Sobre los antecedentes de la lingüística de corpus

Guillermo Rojo
Universidade de Santiago de Compostela

Dada la evidente dependencia que existe entre la lingüística de corpus (LC) y las computadoras, parece lógico pensar que la historia de la LC en sentido estricto no puede remontarse más allá del momento en que comienza el proceso de diseño, construcción y utilización de las primeras máquinas a las que podemos considerar realmente computadoras. Quizá la aparente nitidez de ese corte explica que no se haya prestado demasiada atención al análisis de las primeras fases de la LC ni a sus antecedentes, lo cual ha llevado a presentaciones bastante deficientes e incompletas de un tema que, sin embargo, posee un interés considerable para la historia de la lingüística. En realidad, es una cuestión bastante compleja, constituida por varias facetas relativamente independientes. En primer lugar, los antecedentes lejanos de los corpus y la LC. Más cerca está la cuestión del vínculo entre los corpus presentes en la teoría y la práctica distribucionalistas y los que aparecen a partir de 1963. En tercer término, la historia de los primeros veinte años de la LC, con el contraste entre la marginación que sufre en los Estados Unidos y el desarrollo que en esos mismos años experimenta en Europa. Y por último, el análisis de los antecedentes en tradiciones lingüísticas diferentes de la inglesa, que es la única a la que se ha atendido habitualmente[1]. Me propongo en este trabajo contribuir al

[1] Para una interesante revisión de los desajustes que la historia «oficial» de la LC presenta en algunos de estos aspectos, *cfr*. Jacqueline Léon, «Claimed and Unclaimed Sources of Corpus Linguistics», *Henry Sweet Society Bulletin,* 44 (2005), págs. 36-50.

mejor conocimiento del primero de estos aspectos, añadiendo, siempre que sea pertinente, datos procedentes de la tradición lingüística española.

El estudio de los antecedentes de la LC parte, sin duda, del artículo que W. Nelson Francis publicó en 1992 con el provocador título «Language corpora B.C.» (es decir, «before computer[s]»)[2]. Si se tiene en cuenta que Francis y Kučera son los autores del *Brown Corpus*, que es el que se señala generalmente como el primer corpus electrónico, se comprenderá la importancia que hay que atribuir a esta especie de revisión de precursores del trabajo que se considera fundacional de esta aproximación. Francis orienta su contribución en la línea de demostrar, frente a la idea que supone generalizada en aquel momento, que «many important corpora of English were assembled long before the computer was invented»[3]. A la de haber sido construidos con textos en inglés, Francis añade una segunda restricción, que deriva directamente de su definición de corpus[4]: haber sido elaborados para su empleo en el análisis lingüístico. Como consecuencia de ello, no entra en la consideración de conjuntos como el *Oxford Book of English Verse*, el *Corpus Iuris Civilis* o el *Corpus Glossary*, todos ellos mencionados explícitamente. En aplicación de las restricciones señaladas, se centra en tres líneas tradicionales: las colecciones de datos construidas en lexicografía, dialectología y gramática. En la primera de ellas menciona los grandes proyectos lexicográficos del inglés: el diccionario de Johnson, el OED y el Merriam-Webster. Dedica luego atención a aquellos estudios dialectológicos que, como el de Ellis, realizaron un enorme esfuerzo para la recogida de datos, mediante corresponsales más o menos expertos en muchos de los casos. Este tipo de reunión de materiales técnicos termina, siempre según Francis, en los atlas lingüísticos. Finalmente, en lo que respecta a antecedentes de corpus en estudios gramaticales, cita muy de pasada a autores como Jespersen, Kruisinga

---

[2] W. Nelson Francis, «Language corpora B. C.», en Svartvik, Jan (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82*, Berlín (Mouton - de Gruyter), 1992 [= Trends in Linguistics. Studies and Monographs, 65], págs. 17-31.

[3] Francis, «Language corpora B. C.», pág. 17.

[4] Un corpus es «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, «Language corpora», *ibidem).* Retoma aquí la definición utilizada en W. Nelson Francis, «Problems of assembling and computerizing large corpora», en Johansson, Stig (ed.), *Computer corpora in English language research*, Bergen (Norwegian Computing Centre for the Humanities), 1982, págs. 7-24. La cita, pág. 7.

o Poutsma para centrarse finalmente en el *Survey of English Usage*, dirigido por Randolph Quirk. Una perspectiva muy semejante es la que proporciona Svartvik, que menciona las mismas obras, añade una referencia al trabajo desarrollado por Fries para la preparación de *The Structure of English* y se centra luego en la relación establecida entre Francis y el equipo que trabajaba con Quirk en la construcción del SEU, al que pertenecía el propio Svartvik[5].

Para valorar adecuadamente la orientación adoptada por Francis y seguida por Svartvik hay que tener en cuenta que, de acuerdo con sus propias palabras, su objetivo es mostrar la existencia de corpus ingleses construidos para el análisis lingüístico antes de la introducción de las computadoras. No se trata, por tanto, de buscar antecedentes de la LC, ni siquiera de la LC en inglés. Por tanto, ausencias que serían injustificables en un trabajo sobre antecedentes de la LC resultan comprensibles en una investigación sobre corpus lingüísticos antes de la computadora. Esta distinción, importante, permite entender (no forzosamente compartir) la exclusión del *Corpus Iuris Civilis* y, con él, de todos los integrantes de una larga y nutrida tradición de *corpora* constituidos por materiales lingüísticos reunidos con propósitos diferentes del análisis lingüístico. Sin embargo, esta importante distinción no nos conduce a una valoración más positiva de la línea adaptada por Francis, sino probablemente a movernos en sentido contrario. En efecto, si se trata de localizar corpus precomputacionales (del inglés) y un corpus es «a collection of texts...», habrá que demostrar que el conjunto de papeletas elaboradas en el curso de la preparación de materiales para el *OED* (o cualquier otro diccionario) constituye realmente un corpus en el sentido en que se ha definido. Esto es, si bien parece razonable aceptar que esos millones de papeletas pueden ser considerados un antecedente claro de la LC, no es tan seguro que puedan ser considerados como un corpus. *Mutatis mutandis*, algo muy parecido sucede con las papeletas reunidas para los trabajos gramaticales (Jespersen, por ejemplo).[6]

---

[5] Jan Svartvik, «Corpus linguistics 25 + years on», en Roberta Facchinetti, *Corpus linguistics 25 years on,* Amsterdam - Nueva York (Rodopi), 2007, págs. 11-25.

[6] No se me escapa que las papeletas recogidas para la confección de un diccionario pueden ser tratadas como un corpus, que es lo que se ha hecho, por ejemplo, con las del OED. Sin embargo, ese cambio requiere la realización de determinadas operaciones, que son, precisamente, las que marcan la diferencia entre una colección de citas tomadas como ilustraciones del uso de una palabra y el corpus que ese conjunto de citas puede llegar a constituir. *Cfr*. Sebastian Hoffmann, «Using the OED quotations database as a corpus – a linguistic appraisal», *ICAME Journal,* 28 (2004), págs. 17-30.

Todavía menos dudas puede haber, me parece, con respecto a los materiales dialectológicos mencionados por Francis. Los datos laboriosamente recogidos para construir, por ejemplo, un atlas lingüístico son un corpus (esto es, un agregado) de materiales lingüísticos (elementos léxicos, fenómenos morfológicos, variantes fónicas), pero no un corpus lingüístico (es decir, «a collection of texts...»).

Por otro lado, si el concepto de corpus que se maneja permite tomar en consideración conjuntos como los mencionados, resulta sorprendente la exclusión total de la línea consistente en la reunión y análisis de textos enfocada a la producción de listas de frecuencias, casi siempre léxicas y con mucha frecuencia destinadas a la enseñanza de lenguas. La exclusión resulta más curiosa todavía si se tiene en cuenta el hecho, bien conocido, de que la explotación que Francis y Kučera hicieron del corpus que habían construido fue, precisamente, el análisis estadístico de las frecuencias léxicas del *Brown Corpus*[7].

En los años siguientes, este tema es progresivamente integrado en las presentaciones generales de la LC y planteado ya como una búsqueda de antecedentes de esta corriente (aunque las denominaciones utilizadas no siempre lo muestren con claridad). Así, McEnery y Wilson[8] aluden, en solo dos páginas, a una «early corpus linguistics» en la que mencionan trabajos relacionados con la recogida de datos para estudio de la adquisición del lenguaje, frecuencias léxicas para enseñanza de lenguas o comparaciones (estadísticas) entre las frecuencias léxicas y gramaticales en diferentes lenguas. Reconocen aquí que «[t]his type of work was not merely limited to English»[9] y mencionan los trabajos realizados para la confección del francés fundamental. Además, bajo el epígrafe de «spelling conventions» mencionan el corpus de once millones de palabras alemanas analizado por Käding a finales del XIX para obtener las frecuencias de distribución de secuencias de letras[10].

---

[7] Henry Kučera y W. Nelson Francis, *Computational Analysis of Present Day American English,* Providence (Brown University Press), 1967. W. Nelson Francis y H. Kučera, *Frequency Analysis of English Usage: Lexicon and Grammar,* Boston (Houghton Mifflin), 1982.

[8] Tony McEnery y Andrew Wilson, *Corpus Linguistics,* Edimburgo (Edinburgh Univ. Press), 1996.

[9] McEnery y Wilson, *Corpus linguistics,* pág. 4.

[10] Käding era profesor de estenografía, de ahí el interés de la mayor o menor frecuencia de esas combinaciones. *Cfr.* F. W. Käding, *Häufigkeitswörterbuch der deutschen Sprache. Festgestellt durch einen Arbeitsausschuss der deutschen Stenographie-Systeme,* Steglitz bei Berlin (Selbstverlag), 1897-1898.

Un nuevo paso adelante se da en la presentación que hace Graeme Kennedy[11] de los «pre-electronic corpora». Según este autor, la «considerable tradition of corpus-based linguistic analysis»[12] previa a las computadoras se manifiesta en cinco grandes líneas: estudios bíblicos y literarios, lexicografia, estudios dialectales, estudios relacionados con el aprendizaje y enseñanza de lenguas y estudios gramaticales. Dado que tampoco Kennedy presta atención suficiente a lo realizado fuera del inglés, la presentación que hace de las líneas utilizadas por Francis es muy semejante a la que ya hemos examinado[13]. Es de importancia señalar que, dentro de la «corpus-based research» incluye los procesamientos (manuales) de textos destinados a producir listas de lemas y formas destinadas a la enseñanza del inglés, entre los que menciona las realizadas por Thorndike (publicada en 1921) sobre un corpus de 4,5 millones de formas y, algo más tarde, la del propio Thorndike y Lorge (publicada en 1944), ya sobre la muy respetable cantidad de 44 millones de formas. Tras aludir, sin detalles, a la existencia de recuentos similares para otras lenguas, Kennedy cita también el trabajo ya mencionado de Käding.

Dado que se trata de integrar una línea de orígenes y desarrollos alejada de las anteriores, tiene más trascendencia el reconocimiento de la vinculación de los estudios que Kennedy caracteriza como «bíblicos y literarios», cuya presencia entre los antecedentes de la LC se justifica no por su carácter de corpus, sino, más bien, por el empleo de concordancias. Tras caracterizar esta línea de trabajo como «one of the first significant pieces of corpus-based research with linguistic associations»[14], Kennedy menciona las concordancias de los textos bíblicos publicadas por Cruden en 1737, de las que destaca que, en ocasiones, corresponden a combinaciones de palabras no inmediatamente adyacentes.

Algunos años más tarde, en un trabajo centrado en los corpus preelectrónicos, Meyer reajusta las líneas en las que, con sus propias palabras, figuran los

---

[11] Graeme Kennedy, *An Introduction to Corpus Linguistics,* Londres (Longman), 1998.

[12] Kennedy, *An Introduction to Corpus Linguistics,* pág. 13.

[13] Y, claro está, presenta las mismas dificultades. Con respecto a los supuestos corpus de carácter dialectológico, indica que «[m]ost of the work was on lexical variation in the choice of words for particular concepts, and possible variant forms of particular words, both in spelling and pronunciation» (Kennedy, *An Introduction to CL*, págs.15-16).

[14] Kennedy, *An Introduction to Corpus Linguistics,* pág. 13.

«linguistic projects in which preelectronic corpora played an important role»[15]: concordancias bíblicas, gramáticas, diccionarios y el SEU. Como se ve, no aparecen los materiales dialectales ni las listas de frecuencias y, en cambio, las concordancias (bíblicas) están plenamente consolidadas como corpus preelectrónicos. Aunque su atención fundamental en este punto va dirigida a las concordancias de Cruden ya mencionadas, señala la existencia de obras de esta clase considerablemente más antiguas (la de Hugo de San Caro, por ejemplo), sobre otras lenguas (latín, hebreo) y alude también a la ampliación de este tipo de trabajo a obras literarias o autores de especial importancia (Chaucer, entre otros). Por su parte, McCarthy y O'Keefe[16] se refieren, en primer lugar y con cierta extensión, a las concordancias y luego a la papeletización en lexicografía y también a los corpus utilizados por los distribucionalistas estadounidenses.

El análisis somero que hemos realizado de algunas de las presentaciones más conocidas sobre los antecedentes de la LC muestra la existencia de notables oscilaciones acerca de qué tipos de conjuntos textuales pueden merecer la consideración de corpus (preelectrónicos) y, como consecuencia de ello, una cierta confusión con respecto a las líneas a las que deberíamos atender. Siempre se mencionan los materiales reunidos como fondo documental para la confección de diccionarios y casi siempre los utilizados para tratados gramaticales. Las listas de frecuencias y los materiales dialectales son citados en unos casos y no en otros. Por fin, las concordancias realizadas sobre textos como los bíblicos, marginadas por Francis, parecen haberse convertido en el punto de partida en la visión generalizada sobre esta cuestión. Por citar un caso reciente, Aston se remonta a las que Hugo de San Caro hizo en el siglo XIII sobre la Vulgata y señala que «[i]t thus seems right to see him as the first corpus linguist»[17] En una dimensión diferente, pero no menos importante para la comprensión adecuada

---

[15] CHARLES F. MEYER, «Pre-electronic corpora». En Lüdeling, Anke y Merja Kytö (eds.): *Corpus Linguistics. An International Handbook*, Berlin (W. de Gruyter), vol. 2, 2009, págs. 1-14; la cita, pág. 1.

[16] M. MCCARTHY y A. O´KEEFFE, «Historical perspective: What are corpora and how have they evolved». En Anne O'Keefe y Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics,* Oxon (Routledge), 2010, págs. 3-13.

[17] GUY ASTON, «Applied Corpus Linguistics and the learning experience», en Vander Viana, Sonia Zyngier y Geoff Barnbrook (eds.), *Perspectives on Corpus Linguistics*, Amsterdam (John Benjamins), 2011, págs. 1-16; la cita, pág. 1.

de este tema, casi todo lo que se ha publicado se centra en la tradición lingüística inglesa, mientras que las demás tradiciones quedan reducidas a un par de referencias anecdóticas o están totalmente ausentes.

Como consecuencia lógica del desarrollo que ha tenido la LC en los últimos años y la importancia que ha adquirido, la cuestión de sus orígenes y su historia ha dejado la zona marginal en la que se encontraba hasta no hace mucho tiempo y ha pasado a ocupar un lugar central en los tratamientos generales. La primera cuestión que plantean Viana, Zyngier y Barnbrook[18] a los autores que colaboran en el volumen es, precisamente, cuáles consideran que son las raíces de la LC y a qué atribuyen la importancia que ha adquirido. Se trata ya, pues, de un tema de cierta relevancia y, en consecuencia, merece la pena intentar establecer algunas pautas generales que nos permitan tratarlo del modo adecuado.

Aunque pueda parecer una obviedad, para fijar las líneas generales de los antecedentes de la LC no será inútil partir de una caracterización operativa de esta corriente. En realidad, *lingüística de corpus* equivale a 'lingüística basada en el análisis de corpus', que es una aproximación al análisis de los fenómenos lingüísticos tal como se presentan en un corpus textual o en diversos subcorpus textuales. No es necesario insistir en que la cuantificación es un aspecto fundamental en esta perspectiva.

Tal como se entiende hoy habitualmente, un corpus lingüístico es «a set of natural texts (or pieces of texts), stored in electronic form, assumed to be jointly representative of a linguistic variety in some of its components, or in all of them, and grouped together so that they can be scientifically studied»[19]. Se trata, pues, de la utilización de un conjunto amplio de materiales textuales reunidos con un determinado diseño gracias al cual se puede considerar que son representativos de una determinada variedad lingüística. Dicho de otro modo, es la aplicación a los estudios lingüísticos contemporáneos de lo que tradicionalmente se ha entendido como *corpus*. Entre los diversos significados de la palabra latina derivados del de carácter físico, el *Oxford Latin Dictionary* recoge «any structure comparable to a body, a fabric framework» (ac. 6) y «a comprehensive collec-

---

[18] Viana, Zyngier y Barnbrook (eds.), *Perspectives on Corpus Linguistics.*

[19] Guillermo Rojo, «Hispanic Corpus Linguistics», en Lacorte, Manel (ed.), *The Routledge Handbook of Hispanic Applied Linguistics*, Londres (Routledge), 2014, 371-387. La cita, pág. 371.

tion of facts on a given subject; a compendium of scientific, literary or other writings, an encyclopaedia, etc.» (ac. 16). Es precisamente esta última acepción la que fue ampliamente utilizada en Europa occidental después de la caída del imperio romano y aparece en conjuntos bien conocidos como el *Corpus Iuris Civilis* o, para acercarnos a nuestro tiempo, el *Corpus Inscriptionum Latinarum*. Aunque habitualmente se piensa en corpus textuales, no es difícil encontrar ejemplos de corpus constituidos por objetos de otros tipos, como el *Corpus vasorum antiquorum*[20] o el *Corpus vitrearum*[21]. La idea general, pues, es que un corpus consiste en un conjunto de objetos (casi siempre textos) reunidos con la intención de facilitar el análisis de su contenido.

En el caso que nos interesa aquí, construir un corpus textual implica fijar unos objetivos generales, establecer un diseño, seleccionar los textos adecuados y, cuando es necesario, decidir cuál es la versión más recomendable de cada uno de ellos. En distintos grados y con diferentes consecuencias, estos rasgos están presentes en la construcción del corpus homérico, el corpus bíblico, el corpus legislativo promovido por Justiniano, las Siete Partidas o el *Corpus del español del siglo XXI (CORPES)* de la RAE. En efecto, el rasgo distintivo de los corpus lingüísticos que con mayor insistencia se ha venido destacando en los últimos años es, precisamente, el del diseño, que en cierto modo implica todos los demás. Frente a lo que en la LC tradicional se llamaba un archivo (esto es, un conjunto de textos desvinculados entre sí y reunidos en un recurso único), un corpus supone la búsqueda de la representatividad y el equilibrio, que los textos que lo integran han sido seleccionados en función de determinados parámetros y que, en definitiva, constituyen un todo organizado y estructurado. Esa es la caracterización con la que deben ser contrastados los candidatos a corpus preelectrónicos, aceptando, como es natural, diferencias en los objetivos específicos (legislativos, literarios) y, por supuesto, en el soporte.

Debe quedar claro también que hasta ahora hemos hablado de corpus, no de concordancias. Es evidente que, una vez constituido el corpus en cuestión, su extensión y complejidad puede hacer conveniente o casi imprescindible disponer de un sistema rápido y fiable de localización en el corpus de aque-

[20] http://www.cvaonline.org/cva/.

[21] http://cvi.cvma-freiburg.de/.

llos fragmentos que pueden interesar por alguna razón. Dado que hemos de situarnos inicialmente en la tradición bíblica cristiana, lo esperable es que el interés se establezca fundamentalmente en llevar a cabo un inventario fácil de manejar en el que aparezcan identificados los pasajes en los que los textos bíblicos se refieren a ciertos temas. Es decir, los diferentes fragmentos que *concuerdan*[22] en tratar los mismos temas y, claro está, en hacerlo de modo parecido. Las llamadas «concordancias reales» (*concordantiae rerum*) bíblicas pueden remontarse a períodos anteriores, pero su aparición en forma próxima a la definitiva se sitúa en la primera mitad del siglo XIII, con las realizadas por Antonio de Padua[23]. De la misma época datan las primeras concordancias verbales, elaboradas por el dominico Hugo de San Caro, ya organizadas a partir de las palabras[24]. En realidad, se trata de lo que hoy llamamos índices, es decir, se limitan a señalar el lugar aproximado *(cfr*. nota anterior) en el que se localiza la palabra que encabeza la entrada. A estas *Concordantiae breves* siguieron unos veinte años después las llamadas *Concordantiae majores* o *Con-*

[22] El *Diccionario de autoridades* define las *concordancias* como «[l]as tablas de lugares semejantes en razónes ù dicciones: como son las concordancias de la Biblia» (s. v. *concordancia).*

[23] «Esta especie de concordancias distribuye los materiales de la Sagrada Escritura en cierto número de epígrafes, por ejemplo: caridad, fe, redención, infierno, justicia, etc. y, disponiéndolos en orden alfabético, facilitan a los predicadores, teólogos, etc. […] el hallazgo de los pasajes de la Sagrada Escritura donde se tratan las materias que quieren estudiar. El inventor de este género de concordancias fue san Antonio de Padua (1195-1231), con su obra *Concordantiarum moralium in S. Biblia Libri V* (*Enciclopedia universal ilustrada europeo-americana*, Bilbao – Madrid – Barcelona (Espasa-Calpe), 1908-1930, s.v. Versión electrónica del artículo sobre concordancias en http://www.filosofia.org/enc/eui/e610155.htm [consultado 8/2/2014].

[24] *Cfr*. SUZANNE HANON, «La concordance»», en Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta (eds.), *Wörterbücher. Dictionaries. Dictionnaries. Ein internationales Handbuch zur Lexicographie.* Berlin (W. de Gruyter), II, 1990, págs. 1562-1576; la cita, pág. 1563. Hugo de San Caro (Hugues de Saint-Cher, Hugo de Sancto Caro), fallecido en 1263, fue un dominico (llegó a cardenal) que invirtió diez años en esta tarea, auxiliado, al parecer, por quinientos frailes de su misma orden. Además de la indicación del libro, las referencias incluyen el capítulo (estructuración introducida poco antes por Stephen Langton, arzobispo de Canterbury). Y para lograr indicaciones más precisas, cada capítulo fue dividido en siete fragmentos de aproximadamente la misma extensión (*a, b, c*, etc.). La organización en versículos fue introducida por Robert Estienne en 1525. *Cfr*. *Catholic Encyclopedia* (1907-1912); versión digitalizada en http://www.catholic.org/encyclopedia/ [consultada 8/2/2014]. Puede encontrarse una versión digitalizada del manuscrito de la obra de San Caro en la Bayerische StaatsBibliotek (http://bildsuche.digitalesammlungen.de/?c=viewer&bandnummer=bsb00034253&pimage=00001&v=100&einzelsegmentsuche&mehrsegmentsuche&l=es).

cordantiae anglicanae[25], que incluían ya un fragmento del texto en que aparecía la palabra en cuestión. Como se puede apreciar, en un período de menos de cincuenta años se establece la estructura conceptual básica para las concordancias y se llega, desde los índices conceptuales a las concordancias verbales, pasando por índices verbales. El paso siguiente, que se demorará varios siglos, consiste en aplicar esta técnica a textos de otro tipo, como los de Shakespeare o Chaucer, que son las llamadas concordancias de autor[26]. El engarce de esta larga tradición con las computadoras está perfectamente claro en la obra que lleva a cabo Roberto Busa a comienzos de la década de los 50 sobre la obra de Tomás de Aquino[27].

No parece que pueda haber dudas de la continuidad de la tradición sobre corpus y la utilización de las concordancias como procedimiento básico para su explotación. Con independencia de que esta se refiera a cuestiones estrictamente lingüísticas o de otro tipo, es evidente que la organización se hace, salvo en el caso de las *concordantiae rerum*, sobre elementos lingüísticos y, por tanto, estamos siempre ante recursos del mismo tipo. La introducción de las computadoras produce una considerable disminución del trabajo manual que supone elaborar las concordancias de una obra o un conjunto de obras, pero no altera la naturaleza de la tarea realizada.

Hasta la llegada de las computadoras, reunir un conjunto de textos y elaborar las concordancias de las palabras contenidas en ellos era un trabajo enorme, pero realizable siempre que se dispusiera del tiempo y los medios necesarios[28].

---

[25] *Cfr*. HANON, «La concordance», pág. 1563.

[26] *Cfr*. OLGA M. KARPOVA, «Author concordances, with special reference to Shakespeare», en R. R. K. Hartman (ed.), *Lexicography: Critical Concepts*, Londres (Routledge), vol. III, 2003, págs. 112-123.

[27] ROBERTO BUSA, *Index Thomisticus: Salncti Thomae Aquinatis operum indices et concordantiae* Stuttgart (Frommann-Holzboog), 56 vols., 1974-1980. R. BUSA, «The Annals of Humanities Computing: The Index Thomisticus», *Computers and the Humanities,* 14 (1980), págs. 83-90.

[28] Las dirigidas por Hugo de San Caro supusieron, según parece, varios años de trabajo de unos 500 monjes. Cruden hizo lo mismo para la versión inglesa (la Biblia del rey Jaime) en dos años de trabajo a dieciocho horas diarias y siete días a la semana (*cfr*. MEYER, «Pre-electronic corpora», pág. 2). Según las interesantes informaciones estadísticas que se facilitan en el proyecto *Perseus* (http://www.perseus.tufts.edu/hopper/), el texto de la *Vulgata* consta de 620 930 formas (consultado el 11/03/2014). La Biblia del rey Jaime consta de algo más de 780 000 palabras *(cfr*. los recuentos por

Pero es evidente que, a partir de un determinado volumen de textos, esta técnica es ya inviable. Los primeros miembros de la RAE no podían pensar en seleccionar las obras que pudieran servir de base para su diccionario y elaborar las concordancias conjuntas o bien obra a obra. Y tampoco podía hacerlo Rufino José Cuervo. Los proyectos lexicográficos que basan sus lemarios y definiciones en lo que se puede encontrar en los textos reales tienen que aplicar una técnica bastante diferente: determinar con carácter más o menos cerrado las obras en las que van a basarse y extraer de ellas los ejemplos que consideran relevantes para el diccionario. No se trata, pues, como en el caso de las concordancias, de poder localizar todos los casos de una determinada forma en el corpus analizado, sino de extraer de un cierto conjunto de textos (potencialmente abierto a la introducción de nuevos casos) aquellos ejemplos que se consideran de interés. La idea clave es, pues, la selección. Son bien conocidos los problemas que este procedimiento supone, pero no es este el lugar de tratarlos[29]. Lo que cuenta para nuestro propósito en este trabajo es que creación de corpus y producción de concordancias por un lado y despojo de las citas que se consideran de interés para la obra que se está preparando son dos procedimientos considerablemente distintos y no parece que el segundo pueda ser tenido en cuenta como antecedente de la LC si nos movemos con criterios estrictos[30]. Lo mismo, *mutatis mutandis*, cabe decir de los ficheros reunidos como fondo documental para tratados gramaticales, puesto que también en este caso se trata de hacer fichas de aquellos fragmentos que, por alguna razón, resultan relevantes.

---

libros que figuran en Nic Kizziah, «King James Bible Statistics» (http://www.biblebelievers.com/believers-org/kjv-stats.html; consultado el 11/03/2014).King

[29] Quirk señala con toda claridad las diferencias. Es posible, dice, que gramáticos o lexicógrafos usen un corpus «as a convenient source for 'good examples' to put in their grammar. But that is not where the value or the challenge of a corpus will lie. If we ignore the value and evade the challenge of total accountability, our use of a corpus will be no advance on Jespersen's use of his voluminous collections of slips or Murray's use of those file boxes bursting with marked-up quotations for the OED. Such scholars certainly ensured that everything in their published volumes was firmly anchored in textual reality, but not that everything in their samples of textual reality was reflected in those published volumes» (Randolf Quirk, «On corpus principles and design», en Jan Svartvik [ed.], *Directions in Corpus Linguistics,* págs. 457-469; la cita, p. 467).

[30] El caso paradigmático de la utilización de un corpus para la confección de un diccionario es, por supuesto, el COBUILD. *Cfr*. John Sinclair, «Introduction» a *Collins Cobuild English Language Dictionary*, Londres (Harper Collins Publishers), 1987.

En resumen, de esta tradición de estudios lexicográficos y gramaticales podemos decir que trabaja con casos seleccionados de un conjunto de textos más o menos amplio y que, en ese sentido, emplea un corpus textual y construye un corpus de ejemplos. Es evidente que, en la expresión anterior, la palabra *corpus* no tiene el sentido con que la hemos empleado anteriormente. Y también lo es que los recursos que lexicógrafos y gramáticos podían utilizar antes de la difusión de las computadoras no permitían pasar de esta fase. Es cierto que se puede trabajar con técnicas semejantes a las empleadas en LC sin computadoras, pero, en ese caso, el corpus tiene que ser de tamaño reducido y, además, se precisa una lectura completa para cada uno de los fenómenos que van a ser estudiados. Es viable, pues, pero con fuertes limitaciones que restringen considerablemente su empleo real. El diseño inicial del SEU es, en mi opinión, el que más se acerca a esta posibilidad: fichado (casi) exhaustivo del contenido de un corpus de un millón de formas[31].

Lo visto hasta aquí debería dejar claro que la reunión de materiales dialectales del estilo de los señalados a veces en la bibliografía (*cfr*. supra) no reúne las condiciones precisas para poder ser considerado entre los antecedentes de la LC ni constituyen un corpus.

Un caso diferente se plantea con los estudios de frecuencias, que algunos autores han incluido entre los antecedentes de la LC. Lo que se hace en esta orientación consiste en analizar exhaustivamente un corpus de textos que, dada la finalidad estadística que se persigue, tiene una distribución que se supone representativa y equilibrada del universo que se pretende incluir en el estudio. Es decir, posee el rasgo de diseño que hemos señalado como la característica básica de un corpus textual en el sentido técnico. Lo que sucede aquí es que, de nuevo por las limitaciones propias de la era preelectrónica, el resultado se reduce al recuento de cada uno de los elementos o fenómenos analizados. Por supuesto, eso es lo que se pretende, de modo que no cabe considerarlo como una deficiencia de planteamiento. Pero, para lo que aquí se persigue, es importante tener en cuenta que en esta línea se configura un corpus textual y se realizan recuentos exhaustivos de sus elementos léxicos o, en pocos casos, de estructuras gramaticales; se conserva el resultado

[31] *Cfr.* http://www.ucl.ac.uk/english-usage/about/quirk.htm.

del recuento y se desechan los textos (salvo, por razones perfectamente comprensibles, la inclusión de algunas referencias o ejemplos ilustrativos en el caso de los recuentos gramaticales). En lo que respecta al español, existe un grupo relativamente bien nutrido de estudios de frecuencias léxicas, el más elaborado de los cuales es, sin duda, el *Frequency Dictionary of Spanish Words*[32], que pudo beneficiarse ya de la utilización de computadoras para la realización de los cálculos. En cuanto a los estudios de frecuencias de estructuras gramaticales, mucho menos abundantes por razones obvias, las dos contribuciones de Keniston[33] siguen constituyendo un ejemplo deslumbrante que nadie ha podido superar todavía.

Así pues, la línea de antecedentes que lleva hasta la constitución de la LC tal como la conocemos hoy en día resulta bastante más clara e interesante de lo que se ve habitualmente en la escasa bibliografía existente. En la vía central, la tradición del diseño, construcción y explotación (concordancias) de corpus textuales de diferentes tipos y finalidades. Como hemos visto, su aplicación estricta está limitada por la imposibilidad de llevar a cabo el análisis exhaustivo de textos con un tamaño total excesivo. A su lado, por una vía lateral, la línea consistente en la lectura de cantidades considerablemente mayores de textos para extraer de ellos los fragmentos que se consideran de interés para delimitar el significado de una palabra e ilustrar su uso o bien las características de un determinado fenómeno gramatical. Relajando un tanto las características exigibles a un corpus, se puede decir que consiste en la selección de casos ilustrativos de un elemento o fenómenos presentes en un corpus abierto. Por fin, por otra vía lateral, la línea consistente en el recuento de los casos de un cierto elemento o fenómeno que se encuentran en un conjunto determinado de textos. Como en el caso anterior, podría hablarse de un corpus abierto del cual se extraen resultados, pero no se conserva el texto.

---

[32] Alphonse Juilland y E. Chang-Rodríguez, *Frequency Dictionary of Spanish Words,* La Haya (Mouton), 1964.

[33] H. Keniston, *The Syntax of Castilian Prose. The Sixteenth Century.* Chicago, (The University of Chicago Press), vol. IL, 1937. *Vid.* también *Spanish Syntax List: A Statistical Study of Grammatical Usage in Contemporary Spanish Prose on the Basis of Range and Frequency,* Nueva York (H. Holt and Co.), 1937.

La llegada de las computadoras hace que estas tres líneas puedan superar los límites existentes hasta ese momento y confluir. Las computadoras permiten que la edición del contenido de un texto en forma de concordancias se pueda llevar a cabo de forma mucho más rápida y cómoda (no tanto como en 2014, por supuesto), con lo que la reunión de materiales para trabajos lexicográficos y gramaticales se simplifica enormemente. Lo mismo, claro está, en lo que se refiere a la realización de recuentos léxicos. Durante unos cuantos años, la construcción de corpus (Brown Corpus, Lancaster-Oslo-Bergen Corpus, etc.), que en muchos casos son explotados casi exclusivamente para estudios de estadística léxica, alterna con la introducción en computadora de textos para producir listas de frecuencias léxicas y léxicos inversos, como sucede, en el caso del español, con los proyectos ONE71[34] o PE77[35] o bien como apoyo a la elaboración de diccionarios, como el *Dictionary of Old Spanish Language*.[36] Por una línea diferente, 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), pretendía reunir materiales que permitieran analizar la variación en las grandes ciudades del mundo hispánico.[37] De forma totalmente natural, muchos de esos proyectos terminan generalizando su uso y convirtiéndose en corpus (Spanish on Line, SOL, en los dos primeros casos, la *Biblioteca digital de textos de español antiguo* en el tercero y el *Macrocorpus de la norma lingüística culta de las principa-*

[34] *Once novelas españolas*. Proyecto realizado por Mighetto y Rosengren en la Universidad de Goteburgo, consistente en introducir en computadora el texto de once novelas españolas de los años 1951 a 1971, con un total ligeramente superior al millón de formas. *Cfr*. David Mighetto, *ONE71. Banco de datos de once novelas españolas 1951-1971* (Univ. de Göteborg), 1985.

[35] *Prensa española de 1977*. Obra también de Mighetto y Rosengren, consistente en introducir el texto de unos 3000 artículos procedentes del periódico *El País* y la revista *Triunfo* en 1977. Consta de algo más de dos millones de formas. *Cfr*. David Mighetto y Per Rosengren, *Banco de datos de Prensa española 1977. Concordancia lingüística y texto fuente* (Universidad de Göteborg), 1982.

[36] Dirigido en la década de los 70 por Lloyd A. Karsten y John J. Nitti en el *Hispanic Seminar of Medieval Studies* de la Universidad de Wisconsin - Madison.

[37] *Cfr*. Juan M. Lope Blanch, «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), 1967, págs. 255-264, Juan M. Lope Blanch, *El estudio del español hablado culto. Historia de un proyecto.* México (UNAM), 1986.

les ciudades del mundo hispánico como conversión parcial del último)[38]. La facilidad creciente para la digitalización de los textos, la evolución de la capacidad de memoria y de la velocidad de cálculo de las computadoras por un lado y las ventajas derivadas de la aparición de internet por otro hacen el resto.

---

[38] SOL, *Spanish On Line* (http://spraakbanken.gu.se/konk/rom2/), *Biblioteca Digital de Textos del Español Antiguo* (http://www.hispanicseminary.org/textconc-es.htm). J. A. Samper Padilla, C. E. Hernández Cabrera y M. Troya Déniz (eds.), *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*, Las Palmas (Universidad de Las Palmas de Gran Canaria), 1998. Para más detalles sobre estos proyectos que se sitúan a ambos lados de la línea de separación, *cfr*. G. Rojo, «Los corpus textuales del español», en Javier Gutiérrez-Rexach (ed.), *Enciclopedia lingüística hispánica*, Londres (Routledge). En prensa.