferentes subdisciplinas lingüísticas, con intereses muy variados y orientaciones diversas.

Por todo ello resulta imposible trazar un panorama relativamente detallado de los corpus existentes en español en un trabajo con las limitaciones de espacio esperables en una enciclopedia. Dado, además, que los lectores interesados pueden encontrar información detallada sobre recursos lingüísticos en la página mantenida por Joaquim Llisterri (http://liceu.uab.es/~joaquim/applied_linguistics/new_technologies/LengEsp_Materiales_WWW.html#recursos_linguisticos) y una amplia y actualizada relación de corpus de español en el magnífico informe elaborado por Briz y Albelda (2009), intentaré aquí proporcionar más bien una perspectiva general de las líneas maestras por las que ha discurrido hasta el presente la LC en español, citando en cada bloque únicamente los corpus más significativos.

2. LOS ORÍGENES

La LC en español comenzó de forma relativamente tardía, pero ha experimentado un desarrollo muy rápido e intenso que, hasta cierto punto, se explica precisamente por el hecho de haber iniciado su desarrollo en un marco tecnológico más evolucionado, lo cual permite entender también algunos factores característicos de la LC en español frente a lo habitual en otras tradiciones. Por una parte, el predominio de los corpus de referencia (CREA, CORDE, CE y CORPES, cf. infra), todos ellos de tamaño medio o grande, que permiten obtener la visión general de un determinado fenómeno no solo en un momento determinado, sino también a lo largo del tiempo o del espacio, lo cual resulta

especialmente importante en el caso del español, lengua habitual de la mayor parte de la población de un gran número de países (cf. Moreno y Otero 2007) y un grado de homogeneidad compatible con la existencia de una norma policéntrica. Por otra, esos corpus, concebidos y desarrollados ya en las condiciones de acceso que brinda la existencia de internet, están abiertos a la consulta externa desde su aparición, aunque, por las limitaciones que impone la existencia de derechos sobre los textos, la obtención de resultados está habitualmente limitada a los fragmentos que contienen el fenómeno que se desea analizar. En todo este desarrollo ha tenido una importancia crucial la decisión de acometer la construcción del CREA y, muy poco después, la del CORDE, adoptada por la Real Academia Española en 1995. De acuerdo con el papel atribuido tradicionalmente a las academias de la lengua en el mundo hispánico, su deseo de convertirse también en centros de recursos lingüísticos ha proporcionado instrumentos que, sin duda, han dado un giro considerable a las investigaciones sobre el español de todas las épocas, lugares y tipos.

Lo mismo que sucede en otras tradiciones, en los orígenes de la LC en español se entrecruzan líneas próximas a lo que luego será la LC, pero que no emplean recursos electrónicos, con las que emplean medios informáticos para automatizar tareas que, en principio, están menos vinculadas a la LC en sentido estricto. La primera de estas líneas es, sin duda, la representada por el *Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades de Iberoamérica y de la Península Ibérica*, propuesto inicialmente por Lope Blanch en el simposio de Bloomington (1964) (cf. Lope Blanch 1967, 1986; cf. también Spitzová 1991 y Rabanales 1992). Si lo comparamos con el que puede ser considerado su paralelo en la lingüística inglesa, el *Survey of English Usage* (SEU), dirigido por Randolph Quirk, vemos que ninguno de los dos fue concebido inicialmente como un corpus en el sentido actual del término, pero responden a objetivos muy próximos. El SEU, más reducido, caminaba en la línea de integrar los materiales (un millón de formas) en un conjunto único. El *Proyecto de la norma culta*, mucho más amplio en su diseño, se centraba en la publicación independiente de textos recogidos en diferentes ciudades. Esto es, carecía del rasgo de integración, fundamental en la construcción de un corpus, pero primaba, en cambio, el análisis de la variación lingüística. Como es de esperar por los rasgos señalados, el resultado de ambos proyectos fue reconvertido posteriormente en un corpus en el sentido habitual (aunque solo una pequeña parte en el caso del *Proyecto de la norma culta*, cf. infra).

Por una vía diferente se sitúan los trabajos que, en términos generales, suponen el aprovechamiento de las ventajas que proporcionan las recién nacidas computadoras para automatizar y

facilitar algunas tareas como la preparación de listas de frecuencias, índices o concordancias mediante la informatización de un conjunto más o menos amplio de textos. El *Hispanic Seminar of Medieval Studies* fue pionero, ya en la década de los 70, en la conversión a formato electrónico de textos medievales españoles para la redacción del *Dictionary of Old Spanish Language*, dirigido por Lloyd A. Karsten y John J. Nitti. Es fácil imaginar la enorme cantidad de dificultades que sus responsables tuvieron que vencer para recoger la complejidad y diversidad de las grafías medievales o el tratamiento del desarrollo de abreviaturas con los sistemas informáticos de la época. Con el paso de los años, todos los textos transcritos y procesados han pasado a formar parte de los corpus integrados en la *Biblioteca digital de textos del español antiguo* (y una parte de ellos también del CORDE). Un poco más tarde surgen dos proyectos desarrollados en la Universidad de Göteborg por David Mighetto y Per Rosengren. Primero nueve, luego once, novelas españolas de los años 1951 a 1971 (proyecto ONE71, con algo más de un millón de formas; cf. Mighetto 1985) y luego unos 3000 artículos publicados en el periódico *El País* y la revista *Triunfo* en 1977 (proyecto PE77, con casi dos millones de formas; cf. Mighetto y Rosengren 1982, 1983, 1985) a partir de los cuales editaron listas de frecuencias, concordancias y diccionarios inversos. Algún tiempo después, ambos conjuntos textuales pasaban a constituir sendos corpus integrados en el proyecto SOL (que, en la versión consultable en 2014, contiene también, bajo el nombre de COR92, el CORLEC preparado por Francisco Marcos Marín en 1992, cf. infra). Por fin, por esta misma época Hiroto Ueda informatizaba los textos de 30 obras teatrales españolas (cf. Ueda, 1987 y 1989-1997).

3. LOS PRIMEROS CORPUS DE ESPAÑOL

Entre 1980 y 1995 aparecen los primeros corpus de español concebidos y desarrollados ya con criterios comparables a los que en ese momento se estaban llevando a cabo para otras lenguas. Cabe distinguir, para esta época, cinco bloques fundamentales. En primer lugar, corpus de tamaño reducido (incluso para los estándares del momento), resultado de proyectos individuales o de grupos de investigación y que sirven fundamentalmente como fuentes de datos para análisis de diferente naturaleza. Destaca en este grupo el corpus de Lovaina, formado por 39 textos escritos españoles y americanos (entre 1922 y 1988) construido bajo la dirección de Josse de Kock en esa universidad. En realidad, son dos conjuntos textuales formados por 19 y 20 textos, respectivamente, de unas 100 000 formas cada uno de ellos. El contenido de ambos corpus fue publicado entre 1990 y 1992, complementados con los índices

alfabéticos, inversos y listas de frecuencias de cada uno de ellos (cf. De Kock et al. 1990-1992; De Kock 2001a). A pesar de su tamaño reducido, han sido la base de una importante cantidad de publicaciones, con centro en la serie *Gramática española. Enseñanza e investigación* publicada por la Universidad de Salamanca. De 1990 data también el corpus ENTREVIS90, construido por Kjær Jensen en la Universidad de Århus, constituido por un total de 725 000 formas procedentes de entrevistas publicadas en las revistas *Tiempo* y *Cambio 16* de 1990 (cf. Jensen 1991 y 2001). Poco tiempo después, Jensen construyó ENTREVIS95, con unas 569 000 palabras tomadas de entrevistas aparecidas en estas dos revistas a lo largo de 1995. A estos corpus podemos añadir el ya mencionado *Spanish on line* (SOL), disponible para su consulta a través de Internet desde 1998.

El segundo bloque está constituido por los corpus que se desarrollan en esta misma época para servir a propósitos lexicográficos, siguiendo la estela marcada por el COBUILD. El *Corpus Vox-Biblograf* (CVB), dirigido por Manuel Alvar Ezquerra, derivado del corpus de español diseñado en el interior del proyecto europeo NERC (cf. infra), constaba en 2001 de unos diez millones de formas procedentes de textos escritos y estaba prevista, pero todavía no finalizada, la inclusión de textos orales (cf. Alvar y Corpas 2001: 230-231). En una línea similar, el corpus CUMBRE, dirigido por Aquilino Sánchez, sirvió para la redacción del diccionario CUMBRE, editado por la editorial SGEL, presentado como el primer diccionario español basado en un corpus (GDUEsA: 7-8). Consta de unos veinte millones de formas, procedentes de textos escritos y orales de España y América. A partir de dos millones de formas de este corpus, etiquetadas y desambiguadas, se publicó un diccionario de frecuencias (Almela et al. 2001). También cabe destacar el *Corpus del español mexicano contemporáneo* (CEMC), formado por 996 muestras de unas 2000 formas procedentes de textos escritos y orales producidos entre 1921 y 1974 (cf. http://www.corpus.unam.mx:8080/cemc/). A partir de los datos de este corpus se publicaron varios diccionarios de español mexicano, dirigidos todos ellos por Luis Fernando Lara.

El tercer grupo está formado por corpus de tamaño pequeño que se construyen en el marco de proyectos europeos. El proyecto *Corpus Resources and Terminology Extraction* (CRATER) dio lugar a un corpus multilingüe (inglés, francés y español) constituido por textos alineados destinados a facilitar la extracción y traducción de términos técnicos. El proyecto *Network of European Reference Corpus* (NERC) tenía como finalidad fundamental la de establecer los rasgos básicos para la construcción de corpus de características semejantes en diversas lenguas europeas. El proyecto PAROLE pretendía la creación de corpus de construcción similar de diversas lenguas europeas con un tamaño aproximado de

veinte millones.

El cuarto bloque está formado por varios corpus de carácter general y tamaño relativamente pequeño. En primer lugar, los dirigidos por Francisco Marcos Marín en diversas acciones patrocinadas por la Sociedad Estatal del Quinto Centenario. Son el *Corpus lingüístico de referencia de la lengua española en Argentina*, el *Corpus lingüístico de referencia de la lengua española en Chile*, cada uno de ellos con alrededor de dos millones de formas y el CORLEC (cf. infra). El corpus LEXESP, resultado de la colaboración de diversos equipos de lingüistas y psicólogos, tiene alrededor de 1,5 millones de formas de textos escritos sobre las que se elaboró el *Diccionario de frecuencias de las unidades lingüísticas del castellano* (Alameda y Cuetos 1995). Hay que mencionar aquí también el *Corpus of Contemporary Spanish* reunido por Barry Ife y constituido por algo más de 5 millones de formas, distribuido en CD-ROM.

Por fin, en la dimensión diacrónica, Francisco Marcos Marín, Charles Faulhaber, Ángel Gómez Moreno y Antonio Cortijo Ocaña son los responsables del proyecto *ADMYTE*, que reúne las transcripciones de una notable cantidad de textos medievales españoles. Publicado inicialmente en CD (Admyte 0 en 1991 y Admyte 1 en 1992), los textos de este corpus (unos 12 millones de formas en total) son consultables ahora en red.

4. LA SITUACIÓN ACTUAL

Como ya se ha señalado, la Real Academia Española tomó en 1995 la decisión de emprender la construcción de un corpus del español contemporáneo, el *Corpus de referencia del español actual* (CREA), y, pocos meses después, de su complemento histórico, el *Corpus diacrónico del español* (CORDE). En las versiones que se pueden considerar cerradas a finales de 2013, el CREA contiene unos 160 millones de formas procedentes de textos, tanto escritos como orales, de todos los países hispánicos, publicados entre 1975 y 2004. El CORDE, por su parte, está formado por algo más de 250 millones de formas que proceden de textos que van desde los orígenes de la lengua hasta 1974. Ninguno de los dos corpus está anotado en sus versiones públicas, pero la aplicación de consulta presenta una gran versatilidad, que permite hacer recuperaciones selectivas mediante la combinación de diferentes valores de cualquiera de los muchos parámetros que han entrado en la configuración de los corpus. La necesidad de mejorar y ampliar sus bancos de datos ha llevado a la RAE a construir el *Corpus del diccionario histórico* (CDH), pensado fundamentalmente para ser el fondo documental del

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.

Nuevo diccionario histórico del español y también el *Corpus del español del siglo XXI* (CORPES), que en su primera fase comprenderá trescientos millones de formas procedentes de textos escritos y orales producidos entre 2001 y 2012. La versión provisional consultable en mayo de 2015 (la 0.8) consta de unos doscientos siete millones de formas, todas ellas de textos escritos anotados y lematizados.

En la categoría de los corpus de referencia entra también el *Corpus del español* (CE), construido por Mark Davies, formado por unos cien millones de formas desde los orígenes de la lengua hasta finales del siglo XX. El CE está parcialmente anotado y lematizado, lo cual le confiere su mayor valor. La aplicación de recuperación de información es muy rápida y vistosa, pero solo admite la selección por siglos y, en los del XX, también por tipos de texto.

Carácter general tiene también el *Corpus del español actual* (CEA), construido por Carlos Subirats y Marc Ortega, integrado por unos 450 millones de formas procedentes de textos parlamentarios, documentos oficiales de la ONU y de la parte española de la *Wikipedia*, anotados automáticamente. El *Corpus dinámico del castellano de Chile* (CODICACH), desarrollado por Scott Sadowsky, consta de ochocientos millones de formas ortográficas, casi todas procedentes de textos de prensa (pero solo hay acceso abierto a las listas de frecuencias léxicas derivadas). Mucho más reducido en tamaño, aunque con una gran cantidad de información gramatical y léxica asociada, es el corpus *Ancora-Es*, del grupo CLIC de la Universidad de Barcelona, dirigido por M.ª Antònia Martí.

Como consecuencia de la prioridad atribuida en el *Proyecto de la norma culta* al análisis de la lengua oral y su temprano cruce con la sociolingüística de orientación laboviana, surgió muy pronto una notable cantidad de proyectos de recogida de materiales orales en un gran número de ciudades a uno y otro lado del Atlántico (cf. Bentivoglio 1991, 1998: 40; Briz y Albelda 2009, § 2), pero solo una parte ha sido publicada en formato impreso y una parte todavía menor ha sido digitalizada. Afortunadamente, la *Asociación de lingüística y filología de América latina* (ALFAL) acometió el proyecto de digitalizar y dotar de una codificación unificada una parte de los materiales recogidos (12 ciudades con 14 encuestas de cada una de ellas, cf. Samper 1995) en el denominado *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico* (Samper et al. 1998), distribuido en CD e incorporado íntegramente al componente oral del CREA. Algunos de esos materiales, a los que se añadieron más entrevistas grabadas en algunas ciudades americanas, fueron utilizados en el proyecto *Estudio gramatical del español hablado en América* (EGREHA), dirigido por César Hernández Alonso (2009). La unión de estas dos tradiciones y su adaptación a las características actuales de la investigación en este terreno se da en el *Proyecto de estudio sociolingüístico del español*

de España y de América (PRESEEA), coordinado por Francisco Moreno Fernández. En julio de 2014 forman parte de este proyecto 39 equipos, que han recogido y transcrito entrevistas semidirigidas en otras tantas ciudades del mundo hispánico. Los textos tienen una codificación común y se puede consultar ya una versión provisional que permite la recuperación selectiva de algunos de los subcorpus. Una parte considerable de esos materiales pasará a formar parte del CORPES.

A los anteriores, de orientación netamente sociolingüística, hemos de añadir los corpus de lengua oral construidos con otras finalidades de estudio, compatibles en muchos casos en el enfoque sociolingüístico. Con carácter de corpus de referencia podemos citar el *Corpus oral de referencia de la lengua española contemporánea* (CORLEC), que contiene la transcripción de 1 100 000 formas grabadas en 1990-1992 y que ha sido integrado también en el CREA. C-ORAL-ROM, distribuido en CD, es el resultado de un proyecto europeo en el que se han transcrito unas 300 000 formas de cada una de las cuatro lenguas participantes, el español entre ellas. El *Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante* (ALCORE y COVJA), dirigido por Dolores Azorín y el *Vernáculo Urbano Malagueño* (VUM), dirigido por Villena Ponsoda están también en este bloque. Corpus consistentes en conversaciones son el *Corpus de conversaciones coloquiales* reunido por el grupo Val.Es.Co, que en su versión 2.0. (julio de 2013) muestra 46 conversaciones con un total de 120 000 palabras. El grupo, dirigido por Antonio Briz, ha desarrollado un sistema de codificación que ha sido adoptado por otros muchos proyectos. Hay que citar también el *Corpus conversacional de Alcalá*, construido por Ana Cestero, integrado en el CREA, y el *Corpus conversacional de Barcelona y su área metropolitana*, proyecto dirigido por Rosa Vila. Sobre lengua oral, pero con propósito de utilización para fines didácticos (ELE), el *Spanish Oral Language Archive Project*, construido en la Universidad Carnegie Mellon, y el *Corpus oral didáctico anotado lingüísticamente* (C-Or-DiAL), construido por Carlota Nicolás Martínez (2012), vinculado a C-ORAL-ROM; se distribuye en un CD que contiene las grabaciones y las transcripciones, anotadas, con un total de casi 120 000 formas.

El proyecto *Difusión internacional del español por los medios* (DIES-M), coordinado por Raúl Ávila, que integra los antiguos DIES-RTV y DIES-RTP, reúne y transcribe textos procedentes de los más diversos medios de comunicación (prensa escrita, radio y televisión). Con materiales de habla juvenil figuran el COVJA, ya mencionado, el *Corpus de habla de los universitarios salmantinos* (CHUS), dirigido por Julio Borrego y Carmen Fernández Juncal, y el *Corpus oral de lenguaje adolescente* (COLA), construido por Annete Myre Jörgensen, de la Universidad de Bergen, con transcripciones y sonido alineado procedentes de muestras tomadas en Madrid, Santiago de Chile,

Buenos Aires y Ciudad de Guatemala del habla de jóvenes con edades comprendidas entre los 13 y los 19 años. Por fin, orientado al español rural y al estudio de fenómenos gramaticales, el *Corpus oral y sonoro del español rural* (COSER), dirigido por Inés Fernández-Ordóñez en la Universidad Autónoma de Madrid.

El Grial, dirigido por Giovanni Parodi en la Pontificia Universidad Católica de Valparaíso, se centra en el análisis de las diferencias entre textos correspondientes a diferentes registros del español. Sobre el español técnico de diferentes países, orientados casi siempre a la obtención de terminología especializada, hay que mencionar en primer lugar el *Corpus técnico del Institut universitari de lingüística aplicada* de la Universidad Pompeu Fabra, con textos escritos (en español, catalán, inglés, francés y alemán) correspondientes a los ámbitos del derecho, la economía, la genómica, la medicina y el medio ambiente, a los que se añade un corpus de prensa como elemento de contraste. El corpus *Iberia*, realizado en el Consejo superior de investigaciones científicas y dirigido por Ignacio Ahumada, contiene cerca de 90 millones de formas procedentes de textos técnicos variados publicados entre 1985 y 2011. Hay que mencionar en este apartado las dos secciones españolas del proyecto *Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies* (PANACEA), dedicadas a medio ambiente y legislación laboral. El corpus PAAU92, diseñado y construido por Paz Battaner y Sergi Torner recoge las pruebas escritas de 700 exámenes de diversas materias en las pruebas de acceso de 1992 en seis universidades españolas (Barcelona, Complutense de Madrid, Murcia, Oviedo, Salamanca y Sevilla). Ha sido incorporado al CREA.

Se están desarrollando también algunos corpus formados con textos producidos por estudiantes de español como lengua extranjera. El *Spanish Learner Language Oral Corpora* es un corpus oral de español L2 compilado por investigadores de las universidades británicas de Southampton, Newcastle y Greenwich. El *Corpus escrito de español L2* (CEDEL2), dirigido por Cristóbal Lozano en la Universidad de Granada, forma parte de un proyecto centrado en el análisis del orden de palabras en español e inglés. Por fin, el Instituto Cervantes está (en julio de 2014) a punto de publicar el *Corpus de aprendices de español* (CAES), formado por muestras de más de 1400 estudiantes de ELE con árabe, chino mandarín, francés, inglés, portugués y ruso como L1.

Entre los corpus de orientación diacrónica, además de la *Biblioteca Digital de Textos del Español Antiguo* y el CORDE, ya citados, hay que mencionar el *Corpus histórico del español de México* (CHEM), el *Corpus de documentos españoles anteriores a 1700* (CODEA), dirigido por Pedro

Sánchez-Prieto, con unos 1500 documentos transcritos hasta el momento según las directrices seguidas en el proyecto *Corpus hispánico y americano en la red: textos* antiguos (CHARTA). Carácter más restringido por el tipo de textos incluidos (solo traducciones de la Biblia al castellano), pero mucho más amplio en cuanto a posibilidades de recuperación de datos y análisis tiene el proyecto *Biblia medieval*, dirigido por Andrés Enrique-Arias, con unos cinco millones de formas. Para finales de 2015 está prevista la publicación del *Corpus diacrónico y diatópico del español de América* (CORDIAM), patrocinado por la Academia mexicana de la lengua y dirigido por Concepción Company y Virginia Bertolotti. Contendrá la transcripción de unos 3000 documentos, no literarios, procedentes de todos los países americanos, con un total de unos cuatro millones de formas.

Para el análisis del lenguaje infantil cabe citar, en primer lugar, la parte española del conocido proyecto *Child Language Data Exchange System* (CHILDES). A ello podemos sumar el *Corpus de habla infantil espontánea del español* (CHIEDE), con algo menos de 60 000 formas procedentes de 30 grabaciones, que se distribuye a través de ELRA.

Hay gran cantidad de textos en español en corpus multilingües, entre los que cabe destacar el *Catalan-Spanish Parallel Corpus* (ELRA-W0053), con unos cien millones de formas procedentes de la edición en ambas lenguas de *El periódico de Cataluña*; el *MultiUN: Multilingual UN Parallel Text 2000—2009*, con trescientos cincuenta millones de formas procedentes de textos oficiales de la ONU; el *European Parliament Proceedings Parallel Corpus 1996-2011* y el *Wikicorpus, v. 1.0*, con elementos procedentes de las versiones en español, catalán e inglés de la Wikipedia, anotados con FreeLing.

Al lado de los corpus constituidos por la transcripción de textos orales (entrevistas, conversaciones, tertulias, discursos, etc.) hay que situar los integrados por grabaciones, destinados a proporcionar los datos necesarios para las aplicaciones de análisis y síntesis de voz. Además de la gran cantidad de grabaciones de los más diversos tipos que figuran entre los materiales ofrecidos por ELRA, hay que mencionar *Albayzin*, *EUROM*, *SpeechDat*, *ACCOR*, *MULTEXT* o *MATE* (cf. la amplia información sobre estos corpus en la parte correspondiente de la página *http://liceu.uab.es/~joaquim/language_resources/spoken_res/Corp_oral_esp.html*, mantenida por Joaquim Llisterri.

La corriente conocida como *Web as Corpus* ha creado, como es lógico, conjuntos con textos escritos en español. El carácter automático y escasamente selectivo con que se han venido construyendo estos recursos hace que, en principio, deban ser considerados más bien como archivos (cf. supra). Sin embargo, la selección y el filtrado de textos se ha ido haciendo de forma cada vez más

refinada. Hay que mencionar en este grupo la parte española de los *Internet corpora* construidos en la Universidad de Leeds, con unos 143 millones de formas y, sobre todo, el corpus *EsTenTen*, que forma parte de un proyecto mucho más general, dirigido por Adam Kilgarrif (2013) y que tenía, en diciembre de 2013, algo más de 8300 millones de formas, etiquetadas, procedentes de todos los países hispánicos.

Los corpus son recursos básicos que, además de la explotación directa, constituyen el punto de partida de numerosas aplicaciones mediante las adaptaciones y refinamientos necesarios. Los *tree-banks*, por ejemplo, son conjuntos de secuencias analizadas sintácticamente en los que es posible recuperar las que reúnen unas determinadas condiciones estructurales. Hay que mencionar en este bloque el *RST Spanish Treebank*, desarrollado en la Universidad Autónoma de México, con anotación de relaciones discursivas; el *Spanish Tree-Bank* de la Universidad Autónoma de Madrid, con 1500 oraciones anotadas sintácticamente o el *LSP Spanish Treebank* desarrollado en el IULA de la Universidad Pompeu Fabra, con unas 42 000 oraciones analizadas, procedentes del corpus técnico (vid. supra). El corpus *Araknion-es*, construido por el grupo CLIC (cf. supra), contiene algo más de tres millones de oraciones con los análisis sintácticos correspondientes en forma de árboles de dependencias. Con un enfoque bastante diferente, la *Base de datos sintácticos del español actual* (BDS), construida en la Universidad de Santiago de Compostela, almacena el resultado del análisis sintáctico manual de las aproximadamente 160 000 cláusulas contenidas en un corpus de casi un millón y medio de formas del español actual. A partir de la BDS se han desarrollado recursos como el *Corpus sintácticamente anotado* (CSA), dentro del proyecto de *Desarrollo de recursos para el análisis sintáctico del español* (DRASAE), dirigido por M.ª Paula Santalla del Río en la Universidad de Santiago de Compostela, o la base de datos sobre *Alternancias de diátesis y esquemas sintáctico-semántico del español* (ADESSE) dirigida por José María García-Miguel en la Universidad de Vigo.

Relación de corpus y otros recursos electrónicos mencionados en el texto

ACCOR: *Articulatory-Acoustic Correlations in Coarticulatory Processes - A Cross-Linguistic Investigation* (http://www.cstr.ed.ac.uk/research/projects/artic/accor.html).

ACUAH : *Análisis de la conversación. Universidad de Alcalá de Henares.*

ADESSE: *Alternancias de diátesis y esquemas sintáctico-semantícos del español* (http://adesse.uvigo.es/)

ADMYTE (http://www.admyte.com/admyteonline/home.htm).

Albayzín. Base de datos para el reconocimiento del habla en español (http://catalog.elra.info/product_info.php?

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.

products_id=746&osCsid=7a272af9a54b96add9f69ac305a7ed28).

ALCORE: *Alicante corpus oral del español*; integrado en el *Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante*.

AnCora-ES (http://clic.ub.edu/corpus/ancora).

Araknion-es (http://clic.ub.edu/corpus/araknion-es).

BDS: *Base de datos sintácticos del español actual* (http://www.bds.usc.es).

Biblia medieval (http://www.bibliamedieval.es/index.php).

Biblioteca Digital de Textos del Español Antiguo (http://www.hispanicseminary.org/textconc-es.htm).

Brown Corpus: *The Standard Corpus of Present-Day Edited American English* (www.helsinki.fi/varieng/CoRD/corpora/BROWN/).

C-Or-DiAL: *Corpus oral didáctico anotado lingüísticamente* (http://lablita.dit.unifi.it/corpora/cordial).

C-ORAL-ROM: *Integrated reference corpora for spoken romance languages* (http://lablita.dit.unifi.it/coralrom/index.html; http://www.lllf.uam.es/ESP/Coralrom.html).

CAES: *Corpus de aprendices de español L2* (www.cervantes.es/lengua_y_ensenanza/informacion.htm).

*Catalan-Spanish Parallel Corpus (*http://catalog.elra.info/product_info.php?products_id=1122).

CDH: *Corpus del nuevo diccionario histórico* (http://www.rae.es/recursos/banco-de-datos/cdh).

CE: *Corpus del español* (www.corpusdelespanol.org/).

CEA: *Corpus del español actual* (http://sfn.uab.es:8080/SFN/tools/cea/spanish).

CEDEL2 : *Corpus Escrito de Español L2* (http://www.uam.es/proyectosinv/woslac/cedel2.htm).

CEMC: *Corpus del español mexicano contemporáneo* (http://www.corpus.unam.mx:8080/cemc/).

CHARTA: *Corpus hispánico y americano en la red: textos antiguos* (http://www.charta.es).

CHEM: *Corpus histórico del español en México* (http://www.iling.unam.mx/chem/).

CHIEDE: *Corpus de Habla Infantil Espontánea del Español* (http://catalog.elra.info/product_info.php?products_id=1090).

CHILDES: *Child Language Data Exchange System* (http://childes.psy.cmu.edu/).

CHUS: *Corpus de habla de los universitarios salmantinos* (cf. Briz y Albelda 2009).

CODEA: *Corpus de documentos españoles anteriores a 1700* (http://demos.bitext.com/codea/).

CODICACH: *Corpus dinámico del castellano de Chile* (http://sadowsky.cl/codicach-es.html).

COLA: *Corpus oral del lenguaje adolescente* (www.colam.org/om_prosj-espannol.html).

CORDE: *Corpus diacrónico del español* (http://rae.es/recursos/banco-de-datos/corde).

CORLEC: *Corpus oral de referencia de la lengua española contemporánea* (http://www.lllf.uam.es/ESP/Corlec.html).

CORPES: *Corpus del español del siglo XXI* (http://rae.es/recursos/banco-de-datos/corpes-xxi).

Corpus conversacional de Barcelona y su área metropolitana (cf. Vila Pujol 2001).

Corpus de Lovaina: cf. De Kock et al. (1990-1992).

Corpus lingüístico de referencia de la lengua española en Argentina (http://www.lllf.uam.es/ESP/Argentina.html).

Corpus lingüístico de referencia de la lengua española en Chile (http://www.lllf.uam.es/ESP/Chile.html).

Corpus of Contemporary Spanish (http://www.kcl.ac.uk/artshums/depts/ddh/research/projects/completed/ccs.aspx).

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.

Corpus orales para la fonética y las tecnologías del habla en español. Página mantenida por Joaquim Llisterri (*http://liceu.uab.es/~joaquim/language_resources/spoken_res/Corp_oral_esp.html*).

Corpus técnico del Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra (IULA) (http://www.iula.upf.edu/recurs01ca.htm).

COSER: *Corpus Oral y Sonoro del Español Rural* (http://www.lllf.uam.es:8888/coser/).

COVJA: *Corpus oral de la variedad juvenil universitaria del español de Alicante*; integrado en el Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante)

CRATER: *Corpus Resources and Terminology Extraction* (http://ucrel.lancs.ac.uk/projects.html#crater).

CREA: *Corpus de referencia del español actual* (http://rae.es/recursos/banco-de-datos/crea).

CSA: *Corpus sintácticamente anotado* (vid. DRASAE).

CVB: *Corpus Vox-Bibliograf* (cf. Alvar y Corpas, 2001).

DIES-M: *Difusión internacional del español en los medios* (www.colmex.mx/academicos/cell/ravila/index_archivos/page0003.htm).

DIES-RTP: vid. DIES-M.

DIES-RTV: vid. DIES-M.

DRASAE: *Desarrollo de recursos para el análisis sintáctico automático del español* (http://gramatica.usc.es/proxectos/drasae/).

EGREHA *Estudio gramatical del español hablado en América* (cf. Hernández Alonso 2009).

ELRA: *European Language Resources Association* (http://www.elra.info/).

ENTREVIS90 (cf. Jensen 1991 y 2001).

ENTREVIS95 (cf. Jensen 2001).

Es-Ten-Ten (http://www.sketchengine.co.uk/documentation/wiki/Corpora/esTenTen).

EUROM1: *Multilingual Speech Corpus*: (http://catalog.elra.info/product_info.php?products_id=528&osCsid=e682925cbc0378057a1cb911c485ad67).

European Parliament Proceedings Parallel Corpus 1996-2011 (http://www.statmt.org/europarl/).

FreeLing: *An Open Source Suite of Language Analyzers* (http://nlp.lsi.upc.edu/freeling/).

GRIAL (www.elv.cl/prontus_linguistica/site/edic/base/port/grial.html).

IBERIA: *Corpus de español científico* (www.investigacion.cchs.csic.es/elci/node/8).

IULA Spanish Treebank (http://www.iula.upf.edu/recurs01_tbk_uk.htm).

Leeds collection of Internet Corpora (corpus.leeds.ac.uk/internet.html).

LEXESP (http://www.psico.uniovi.es/Dpto_Psicologia/metodos/soft/corpus/).

Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico (cf. Samper et al. 1998).

Lingüística y lengua españolas. Recursos en internet. Página mantenida por Joaquim Llisterri (http://liceu.uab.es/~joaquim/applied_linguistics/new_technologies/LengEsp_Materiales_WWW.html#recursos_linguisticos).

MATE: *Multilevel Annotation, Tools Engineering (*http://xml.coverpages.org/mate.html).

MULTEXT: *Multilingual Text Tools and Corpora* (http://aune.lpl.univ-aix.fr/projects/multext/index.html).

MultiUN: *Multilingual UN Parallel Text 2000—2009* (http://www.euromatrixplus.net/multi-un/).

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.

NERC*: Network of European Reference Corpora.*

OLA: *Spanish Oral Language Archive Project* (http://ml.hss.cmu.edu/mlrc/ola/spanish/index.html).

PAAU 1992 : *Pruebas de acceso a la universidad* (http://www.iula.upf.edu/rec/corpus92/).

PANACEA Environment Spanish monolingual corpus (http://catalog.elra.info/product_info.php?products_id=1192&language=en)..

PANACEA Labour Spanish monolingual corpus (http://catalog.elra.info/product_info.php?products_id=1193).

PRESEEA: *Proyecto para el estudio sociolingüístico del español de España y de América* (preseea.linguas.net/).

RST Spanish Treebank (http://www.corpus.unam.mx/rst/index_es.html).

SEU: *Survey of English Usage* (http://www.ucl.ac.uk/english-usage/index.htm).

SOL: *Spanish On Line* (http://spraakbanken.gu.se/konk/rom2/).

Spanish Learner Language Oral Corpora (http://www.splloc.soton.ac.uk/index.html).

SPEECHDAT: *Spoken Language Resources* (http://catalog.elra.info/product_info.php?products_id=721&osCsid=9289223575b55f27c187a5a97951476a; http://catalog.elra.info/product_info.php?products_id=722&osCsid=6b68023ac61990e6b690d6f0f41fa9c9).

UAM Spanish Treebank (http://www.lllf.uam.es/ESP/Treebank.html).

Val.Es.Co: *Corpus de conversaciones coloquiales* (http://www.valesco.es/).

VUM: *Vernáculo oral malagueño*.

Wikicorpus, v. 1.0: Catalan, Spanish and English portions of the Wikipedia (http://www.lsi.upc.edu/~nlp/wikicorpus/).

Referencias bibliográficas

Alameda, J. R. y Cuetos, F. (1995). *Diccionario de frecuencias de las unidades lingüísticas del castellano.* Oviedo: Universidad de Oviedo.

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R. y Almela, M. (2005): *Frecuencias del español : diccionarios y estudios léxicos y morfológicos.* Madrid: Universitas.

Alvar Ezquerra, M. y Corpas Pastor, G. (2001): "Usos y valores de *para nada* en un corpus de español peninsular actual", en De Kock (2001b), 229-243.

Bentivoglio, P. (1991): " El estudio de la lengua hablada en Venezuela", en *Abralin. Boletim da Associacão brasileira de lingüística*, 12, 61-74.

Bentivoglio, P. (1998): "La variación sociofonológica", *Español actual*, 69, 29-42.

Briz, A. y Albelda, M. (2009). "Estado actual de los corpus de lengua española hablada y escrita: I+D". En *El español en el mundo. Anuario del Instituto Cervantes 2009*. Madrid: Instituto Cervantes, 165-226.

De Kock, J. (2001a): "Un corpus informatizado para la enseñanza de la lengua española. Punto de partida y término", en *Hispanica Polonorum*, 3, 60-86

De Kock, J. (ed.) (2001b): *Lingüística con corpus. Catorce aplicaciones sobre el español*. Universidad de Salamanca. (= *Gramática española. Enseñanza e investigación*. I.7)

De Kock, J. et al. (1990-1992): *Gramática española. Enseñanza e investigación.* Salamanca: Universidad de Salamanca. [Tomo III.I: De Kock, Verdonk, R., Gómez Molina, C.: *19 textos*, 1991 (reimp. 1996); tomo III.2: De Kock, J. Gómez Molina, C. y Delbecque, N.: *20 textos*, 1992; tomo IV.1: De Kock, J.: *Índice alfabético, alfabético inverso y de frecuencia de 19 textos*, 1991; tomo IV.2: De Kock, J.: *Índice alfabético, alfabético inverso y de frecuencia de 20 textos*, 1992; tomo V.I. De Kock, J.: *Concordancia alfabética de 19 textos*, 1990 (solo consultable en forma de listado); tomo V.II. De Kock, J.: *Concordancia alfabética de 20 textos*, 1990 (consultable solo en forma de listado).]

GDUEsA: *Gran diccionario de uso del español actual*. Dir. por A. Sánchez. Madrid: SGEL, 2001.

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.

Hernández Alonso, C. (ed.) (2009): *Estudios lingüísticos del español hablado en América*. Madrid : Visor Libros, 4 vols..
Jensen, K. (1991): "ENTREVIS – a Spanish machine-readable text corpus". *Hermes, Journal of Linguistics*, 7, 81-85.
Jensen, K. (2001): "El verbo *caer*: Estudio semántico-sintáctico", en De Kock (2001b), 245-254.
Kilgarrif, A. y Renau, I. (2013): "*EsTenTen*, a Vast Web Corpus of Peninsular and American Spanish". En *Procedia - Social and Behavioral Sciences*, 95, 12–19. Available online at www.sciencedirect.com.
Lope Blanch, J. M. (1967): "Proyecto de estudio del habla culta de las principales ciudades de Hispanoamérica", en *El simposio de Bloomington. Agosto de 1964. Actas, informes y comunicaciones*. Bogotá: Instituto Caro y Cuervo, 255-264.
Lope Blanch, J. M. (1986). *El estudio del español hablado culto. Historia de un proyecto.* México: UNAM.
Mighetto, D. (1985): *ONE71. Banco de datos de once novelas españolas 1951-1971*. Univ. de Göteborg.
Mighetto, D. & Rosengren, P. (1982): *Banco de datos de Prensa española 1977. Concordancia lingüística y texto fuente*. Univ. de Göteborg.
Mighetto, D. & Rosengren, P. (1983): *PE77. Palabras gráficas españolas: Lista y frecuencias en Prensa Española 77*. Univ. de Göteborg, 4 vols.
Mighetto, D. & Rosengren, P. (1985): *Diccionario reverso DR Reverse Dictionary*. Univ. de Göteborg.
Moreno Fernández, F. (2001): "El corpus *ACUAH*: Análisis de los clíticos pleonásticos". En De Kock (2001b), 353-369.
Moreno Fernández, F. y Otero Roth, J. (2007): *Atlas de la lengua española en el mundo*. 2.ª ed., Madrid / Barcelona: Fundación Telefónica / Ariel.
Nicolás Martínez, C. (2012): *Corpus C-Or-DiAL (Corpus oral didáctico anotado lingüísticamente)*. Madrid: Liceus.
Rabanales, A. (1992): "Fundamentos teóricos y pragmáticos del *Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades del mundo hispánico*", en *BFUCh*, 33, 251-272.
Rojo, G. (2014): "Hispanic corpus linguistics", en Lacorte, Manel (ed.): *The Routledge Handbook of Hispanic Applied Linguistics*. Nueva York: Routledge, 371-387.
Rojo, G. (2015): "Sobre los antecedentes de la lingüística de corpus", en Álvarez Menéndez, A. *et alii*: *Studium grammaticae. Homenaje al Profesor José Antonio Martínez*. Universidad de Oviedo, 675-689.
Samper Padilla, J. A. (1995). "Macrocorpus de la norma lingüística culta de las principales ciudades de España y América". *Lingüística* 7, 263-293.
Samper Padilla, J. A., Hernández Cabrera, C. E. y Troya Déniz, M. (eds.) (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*. Las Palmas: Universidad de Las Palmas de Gran Canaria.
Spitzová, E. (1991): "Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península ibérica: proyecto y realización", en *Studia minora facultatis philosophicae Universitatis Brunensis*, L 12, 61-66.
Tognini Bonelli, E. (2010). "Theoretical overview of the evolution of corpus linguistics". En A. O'Keefe and M. McCarthy (eds.): *The Routledge Handbook of Corpus Linguistics*, 14-27. Oxon: Routledge.
Ueda, H. (1987): *Análisis lingüístico de obras teatrales españoles. Textos e índice de palabras*. Universidad Nacional de Estudios Extranjeros de Tokio.
Ueda, H. (1989-1997): *Análisis lingüístico de obras teatrales españolas. Concordancias*. Universidad de Tokio, V-1-12.
Vila Pujol, M.ª Rosa (2001): *Corpus del español conversacional de Barcelona y su área metropolitana*. Barcelona, Univ. de Barcelona.

PALABRAS CLAVE
Corpus textuales, lingüística de corpus.

ENTRADAS RELACIONADAS
Lingüística de corpus.

LECTURAS COMPLEMENTARIAS
Briz, Antonio y Albelda, Marta (2009).
Lavid, Julia: *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI.* Madrid: Cátedra, 2005.
Parodi, Giovanni (ed.): *Working with Spanish Corpora*. London/New York: Continuum, 2007.
Sinclair, John: *Corpus, Concordance, Collocation.* Oxford, Oxford University Press, 1991.

Borrador final. Pendiente de aparición en Gutiérrez-Rexach, Javier (ed.): *Enciclopedia lingüística hispánica*. Oxon: Routledge, 2016, pp. 285-296.