Capítulo 141: 1151-1163

Aguja de navegar corpus

Guillermo Rojo

En Víctor M. Castel y Liliana Cubo de Severino, Editores (2010)
La renovación de la palabra en el bicentenario de la Argentina.
Los colores de la mirada lingüística.
Mendoza: Editorial FFyL, UNCuyo.
ISBN 978-950-774-193-7

Aguja de navegar corpus

Guillermo Rojo

Universidade de Santiago de Compostela

Santiago de Compostela, España

guillermo.rojo@usc.es

Resumen

Esta ponencia pretende trazar un panorama general de lo que ha supuesto la difusión de la lingüística de corpus (LC) tanto en la investigación lingüística en general como en el ámbito específico de la lingüística hispánica. Se establecen en primer lugar sus rasgos diferenciales con respecto a otras aproximaciones y se sitúa luego el lugar que ocupan los llamados 'corpus de referencia' frente a otros tipos de corpus. Finalmente, se analizan las características fundamentales que presenta la LC en la actualidad.

Introducción

Debo comenzar, y les aseguro que no es simple resultado de la cortesía habitual en estos casos, por agradecer a la Junta Directiva de la SAL y a los organizadores de este congreso la honrosa invitación que me han hecho a participar en él y encomendarme una de las ponencias plenarias. Quisiera aprovechar esta magnífica ocasión que me han brindado para desgranar ante ustedes algunas ideas en torno al estado de la lingüística de corpus en el mundo hispánico. No, por supuesto, en el sentido de la revisión que hice ya en el último congreso de la ALFAL (Rojo, 2008), sino más bien en el de la reflexión acerca de la labor que hemos realizado entre todos, de la situación a la que hemos llegado y de las nuevas características que presenta el trabajo que hacemos los que estamos aquí reunidos, tanto si nos sentimos cómodos bajo la etiqueta 'lingüística de corpus' (LC) como si marcamos distancias con respecto a ella o incluso la rechazamos. Creo que hemos hecho mucho en los últimos treinta o cuarenta años y, en general, lo hemos hecho bien y de prisa, pero creo también que hay algunos factores en los nuevos modos de trabajar sobre los que probablemente no hemos reflexionado lo suficiente. De ahí que me haya permitido beneficiarme del ingenio de Quevedo para dar a esta ponencia el título de *Aguja de navegar corpus*, que utilicé ya para una comunicación, no publicada y muy diferente de esta en cuanto a contenido, que presenté en un coloquio celebrado en Tubinga hace unos cuantos años.

La LC frente a otras aproximaciones

En los primeros párrafos de su introducción a BNCWeb, el programa general diseñado para facilitar la explotación de los datos existentes en el BNC, propone Sebastian Hoffmann (2008) a sus lectores que se pregunten cuáles son sus ideas acerca del significado de la palabra *goalless*, la vigencia en el uso del modal *shall* y si son las mujeres o los hombres quienes hablan más de *cars*. Aunque aquí no importan demasiado las respuestas, el resultado, muy simplificado, consiste en que todos los casos de *goalless* contenidos en el BNC tienen el significado de 'empate a cero' en una competición deportiva, no el de 'sin objetivos'; que el uso de *shall* se mantiene en los grupos jóvenes e incluso tiene índices de empleo más altos que en generaciones de mayor edad en combinaciones del tipo *shall I/we* (pero no en las del tipo *I/we shall*); por último, que la forma *cars* es más usada por mujeres, aunque eso no significa que hablen más de autos que los hombres. Más que las respuestas concretas, que apunto aquí solo para que se puedan valorar los resultados, lo que interesa es el hecho de que llegar a estas conclusiones, que chocan en algunos casos con lo que los propios hablantes ingleses creen saber acerca de su lengua, solo es posible si tenemos la opción de consultar un corpus, de hacer la consulta estableciendo los subcorpus pertinentes en cada caso y, naturalmente, de analizar todos los casos relevantes.

Hoffmann se centra en el BNC, pero resulta sencillo extender sus ejemplos a otros casos igualmente ilustrativos, como diferentes construcciones sintácticas en las que aparece un verbo y frecuencia con que lo hace, evolución de una cierta construcción a lo largo del tiempo, mayor o menor frecuencia de un cierto fenómeno o una forma determinada en diferentes épocas o tipos de texto, frecuencia de palabras o acepciones por países, etc. En general, los corpus textuales constituyen el modo de conocer los usos lingüísticos, la forma en que se usa una lengua en sus diferentes componentes o se ha usado a lo largo de su historia. Este carácter fuertemente basado en los datos empíricos obtenidos de los textos realmente producidos distingue con bastante nitidez la lingüística de corpus de la orientación mentalista, de raíz chomskyana, más preocupada por la competencia, por el conocimiento del hablante. En cambio, no permite ver con la misma claridad una diferencia que, a mi modo de ver, es realmente crucial: la distancia entre la LC y la lingüística descriptiva tradicional, tal como se manifestaba hace treinta o cuarenta años.

Es bien conocido que la LC tuvo una vida complicada en sus primeros años, a causa, sobre todo, de las reacciones negativas que los corpus, el análisis de los textos reales o el papel de la frecuencia producían en los partidarios de la LGT, que, precisamente en esos años, estaba experimentando su mayor crecimiento y expansión. No tiene sentido dedicar aquí tiempo al análisis de una polémica felizmente superada, pero debo decir que, en mi opinión, este tema no se resuelve simplemente con pensar que las tan repetidas afirmaciones de Chomsky acerca del carácter inevitablemente sesgado de los corpus o de la inutilidad de la consideración de la frecuencia en los estudios gramaticales estaban erradas y que el tiempo ha terminado por demostrarlo. Chomsky pensaba en cierto tipo de corpus (el de los distribucionalistas) y en cierto uso de la estadística en la gramática (el de Hockett, por ejemplo). En alguna medida, esas críticas eran correctas, pero unos y otros las han extraído de su contexto originario y, lo mismo que ha ocurrido con tantas otras formulaciones de Chomsky sobre el estructuralismo, las han aplicado a corrientes distintas de las que él tenía en mente en aquel momento, con lo que se ha producido un desajuste considerable.

La división entre las dos aproximaciones sigue ahí, sin duda, pero las formulaciones son hoy muy distintas de las que tuvieron originariamente. Los corpus que utilizamos en la actualidad y los métodos con que los estudiamos están tan lejos del *Brown Corpus* (1964) como los métodos para el establecimiento de los grados de gramaticalidad lo están de la vieja imagen del lingüista desdoblado en hablante-oyente ideal, empeñado en averiguar cuántas transformaciones y en qué orden de actuación sufría una extraña secuencia que nadie habría pronunciado nunca en una situación natural. A pesar de los años transcurridos, sigo creyendo que Fillmore estaba en lo cierto cuando, al oponer los que llamaba 'lingüista de sofá' y 'lingüista de campo', indica

I have two main observations to make. The first is that I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore; all that I have seen are inadequate. The second observation is that every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way. My conclusion is that the two kinds of linguists need each other. Or better, that the two kinds of linguists, wherever possible, should exist in the same body (Fillmore, 1992: 35).

En el año 2001, al hacer la presentación de los resultados de su proyecto de análisis léxico, Fillmore volvía sobre la cuestión y señalaba

[I]n this project, corpus evidence is the main tool, and researchers who are speakers of English are hired to work with this resource and decide how to use what it gives us. Since for our purely lexicographic purposes, corpus evidence and our ability to interpret it provide more lexically specific information than can be found in dictionaries of lexical descriptions known to us, we are daily rewarded with insights about our language that introspection alone, however disciplined, could never direct us to. The limitation to lexical observations, of course, allows us to escape larger-scale and 'deeper' kinds of linguistic facts: our work can proceed with 'canonical' examples of the uses of the lexical units we target for study (Fillmore, 2001: 1).

Dado lo anterior, resulta bastante más interesante explorar un aspecto al que se ha prestado una atención mucho menor, que es el de las diferencias existentes entre la lingüística descriptiva tradicional y la LC tal como la entendemos y practicamos en la actualidad. En mi opinión, la comprensión de las divergencias metodológicas existentes permitirá situar mejor la LC en el panorama de los estudios lingüísticos a comienzos del siglo XXI. La clave está en dar el relieve adecuado a la consideración de que las computadoras han provocado una auténtica revolución instrumental en muchas ciencias, la lingüística entre ellas.

El marco adecuado para comprender este proceso en todas sus dimensiones es, en mi opinión, el trazado por Freeman Dyson. Para Dyson, la ciencia moderna procede de la fusión de dos grandes tradiciones: "la tradición del pensamiento filosófico que se originó en la antigua Grecia y la tradición de los oficios especializados, que es aun más antigua, pero que floreció en la Europa medieval" (Dyson, 1999: 24). Los cambios en la primera línea dan lugar a las revoluciones científicas con las que Kuhn (1962) modificó radicalmente la concepción habitual del progreso científico. La sustitución de un paradigma que ha entrado en crisis –porque no puede explicar las anomalías que se han ido acumulando en un período de ciencia normal– por otro capaz de hacerlo es rebautizada por Dyson como una *revolución conceptual*. Consisten en "explicar cosas antiguas de nuevas maneras" (Dyson, 1997: 50) y también fenómenos que antes no podían ser adecuadamente comprendidos. El caso más citado es, por supuesto, el cambio de la concepción geocéntrica a la heliocéntrica.

Las revoluciones conceptuales son muy llamativas, pero –afortunadamente– se dan muy de tarde en tarde. Mucho más frecuentes y, en la mayor parte de los casos, con mayor repercusión sobre la vida cotidiana y el quehacer científico, son las innovaciones que tienen lugar en la segunda línea, como consecuencia de cambios en las herramientas utilizadas (no solo físicas) o la aparición de nuevos instrumentos. Son las que Dyson denomina *revoluciones instrumentales* (*tool-driven revolutions*), cuyo efecto "es descubrir cosas nuevas que tienen que ser explicadas" (Dyson, 1997: 50). El caso más citado es el que tuvo lugar en el momento en que Galileo contempló el firmamento con el rudimentario telescopio que había conseguido fabricar. Apareció entonces ante sus ojos un panorama mucho más rico y bastante más complejo que el que los seres humanos habían podido observar, a simple vista, hasta ese momento.

Pues bien, las computadoras son las herramientas que producen la gran revolución instrumental en lingüística y, como consecuencia de lo que han puesto a nuestro alcance, han dado lugar a un fuerte cambio en la forma en que entendemos el funcionamiento de las lenguas. Aarts (2000 y 2002) distingue dos grandes tipos de datos lingüísticos: intuitivos y no intuitivos. Los primeros son tanto los introspectivos como los obtenidos de los juicios de otros hablantes. Los no-intuitivos “are provided by what people actually say and write” (Aarts, 2002: 4) y se dividen en ocasionales o asistemáticos (*anecdotal*) y procedentes de corpus. En realidad, como él mismo reconoce a continuación,¹ no se trata de diferentes tipos de datos, sino de distintos modos de reunir y analizar los datos externos, procedentes de actos lingüísticos efectivamente realizados. Aquí reside, me parece, la gran diferencia que se da entre la LC y la lingüística descriptiva tradicional. Se ha apuntado en diversas ocasiones que la LC suponía un enfoque diferente en ámbitos en los que el manejo de los datos reales se ve oscurecido por factores como el peso de un cierto ideal de lengua o el recurso al sentimiento lingüístico, pero no se da esa novedad, en cambio, en las zonas en las que el manejo de datos externos era obligado, como sucede en las investigaciones sobre épocas anteriores de las lenguas. Sin embargo, es precisamente en este terreno donde las dos metodologías muestran sus divergencias con más claridad. Por razones perfectamente comprensibles dadas las circunstancias y los medios con que se trabajaba, la recogida de materiales se hacía habitualmente de modo selectivo, unos cuantos ejemplos de un cierto número de textos (en la tradición hispánica, casi siempre literarios, pero ese es otro problema) o bien todos los ejemplos existentes en unos pocos textos. En cualquier caso, la recolección era el resultado de lo que se sabía o suponía del fenómeno en cuestión precisamente en la fase en que se estaban coleccionando los datos para su estudio, las hipótesis previas de los recolectores, su estimación de los factores relevantes, etc. Como consecuencia de todo ello, los conjuntos de datos reunidos son casi siempre reducidos o muy reducidos, no es seguro que atiendan a todo lo que se debería tener en cuenta y, además, presentan una inevitable sobrerrepresentación de lo extraño, lo inusitado, lo que llama la atención por escasamente habitual. En definitiva, casi siempre una muestra corta y, con frecuencia, peligrosamente sesgada.

El volumen de los materiales utilizados y el sistema de recogida puede ser ilustrado, entre otros muchos casos, con las palabras de Félix Restrepo al presentar la continuación del *Diccionario de Cuervo*. Quiero dejar claro que no aparecen aquí como una crítica fácil, sino como respetuosa descripción de lo que era el modo de trabajar en lo que considero la mejor tradición hispánica de análisis de la realidad lingüística. Según Restrepo, los materiales dejados por Cuervo para las entradas cuya redacción no había llegado a acometer, resultaban, en muchos casos, insuficientes:

Hicimos, D. Pedro Urbano y yo, un considerable esfuerzo para clasificar y redactar la partícula *en*, para la cual dejó Cuervo cerca de 600 ejemplos, y pudimos comprobar que estos ejemplos son pocos si esta partícula, una de las más usadas en castellano, ha de guardar proporción con las ya publicadas *a* y *de*, en cada una de las cuales los ejemplos pasan de 1000. Quedaba pues la partícula *en* muy inferior a las anteriores y por consiguiente no tenemos ninguna garantía de que todos los usos de ella estén representados en los ejemplos coleccionados por Cuervo. En realidad, en la clasificación adoptada por nosotros hay usos importantes que apenas están autorizados por uno o por muy pocos ejemplos, cuando es práctica general de Cuervo apoyar con ejemplos abundantes cada uno de los usos de las palabras que estudia (Restrepo, 1945: 430; cf. también Rojo, en prensa).

En cuanto a la sobrerrepresentación de lo inusual, es bien conocido que, en un proyecto modélico, representativo de la mejor línea de la lexicografía tradicional, Murray tuvo que pedir a sus colaboradores que no olvidasen enviar también ejemplos de palabras corrientes en sus significados habituales. Los apartados 5 y 7 de las *Directions to Readers for the Dictionary* indican:

5. Make a quotation for *every* word that strikes you as rare, obsolete, old-fashioned, new, peculiar, or used in a peculiar way. [...]

7. Make *as many* quotations as *convenient to you* for ordinary words, when these are used significantly, and help by the context to explain their own meaning, or show their use (Murray, 1879a)

Y en las *Additional Notes* a esas instrucciones insiste en este punto:

If Readers will kindly remember that the Dictionary is to contain *all* English words ordinary and extraordinary, that it is to give, if possible, one quotation in each century for every sense or construction of every word, and that it is these quotations that we ask them to supply by their reading, they will at once see why we ask them to give us, not only all the *extraordinary* words or constructions in their books, but also as many *good, apt, pithy* quotations for ordinary words as their time and patience permit. (Murray, 1879a).

Murray hace esta solicitud a partir del conocimiento derivado del análisis de los materiales recogidos en las fases anteriores del proyecto.

In my own opinion, the *Bases of Comparison* formerly issued by the Society were a mistake, and detrimental to the work which they were designed to serve. Their most obvious result, to one who examines the material, is, that while rare, curious, and odd words, are well represented, ordinary words are often most meagrely present; and the editor or his assistants have to search for precious hours for examples of common words, which readers passed by because they happened to find them put down in their 'Basis', as occurring in the Bible or in Burke. Thus of *Abusion*, we found in the slips about 50 instances: of *Abuse* not five, and we had to spend much time in tracing out the early occurrence of this word which readers had omitted to record. This is why we have asked every reader

to give as many *common words* as he conveniently *can*: I had almost asked that rare and odd words should be omitted,-as apparently we have them all- and only common words noted henceforth (Murray, 1879b, 571-572; cursivas en el original; vid. también Murray, 1977: 178 y Rojo, en prensa).

No se olvide, además, que la obtención manual de grandes cantidades de ejemplos, su clasificación, archivo, recuperación y demás operaciones necesarias para la ordenación de los ejemplos suponía una enorme inversión en tiempo, de modo que era inevitable hacer una selección de los materiales. Lo malo, claro está, es que esa selección condicionaba, en muchos casos de forma realmente fuerte, la visión que se podía obtener posteriormente del fenómeno en cuestión.

El manejo de corpus textuales, en cambio, permite superar todos esos inconvenientes y aspirar al análisis exhaustivo de los materiales pertinentes. La recuperación de todos los casos de un cierto fenómeno presentes en un corpus depende de las características del corpus en cuanto a codificación de los textos y anotación de las formas, pero, en cualquier caso, es un proceso que se puede dar con una rapidez y una comodidad totalmente distintas de las que se daban en el mundo tradicional. En otras palabras, se puede aspirar a lo que Leech o Quirk, entre otros, han llamado la 'total accountability'. En palabras de Leech, los datos presentes en un corpus

are used exhaustively: there is no prior selection of data which we are meant to be accounting for and data we have decided to ignore as irrelevant to our theory. This principle of "total accountability" for the available observed data is an important strength of CCL [= computer corpus linguistics, G.R.] (Leech, 1992: 112).

No se trata, por tanto, de la superficial distinción entre trabajar sobre las fichas previamente seleccionadas o los ejemplos obtenidos, también mediante filtro, de un corpus, sino de una marcada diferencia en el modo de enfrentarse con los datos. Quirk lo expresa con total claridad cuando señala que es posible que los gramáticos (o los 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 (Quirk, 1992: 467).

Contemplado desde la situación actual, parece claro que la LC ha cambiado radicalmente el panorama de los estudios lingüísticos. En parte por su propia evolución interna, pero en buena parte también por la continua mejora en las condiciones de trabajo facilitadas por las computadoras² y, de otro lado, por las facilidades que derivan de la existencia de Internet y la enorme masa de información existente en la red. La lingüística española no podía ser una excepción y creo que podemos decir que, aunque se incorporó a esta corriente con un cierto retraso, hemos hecho entre todos un trabajo de calidad más que apreciable, que sitúa a los corpus de español en un nivel no muy distante del que ocupan las lenguas que están en las primeras posiciones, especialmente en el ámbito de los corpus diacrónicos.³ En mi opinión, quizá hemos construido demasiados corpus de intención similar, no hemos coordinado esfuerzos adecuadamente y todavía no hemos profundizado en ellos ni los hemos explotado todo lo que debiéramos.

Los corpus de referencia

A pesar de su corta historia –más reducida todavía en la lingüística hispánica, como acabo de señalar– el trabajo de construcción de corpus lingüísticos (y también los modos de recuperar y estudiar su contenido) ha experimentado fortísimos cambios en sus características y, como consecuencia de ello, lo que debe hacer en este momento un equipo que pretenda montar un corpus se parece muy poco a lo que se describe en los (escasos) trabajos que narran la historia de esta aproximación. En unos cuarenta años hemos pasado de corpus pequeños, que había que componer manualmente en las computadoras, con ciertos trucos para los escasos caracteres especiales de la ortografía del español, a las inmensas posibilidades que la red ofrece hoy para poder construir, con muy poco esfuerzo, un corpus de textos periodísticos con materiales actuales, procedentes de todos los países hispánicos. El uso de escáneres y de los primeros reconocedores ópticos de caracteres, que tanto tiempo y dinero consumían en la época intermedia (digamos, hasta finales de los años noventa) son casi historia en este momento. Aunque, por supuesto, la disponibilidad de formatos digitales es mucho mayor en cualquier tipo de texto, el mayor cambio que me proporcionan mis recuerdos de las primeras fases del CREA está relacionado con la prensa digital. Los textos periodísticos han pasado de ser el componente que mayores esfuerzos consumía, tanto para la recogida de materiales (más difíciles o poco menos que imposibles para los alejados en el tiempo y en el espacio) como para su conversión a formato digital, a la fuente más cómoda de captación de materiales y construcción de corpus muy variados en temática y procedencia. También ha habido cambios importantes, pero mucho menos marcados, en los textos orales, que siguen dando muchísimo trabajo en la transcripción y la codificación. En pocas palabras, la situación ha evolucionado de forma muy rápida y siempre en sentido positivo desde la necesidad de dedicar una enorme cantidad de esfuerzos a las tareas mecánicas de conversión a formato digital a una situación en la que, sin duda, hay que invertir tiempo

en estas labores, pero mucho menos en los corpus que buscan una construcción general representativa y equilibrada y casi nada en los llamados 'corpus oportunistas', que se forman integrando directamente textos disponibles en la red.

De otra parte, esa etapa intermedia, en la que tanto tiempo había que dedicar a las tareas mecánicas, se caracteriza también por un peso excesivo de la codificación y, muy especialmente, de la codificación que pretendía convertirse en un estándar para la construcción de corpus. Entiéndaseme bien. Creo que el SGML primero y el XML más tarde constituyen la mejor forma de conseguir que nuestros textos estén codificados de forma correcta y, además, independiente de los fabricantes de las diversas aplicaciones que tenemos que utilizar, con lo que tenemos cierta garantía de conservación del trabajo realizado aunque haya cambios en los sistemas, desaparezcan las aplicaciones o quiebren las empresas. El camino es, sin duda, un lenguaje general de codificación de la información. Pero otra cosa son los pretendidos estándares a los que todos pretendíamos aproximarnos en el más alto grado posible. Contemplado ya a cierta distancia temporal, me parece claro que una cosa es la codificación en un lenguaje como XML —absolutamente imprescindible—, otra ajustarse a un estándar reconocido —también necesaria— y una tercera el exceso hipercodificador establecido en algunos de los que han sido propuestos como estándares. La codificación muy detallada puede ser importante cuando se va a trabajar sobre un texto o un conjunto reducido de textos en los que quizá resulte de interés, por ejemplo, diferenciar el léxico o las construcciones sintácticas de los diferentes personajes de una obra de teatro o la frecuencia de ciertas formas en una parte de un texto para tratar de resolver algún problema de autoría. Parece claro, sin embargo, que las consultas dirigidas a un corpus general, lo que se llama habitualmente un corpus de referencia, no van en esa dirección y, por tanto, no tiene sentido invertir tiempo y dinero en codificar aspectos que no van a ser luego utilizados en las búsquedas. La experiencia muestra que los factores discriminadores que necesitan quienes consultan un corpus se centran en los parámetros utilizados habitualmente en su construcción: época, tipo de texto, área temática, país, así como situación comunicativa, sexo y edad de los hablantes, si se trata de textos orales. Por otro lado, la hipercodificación aplicada a conjuntos textuales de gran tamaño tropieza con el problema de la dificultad para recuperar luego la información en formas y tiempos razonables, con lo que, con cierta frecuencia, nos hemos encontrado con que una buena parte de los esfuerzos invertidos en codificar los textos no servían de nada en tanto que luego no permitían identificar los fragmentos que presentaban las características buscadas.

Diseñar, construir y preparar para su explotación abierta un corpus de tamaño medio (en este momento, algunos cientos de millones de formas) supone actualmente situarse entre dos corrientes distintas, contrarias entre sí, que se han venido desarrollando en los últimos años y que conviene analizar: la línea conocida como *Web as Corpus* por un lado y la que se ocupa del desarrollo de corpus pequeños y centrados en un determinado tipo de texto por otro.

La primera de ellas postula la inconveniencia de dedicar tiempo y dinero a organizar grandes corpus textuales, dado que hay cientos de millones de documentos en la red que se pueden recuperar mediante la utilización de los buscadores comerciales. No debe confundirse con la utilización de los materiales existentes en la red para construir corpus de las características habituales, sino, insisto, en la utilización directa de lo que hay en la red. Es una consecuencia perfectamente lógica y comprensible del enorme desarrollo que ha tenido Internet en los últimos veinte años. Por reiterar datos conocidos, los ingenieros de Google calcularon a mediados de 2008 que la parte pública de la red había alcanzado ya la cifra de un billón (10^{12}) de páginas únicas. Pues bien, si el 2% de esas páginas están en español (20 000 millones) y contienen una media de 200 palabras por página (menos de medio folio en la presentación habitual), la red alberga un corpus potencial del español de aproximadamente 4 billones (10^{12}) de palabras en páginas no repetidas.⁴ Además, se añade, esa enorme cantidad de materiales se renueva continuamente, de forma que el conjunto no envejece, como sucede en cambio con los corpus de referencia, aunque los documentos antiguos sigan presentes en la red.

La riqueza y variedad de lo que hay en la red es innegable y sería realmente absurdo no utilizarla como fuente de datos en los casos adecuados (por ejemplo, elementos de muy baja frecuencia, difíciles de documentar de modo satisfactorio en los corpus habituales, determinación de grandes tendencias en variantes ortográficas, léxicas, etc.). La mejor prueba del interés del uso de los materiales presentes en la red para la investigación lingüística está en la existencia de programas como WebCorp o KwicFinder, que recurren a los servicios de alguno(s) de los buscadores comerciales para obtener los datos, pero presentan luego los resultados en las formas habituales en lingüística, como líneas de concordancias, por ejemplo, en lugar de limitarse a dar los enlaces a las páginas en las que aparece la expresión buscada.⁵

Sin embargo, no en todos los tipos de investigación lingüística el recurso a los materiales de la red muestra ventaja frente a los corpus de referencia. La causa fundamental de ello es, simplemente, que no podemos hacer recuperación selectiva de la información más que en la escasa medida en que lo permiten los buscadores comerciales, que resulta claramente insuficiente para la mayor parte de nuestros propósitos. Los buscadores han sido construidos con un propósito determinado y funcionan del modo adecuado al logro de esos objetivos. No entra en ellos la utilización de expresiones regulares,

los operadores de distancia ni la selección de los resultados en función del tipo de texto, el tema, la época, etc. El paso de los enlaces a las páginas pertinentes a la zona concreta de la página que interesa a quien hace la consulta se puede solucionar con programas de intermediación del tipo de KwicFinder o WebCorp, aunque siempre estarán condicionados a las facilidades que proporcione el motor. El problema de la selección de los textos de interés, en cambio, no puede ser solucionado por los buscadores comerciales, que no han sido diseñados con ese objetivo.⁶

Por otro lado, a pesar de la enorme riqueza de los materiales que hay en la red a nuestra disposición, en cierto sentido son más pobres que los que podemos encontrar en un corpus. Habitualmente, en las consultas abiertas los corpus facilitan únicamente pequeños fragmentos de texto que contiene la expresión buscada, pero no permiten la descarga del texto completo. Son muchos los propietarios de textos que no tienen inconveniente en una extracción de este tipo, claramente dirigida a ilustrar el uso de una expresión lingüística, pero que, en cambio, no permitirían la descarga del texto completo. Como consecuencia, utilizando los buscadores comerciales no resulta posible acceder a una enorme cantidad de materiales de gran importancia para el análisis lingüístico. Ese defecto se agrava si pensamos en lo que se necesita para investigaciones de tipo diacrónico: hay pocos textos de épocas anteriores a la actual en la parte pública de la red. Curiosamente, esos pocos textos pueden en cambio suponer un problema para quienes están interesados en la lengua actual, puesto que los documentos recuperados van a aparecer mezclados y dispuestos en un orden sobre el que no tiene el menor control quien hace la consulta.

Otra dificultad añadida procede del hecho de que Internet está cambiando continuamente, lo cual significa que los datos se modifican sin cesar, de modo que no es posible contrastar adecuadamente las búsquedas ajenas ni repetir las propias para comprobar los resultados.⁷

En definitiva, el gran problema es que nuestras consultas a lo contenido en la red no van a poder filtrar los documentos que nos interesan por características que no sean las que utiliza el buscador empleado, habitualmente lengua y país, en ambos casos con un notable grado de incertidumbre. Y aquí no cabe esperar a la aparición de un buscador diseñado para la investigación lingüística, puesto que la recuperación por época, tipo de texto, área temática, etc. será imposible en tanto que la información necesaria no está codificada en los documentos, que han sido incorporados a la red con propósitos claramente distintos.

He aludido hace un momento a la diferencia, crucial, entre la tendencia conocida como *Web as corpus* y la posibilidad de utilizar las ventajas de la red para construir un corpus. En este último sentido, la recolección de materiales es solo una parte del proceso y, por tanto, lo que realmente cuenta es lo que se haga después con los documentos incorporados por cualquiera de las vías que estén a disposición de los constructores. De todas formas, creo que conviene también tener en cuenta que existe la posibilidad de reunir grandes cantidades de textos existentes en la red (periodísticos, por ejemplo), agruparlos en una máquina con memoria suficiente y procesarlos superficialmente para obtener de ellos información estadística, con o sin aplicación previa de algún programa de anotación morfosintáctica automática. La más que evidente utilidad de recursos de este tipo no los convierte en corpus textuales en el sentido más habitual de la expresión. Un corpus en sentido estricto debería incluir, como rasgos imprescindibles, la posibilidad de recuperar un contexto suficiente y la de aplicar algún filtro selectivo sobre el tipo de texto, su procedencia, la época, el área temática, etc.

La segunda línea de presión sobre los corpus de referencia es la constituida por los corpus especializados en tipos concretos de textos, habitualmente de tamaño pequeño o incluso muy pequeño. Aunque no es un rasgo inherente a esta línea, lo cierto es que en la mayor parte de los casos se trata de corpus históricos y los constructores son también editores de los textos integrados en ellos, lo cual da el carácter diferencial que nos permite marcar ciertas distancias. Esta línea de trabajo⁸ es de un enorme interés, imprescindible en la investigación histórica del español, que ha estado limitada durante demasiado tiempo a lo que se puede observar únicamente en un subconjunto documental relativamente reducido en tamaño y, sobre todo, concentrado en textos de creación, ciertos autores o incluso ciertas obras. Pero estos otros proyectos se mueven casi siempre en un marco diferente y con unos objetivos distintos, al menos parcialmente. El hecho de que haya que empezar por hacer la edición de los textos limita considerablemente el tamaño que pueden alcanzar estos corpus, pero, en cambio, al no plantear el problema de los derechos de autor ni de distribución, permiten habitualmente el acceso íntegro a los documentos incluidos. Dado que son relativamente pequeños y tienen objetivos más amplios que la investigación puramente lingüística, pueden presentar también la imagen del documento, la transcripción paleográfica, semipaleográfica y crítica, permitir la comparación de copias diferentes, etc., posibilidades todas ellas vedadas en un corpus de referencia, que, como reiteraré dentro de un momento, es, por principio, un trabajo de segundo nivel, que vive de la integración, adaptada, de lo que otros han hecho previamente.

Creo que ayuda a poner las cosas en su sitio considerar que los resultados de estos proyectos son corpus textuales únicamente en el sentido más genérico de la expresión. En realidad, constituyen magníficos ejemplos de lo que puede ser, en los tiempos actuales, la edición crítica de textos, que aprovecha al máximo (o debería hacerlo) las posibilidades que brindan los recursos computacionales que tenemos a nuestra disposición.

Conectado con lo anterior, pero en una línea distinta, están los corpus que presentan un alto grado de especialización en, por ejemplo, el tipo de anotación incorporada. Para usar el modelo clásico, al que volveré dentro de un momento, la investigación o la construcción de un recurso puede crecer en extensión o en profundidad. Dadas las limitaciones habituales en tiempo y medios, es forzoso optar por privilegiar una de ellas, lo cual se convierte irremediabilmente en la preterición de la otra. Los dos son importantes, sin embargo, y su mayor o menor relevancia es más el resultado de una situación o necesidad concreta, que se presenta en un cierto momento, que de una cuestión general. Se puede optar por construir un corpus grande, bien diseñado, equilibrado y con todo lo necesario para hacer recuperación selectiva de la información, pero sin anotación sintáctica, o bien por hacer un corpus mucho más pequeño, bien construido y con anotación sintáctica y semántica. Conseguir las dos cosas al tiempo es mucho más difícil, por no decir imposible.

En resumen, un corpus de referencia es un tipo de recurso con ciertas limitaciones, que no existen en cambio en otros recursos lingüísticos vinculados. Como contrapartida, reúne características que proporcionan ventajas sobre todos los demás en diferentes aspectos. Afortunadamente, la necesidad de optar por una vía u otra es algo que afecta únicamente al equipo constructor del recurso. Quienes los utilizan pueden tomar en cada caso lo que les resulte más ventajoso, más acorde a sus necesidades.

Características de la LC en la actualidad

Como resumen de lo analizado hasta este momento, podemos decir que la lingüística basada en el análisis de corpus es la vía 'natural' de acceso al estudio de los fenómenos lingüísticos para quienes defienden una aproximación empírica en la que los objetivos generales no se limiten a la mera descripción de lo que se encuentra en un corpus más o menos amplio, sino que crean que los datos externos a cuyo conocimiento podemos acceder son el punto de partida y el contraste final para intentar explicar lo que sucede en una variedad lingüística determinada. Se diferencia, por tanto, de la lingüística racionalista en sus formulaciones clásicas, pero también de la lingüística descriptiva tradicional. Por supuesto, no se opone a la lingüística teórica, sino que le proporciona la base empírica necesaria para la formulación de hipótesis y el contraste de sus deducciones. Evidentemente, no es una disciplina, pero creo que tampoco puede ser considerada como una metodología en el sentido más habitual de esa expresión. Es una forma diferente de afrontar el estudio de los fenómenos lingüísticos,⁹ compatible en las fases posteriores con diferentes asunciones acerca de lo que puede ser relevante. Según Gries (2006a: 4), la mayor parte de los trabajos sobre lingüística de corpus presentan las características siguientes:

- El análisis se basa en uno o varios corpus de textos producidos de forma natural, en formato electrónico, de modo que la recuperación de datos se hace automáticamente.
- El corpus es (o se supone que es) representativo y equilibrado con respecto a la modalidad lingüística con la que se va a trabajar.
- El análisis es (o intenta ser) sistemático y exhaustivo, lo cual significa que el corpus no es simplemente una base de datos de la cual se extraen unos cuantos ejemplos y se rechazan otros, sino que la totalidad del corpus es tomada en consideración.
- El análisis pretende ir más allá de un simple 'o...o'. Utiliza datos estadísticos para tratar de cubrir también los casos intermedios entre los valores extremos.
- El análisis se realiza sobre la base de listas de frecuencias (de palabras, formas, esquemas gramaticales, etc.), concordancias y colocaciones.

Desde este marco general, veamos, para terminar, las características más importantes de la situación actual de la LC, con especial atención al ámbito hispánico. En primer término, no se debe olvidar que un corpus general, un corpus de referencia es, por naturaleza, un trabajo que podemos considerar de segundo nivel, esto es, un recurso que depende de la labor realizada previamente en muy diversos aspectos. Quien construye un corpus histórico, por ejemplo, depende de los textos que hayan sido editados con anterioridad, tanto en lo que se refiere a los propios textos como a las convenciones gráficas adoptadas. En otras palabras, no se puede pedir a quien pretende producir un corpus general, formado por miles de textos y cientos de millones de palabras, que acometa también la reedición de textos ya publicados, edite los que no lo han sido, solucione problemas textuales, resuelva incógnitas de autoría, fecha de composición, etc. El corpus histórico general refleja el saber común acerca de los textos y lo utiliza para facilitar recursos de investigación, pero no puede, en sí mismo, ir más allá de lo comúnmente aceptado en el conocimiento de los textos ni de sus autores. De modo similar, en un ámbito totalmente distinto, el proceso de anotación gramatical y lematización automáticas no es el lugar en el que se puede dar solución a problemas que la investigación gramatical no ha resuelto con acuerdo general. Los programas que realizan esa tarea reflejan, con un porcentaje mayor o menor de fallos, lo que es conocimiento común o bien, en caso de discrepancia, uno de los enfoques posibles para una cuestión polémica. La obligación del equipo constructor de corpus se reduce en este punto a hacer explícitos los criterios que se han adoptado y a aplicarlos de forma constante y congruente.

En segundo lugar, la bibliografía clásica repite continuamente la afirmación de que un corpus debe ser representativo y estar equilibrado. La argumentación habitual sobre la representatividad es muy clara: sabemos que en un corpus no puede estar todo y, por tanto, hemos de concebirlo como una muestra de la población de la que ha sido extraído y, en consecuencia, debe ser representativo en el sentido habitual en estudios estadísticos.

Como es bien sabido, una muestra tiene que ser representativa de la población a la que corresponde, puesto que solo de este modo tendremos las garantías esperables a la hora de proyectar a la población general los resultados obtenidos en el análisis de la muestra. En estudios de población, el carácter representativo de la muestra se establece en función de su composición con respecto a parámetros que se consideran relevantes para los fenómenos que se van a estudiar. Los estudios sociológicos, las encuestas de opinión política, por ejemplo, toman en cuenta la distribución geográfica, económica, sociocultural, el sexo, la edad y los factores derivados de los anteriores o conectados con ellos. Por otro lado, hay acuerdo general en que las muestras estadísticas son adecuadas para conocer tendencias, obtener medias, hacer valoraciones generales, etc., pero no constituyen, en cambio, el camino adecuado para estudiar fenómenos raros o casos individuales.

¿Cómo transferir la técnica de construcción de muestras a la distribución de los textos que van a constituir un cierto corpus? La bibliografía está llena de indicaciones en las que los responsables del diseño de los corpus pretenden dar aire 'objetivo' a la agrupación de textos que presentan. Sin embargo, el diseño responde a factores diversos, que empiezan por la mayor o menor facilidad para disponer de los materiales en formato electrónico o el coste que supone introducirlos (prensa de época anterior a los últimos diez años, por ejemplo) y terminan en necesidades específicas. En la misma línea se encuentra la introducción de textos orales, que nunca superan el 10% del total en un corpus general, a pesar de que lo que sabemos del comportamiento de las comunidades lingüísticas nos llevaría a dar a lo oral el peso mayoritario.

¿Qué peso habría que dar a cada época, a cada país, a cada tipo de texto para que el corpus resultante fuese representativo? ¿Cuál debería ser la relación entre el tamaño de la muestra y el tamaño de la población en ella representada? Es poco más que un ejercicio académico, pero ya que resulta imposible hacer el cálculo de la producción oral de los hispanohablantes en un año, digamos, al menos, que las estadísticas oficiales indican que en el año 2004 se publicaron en España algo más de 60 000 títulos. Descontando las 14 500 reimpresiones y teniendo en cuenta que solo el 78,6% de los títulos estaban escritos en español, nos quedan unos 35 700 títulos. Si, en un cálculo prudente, suponemos una media de 20 000 palabras por título,¹⁰ llegamos a la conclusión de que en 2004 los libros y folletos publicados en España suponen alrededor de 714 millones de formas. A ellas habrá que sumar, claro, todo lo publicado en todos los demás países de habla hispana y la producción escrita no incluida en la estadística anterior, fundamentalmente la prensa. ¿Qué significan, proyectados sobre ese fondo, los 7,5 millones de formas procedentes de textos editados o producidos ese año que contiene el CREA o las 500 000 formas que podrían corresponderle en una hipotética ampliación del corpus del español construido por Mark Davies? Evidentemente, no mucho. Y esa enorme desproporción entre población y muestra y, sobre todo, el desconocimiento de las características de la población, es el escenario en que hemos de situar el debate acerca de la representatividad de nuestros corpus.

En realidad, el problema de la representatividad es en buena parte una consecuencia del tamaño reducido que poseían los primeros corpus de trabajo (un millón de palabras en el formato inicial) y, sobre todo, de la imposibilidad de trabajar con subcorpus establecidos según las necesidades de una investigación concreta. El primero de los factores, el tamaño reducido, se relaciona también con la distinción clásica entre corpus de fragmentos y corpus de textos completos. Es evidente que un corpus de propósito general constituido por un millón o diez millones de formas no debería contener textos completos, puesto que eso reduce las características de la muestra y, por tanto, recorta gravemente su carácter representativo, la posibilidad de examinar textos de carácter distinto, etc. Por otro lado, no se ve claro cómo hacer la selección que nos lleve a una muestra representativa de lo que es una variedad lingüística en un momento determinado de su historia. ¿Los libros más vendidos? ¿Los autores más reconocidos? ¿Los periódicos de más difusión? ¿Los programas de radio o televisión con mayores índices de audiencia? Todos esos problemas desaparecen o, al menos, quedan muy reducidos en importancia cuando pasamos a cientos de millones de formas, que es el tamaño con el que trabajamos en la actualidad.

Todavía más importante es el segundo factor, que tiene implicaciones más profundas. Normalmente, en una investigación lingüística no interesa la frecuencia absoluta de un fenómeno, sino su frecuencia relativa, tanto en lo que respecta al volumen general como a los posibles subcorpus. Esto es, nos interesa la frecuencia normalizada de un fenómeno (apariciones por millón de formas, por ejemplo) en diferentes tipos o subtipos de texto. Eso es lo que nos permite ver en qué difieren textos de distintos tipos, países o épocas con independencia del volumen de cada uno de ellos. Adelantando algo que veremos más tarde, ese es el factor que justifica la enorme importancia que tiene el hecho de que el sistema de consultas esté construido de modo que se puedan obtener las combinaciones de características que resulten adecuadas según la investigación. En otras palabras,

no por siglos para periodizar la historia del español. Contemplado desde esta perspectiva, el problema de la representatividad se diluye en el aspecto cuantitativo y se traduce en la necesidad de que el corpus contenga, en cantidad suficiente, textos de diferentes tipos y subtipos, de modo que la comparación de datos de diferentes subcorpus pueda realizarse con garantías. La posibilidad de obtener de forma sencilla y cómoda los resultados correspondientes a diferentes subconjuntos de textos –por ejemplo, noticias de economía publicadas en periódicos argentinos entre 1998 y 2004 frente a libros sobre economía (o cualquier otro tema) publicados en Argentina (o cualquier otro país) en ese período (o cualquier otro)— se convierte en un rasgo fundamental, que da a la consulta del corpus todo su sentido, puesto que es la que permite realmente afinar la comparación de los resultados obtenidos.

En la línea habitual de evolución de las corrientes metodológicas, la lingüística de corpus ha experimentado progresos importantes tanto en extensión como en profundidad, para usar la estructuración tradicional. El progreso en extensión se refleja, por supuesto, en el enorme incremento del tamaño de los corpus, situados ahora con mucha frecuencia en cientos de millones o incluso miles de millones de formas. El aumento de tamaño no es un factor exclusivamente cuantitativo, como ya he indicado: la mayor parte de los problemas señalados tradicionalmente para el análisis de los resultados de un corpus desaparecen cuando podemos trabajar con doscientos o trescientos millones de formas en un conjunto bien equilibrado. Por supuesto, los elementos o fenómenos de muy baja frecuencia seguirán presentando problemas para su documentación adecuada, pero para esos casos siempre podemos recurrir a la masa informe de datos contenidos en la red.

El aumento del tamaño de los corpus ha sido posible gracias a los avances en la capacidad de las máquinas y el aumento de su velocidad de proceso, pero también como consecuencia de los cambios que esos avances han provocado en la gestión de los contenidos digitales. Hace apenas unos años, los constructores de corpus tenían que dedicar la mayor parte de su tiempo a la conversión a formato digital de los textos previamente publicados, a su (re)codificación, etc. Estas tareas suponen hoy un esfuerzo considerablemente menor, de modo que es posible hacer corpus más voluminosos y también mejor estructurados.

Entramos así en la otra dimensión: el crecimiento en profundidad. El viejo modelo del *Brown corpus*, constituido por un millón de formas, analizable solo en conjunto y sin apenas codificación asociada a los textos está muy alejado de lo que tenemos hoy a nuestra disposición. Son corpus de tamaño mucho mayor, normalmente codificados de forma que se puede hacer recuperación selectiva del subcorpus que nos interesa y con mucha frecuencia enriquecidos con anotación morfológica y sintáctica (con menor frecuencia, semántica y pragmática). En paralelismo, los estudios basados en corpus han dejado de referirse casi exclusivamente a cuestiones de frecuencias de elementos léxicos y se extienden hoy a lo largo y ancho de los componentes lingüísticos. La evolución de las técnicas computacionales ha facilitado enormemente lo que se puede añadir a un texto, de modo que quienes trabajan en el análisis de la lengua oral pueden oír el sonido, ver la imagen asociada a él, incorporar a esos dos componentes, en niveles sucesivos que se pueden mezclar o aislar, la transcripción fonética, la fonológica, una transcripción ortográfica a la que se puede conectar el resultado de la anotación morfológica, sintáctica, semántica y pragmática. Así pues, lo tenemos todo, los diferentes aspectos están adecuadamente vinculados, pero podemos observar únicamente lo que nos interesa en cada caso.

Estos cambios son especialmente importantes para quienes, como la mayoría de los que estamos aquí, se dedican a la gramática, tanto en orientación sincrónica como diacrónica. Hace ya muchos años que los datos extraídos de los corpus han dejado de servir únicamente para hacer listas de frecuencias léxicas o estudiar los valores que una expresión presenta en los usos reales. En la posibilidad de usar corpus textuales en los estudios sintácticos, ampliamente utilizada en este momento, han influido varios factores, además de los aspectos generales, derivados de la extensión de este enfoque y la demostración repetida de su validez metodológica. El primero de ellos ha sido mencionado ya: la existencia de corpus amplios, con anotación gramatical de diferentes tipos y profundidad, permite obtener cómodamente datos seguros sobre fenómenos de los que antes no se podían hacer recuperaciones automáticas o eran posibles solo mediante el recurso a consultas reiterativas basadas en las características ortográficas. Son muchos los corpus en los que hoy cabe hacer consultas que manejan, por ejemplo, la clase de palabras o ciertos valores de las categorías gramaticales, de modo que se hace posible el acceso a secuencias definidas por sus características más abstractas. De todas formas, dado que, como reza uno de los postulados de Leech, toda anotación supone una interpretación, manejar estos datos supone siempre los riesgos que derivan de las posibles discrepancias con la anotación añadida y también el mucho más frecuente de los inevitables errores de un proceso automático de anotación. De otra parte, como acaba de señalar Schulte (2009), los corpus diacrónicos bien anotados y con posibilidades reales de recuperar información de forma selectiva (esto es, por tipos de texto, países, tramos temporales, etc.) no abundan todavía, de modo que, en estos casos, es necesario familiarizarse con la utilización de, por ejemplo, expresiones regulares en las consultas, con lo que la recuperación de los casos interesantes se hace mucho más cómoda y rápida.

El segundo gran factor se relaciona con la reconsideración del papel de la frecuencia en una buena parte de los estudios gramaticales. Sin entrar en el núcleo de la cuestión, que nos llevaría demasiado lejos, la frecuencia ha pasado de ser un factor al que apenas se prestaba atención o cuya importancia para los estudios gramaticales era negada, a convertirse en un aspecto central, tanto de los fenómenos morfológicos y sintácticos como de su evolución a través del tiempo, en tanto que factor básico en el cambio lingüístico. Naturalmente, la frecuencia es indesligable del trabajo con corpus, puesto que solo el conocimiento de lo que sucede en la totalidad de un corpus amplio y debidamente equilibrado puede servir de base para lo que necesitamos saber. Por supuesto, la lingüística descriptiva tradicional trabajaba con frecuencias, pero el ya mencionado filtro inevitable las reducía a unos pocos casos (sesgados) de muchos textos o bien a todos los casos de pocos o muy pocos textos. Ahora, en cambio, disponemos de datos en abundancia y creo que eso nos va a permitir reescribir varios capítulos de la gramática española y de la historia de la lengua. Es cierto que los estudios sintácticos requieren en muchos casos un nivel de anotación que no suele encontrarse en corpus de gran tamaño o bien la utilización de bases de datos sintácticos, construidas con este propósito específico, como la BDS.¹¹ Sin embargo, es claro también que toda la línea de trabajo que se mueve en la consideración de una relación estrecha entre léxico y gramática tiene grandes posibilidades de análisis de datos en corpus que no pasan de la llamada anotación morfosintáctica, como muestran, entre otros muchos, los análisis de Gries y Stefanowitsch (cf. para una perspectiva de conjunto, Stefanowitsch & Gries, 2009). Por cierto, que esta línea de investigación muestra que existe una cierta deficiencia en la formación estadística de la mayoría de quienes trabajamos en lingüística, como ha señalado el propio Gries en varias ocasiones (cfr. p.e., Gries, 2006b). De otra parte, es claro que lo que nos interesa es la frecuencia relativa de los fenómenos estudiados, con lo que, de nuevo, se pone de relieve la necesidad de que el sistema de explotación del corpus permita la obtención de datos construyendo los subcorpus que las características del fenómeno estudiado hagan relevantes.

Termino ya. Evidentemente, los corpus textuales no son más que un magnífico recurso que pone en nuestras manos, con gran rapidez y comodidad, grandes cantidades de información disponible sobre el fenómeno que nos interesa. Nótese que he dicho **información**, que, sin entrar ahora en un terreno excesivamente complejo, se opone a **conocimiento** como lo informe frente a lo estructurado, lo externo a lo interiorizado. Con palabras de Mario Bunge (2003), "para transformarse en conocimiento, la información debe ser evaluada como verdadera o falsa, pertinente o irrelevante, práctica o impráctica, interesante o tediosa". Nuestro trabajo, que es precisamente la construcción de conocimiento lingüístico, no es ahora más fácil, pero sí está mejor fundamentado que antes gracias al empleo de corpus textuales.

Notas

1 La diferencia "between anecdotal and corpus data has not to do with the type of data involved, but with the manner of collecting them" (ibidem).

2 Hace ya muchos años, Svensén (1993: 250) explicaba el enorme éxito de la informática en la preparación y confección de diccionarios indicando que la computadora había continuado evolucionando en sentido favorable, "becoming even faster, smaller yet more capacious, and cheaper in relation to what it can do".

3 Para algunas cifras comparativas, permanentemente revisables, cf. Rojo (en prensa). La situación está a punto de cambiar, ya que se anuncia para la segunda mitad de 2010 la disponibilidad (en pruebas) del Corpus of Historical American English (COCA), con unos cuatrocientos millones de formas desde 1800 hasta la actualidad, construido por Mark Davies. Cf. <http://view.byu.edu>.

4 Estos cálculos son intrínsecamente inseguros. Hay que tener en cuenta que la web indexada es muy inferior a las cifras que se dan habitualmente para su volumen total y que, además, cada buscador indexa conjuntos parcialmente diferentes. En octubre de 2009, Google identificaba unos 700 millones de páginas en español. Si suponemos el mismo número medio de palabras por página que se calcula en el texto, el resultado sería un corpus de 140 000 millones de palabras, que sigue siendo enorme. Cf. Rojo y Sánchez (2010: caps. 2 y 3) para más detalles sobre el tamaño de Internet y el número de páginas en español.

5 Sobre esta tendencia pueden verse, entre otros trabajos, la introducción al volumen especial que Computational Linguistics dedicó recientemente al tema (Kilgarrif & Grefenstette, 2003), así como la de Hundt, Nesselfauf y Biewer (2007b) a un volumen colectivo sobre el mismo tema. Interesante visión de conjunto es la que se muestra en Lüdeling, Evert & Baroni (2007). Sobre WebCorp, vid., por ejemplo, Renouf, Kehoe & Banerjee (2007). Sobre KwicFinder, vid. Fletcher (2007).

6 Para la enumeración de las características que habría que exigir a un buscador realmente útil en la investigación lingüística, vid. Volk (2002).

7 Sinclair (2005: 15) diferencia radicalmente entre la WWW y los corpus:

The World Wide Web is not a corpus, because its dimensions are unknown and constantly changing, and because it has not been designed from a linguistic perspective. At present it is quite mysterious, because the search engines, through which the retrieval programs operate, are all different, none of them are comprehensive, and it is not at all clear what population is being sampled. Nevertheless, the WWW is a remarkable new resource for any worker in language [...] and we will come to understand how to make best use of it.

Como han señalado Kilgarrif y Grefenstette (2003: 333-334), debe evitarse la confusión entre una pregunta sobre si tal conjunto es un corpus y otra acerca de si ese conjunto es un corpus adecuado para un determinado propósito. Naturalmente, es necesario examinar la definición de corpus que se utiliza en cada caso.

8 Representada, por ejemplo, por el proyecto CODEA, dirigido por Pedro Sánchez-Prieto (<http://www.textoshispanicos.es/index.php>) o el de edición de las Biblias medievales, dirigido por Andrés Enrique-Arias (<http://corpus.byu.edu/bibliamedieval/>).

9 Para Leech (1992: 106), la LC "defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject". Gries (2009: 1) la considera "a method(ology), no more, but also not less", aunque no cree que "this difference would result in many practical differences".

10 Para que se pueda apreciar que el cálculo está hecho muy a la baja, téngase en cuenta que el texto de la Crónica de una muerte anunciada tiene 27 960 palabras en total.

11 Base de datos sintácticos del español (<http://www.bds.usc.es>).

Referencias

- Aarts, Jan (2000): "Towards a new generation of corpus-based English grammars", en Lewandowska-Tomaszczyk & P. Melia (eds.): PALC '99. *Practical applications in language corpora. Papers from the international conference at the university of Lodz*. Frankfurt am Main: Lang, 17-36.
- Aarts, Jan (2002): "Does corpus linguistics exist? Some old and new issues". En Breivik, Leiv Egil & Angela Hasselgren (eds.): *Language and Computers. From the COLT's mouth... and others*. Amsterdam: Rodopi, 2002. 1-17. Cito por su reed. en Teubert, Wolfgang & Ramesh Krishnamurthy (eds.): *Corpus Linguistics*. Londres: Routledge, 2007, 58-73.
- Bunge, Mario (2003): "Información + Evaluación = Conocimiento", en *Pliegos de Yuste*, 1, 2003. [<http://www.pliegosdeyuste.eu/n1pliegos/mbunge.pdf>].
- Dyson, Freeman (1997). *Imagined Worlds*. Harvard University Press, 1997. Cito por la trad. esp. de Joandomènec Ros: *Mundos del futuro*. Barcelona: Crítica, 1998.
- Dyson, Freeman (1999) : *The Sun, the Genoma, the Internet*. Oxford Univ. Press, 1999. Cito por la trad. esp. de Juan Manuel Ibeas, *El sol, el genoma e Internet. Las tres cosas que revolucionarán el siglo xxi: la energía solar, la ingeniería genética y la comunicación mundial*. Madrid: Debate: 2000.
- Fillmore, Charles J. (1992): " 'Corpus linguistics' or 'Computer-aided armchair linguistics' ", en Svartvik (1992), 35-60.
- Fillmore, Charles J. (2001): "'Armchair linguistics vs. corpus linguistics revisited'", en De Cok, S., G. Gilquien, S. Granger & S. Petch-Tyson (eds.): *ICAME 2001. Future challenges of corpus linguistics. Proceedings of the 22nd ICAME conference*. Louvain-la-Neuve: Université Catholique de Louvain, 1-2.
- Fletcher, William H. (2007): "Concordancing the Web: promise and problems, tools and techniques" en Hundt, Nesselhauf & Biewer (2007) 25-45.
- Gries, Stefan Th. (2006a): "Introduction" a Gries, Stefan Th. & Anatol Stefanowitsch (eds.): *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis*. Berlin: Mouton, 2006, 1-17.
- Gries, Stefan Th. (2006b): "Some Proposals towards a More Rigorous Corpus Linguistics", *Zeitschrift für Anglistik und Amerikanistik*, 54/2, 2006, 191-202.
- Gries, Stefan Th. (2009): "What is Corpus Linguistics", en *Language and Linguistic Compass*, 3, 2009, 1-17.
- Hoffmann, Sebastian (2008): "Looking at language in use: some preliminaries", en Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee & Ylva Berglund Prytz: *Corpus Linguistics with BNCWeb - a Practical Guide*. Frankfurt: Lang, 2008, 1-12.
- Hundt, Marianne, N. Nesselhauf & C. Biewer (2007b): "Corpus Linguistics and the Web", en Hundt, Nesselhauf & Biewer, 2007a, 1-5.
- Hundt, Marianne, N. Nesselhauf & C. Biewer (eds.) (2007a): *Corpus Linguistics and the Web*. Amsterdam / Nueva York: Rodopi, 2007.
- Kilgariff, Adam & G. Grefenstette (2007): "Introduction to the Special Issue on the Web as Corpus", *Computational Linguistics*, 29/3, 2003, 333-347.
- Leech, Geoffrey (1992): "Corpora and theories of linguistic performance", en Svartvik (1992), 105-147.
- Lüdeling, Anke, S. Evert & M. Baroni (2007): "Using web data for linguistic purposes", en Hundt, Nesselhauf & Biewer, 2007a, 7-24.
- Murray, James A. H. (1879a): "An Appeal to the English-Speaking and English-Reading Public to Read Books and Make Extracts for the Philological Society's *New English Dictionary*". Utilizo el facsímil electrónico de la segunda edición de la *appeal* (24/6/1879) que se encuentra en la página del OED: www.oed.com/archive/appeal-1879-06/p1.html [descargado el 29/7/2009].
- Murray, James A. H. (1879b): "Eighth Annual Address of the President to the Philological Society, Delivered at the Anniversary Meeting", *Transactions of the Philological Society*, 1877-79, 561-586.
- Murray, K. M. Elisabeth (1977): *Caught in the Web of Words. James Murray and the Oxford English Dictionary*. Yale University, 1977.
- Quirk, Randolph (1992): "On corpus principles and design", en Svartvik (1992), 457-469.
- Renouf, Antoinette, Andrew Kehoe & Jayeeta Barnerjee (2007): "WebCorp: an integrated system for web text search", en Hundt, Nesselhauf & Biewer (2007), 47-67.
- Restrepo, Félix (1945): "La continuación del Diccionario de construcción y régimen de la lengua castellana", *Thesaurus*, 1/3, 1945, 429-432.
- Rojo, Guillermo (2008): "Lingüística de corpus y lingüística del español", *Actas del XV Congreso de la Asociación de Lingüística y Filología de América Latina*. Montevideo, 2008. Edición en CD.
- Rojo, Guillermo (en prensa): "El papel de los corpus en el estudio de la historia del español", intervención en la mesa redonda sobre Los corpus diacrónicos en la historia de la lengua española celebrada en el VIII Congreso internacional de historia de la lengua española, Universidade de Santiago de Compostela (14-18 de septiembre de 2009). En prensa en las Actas del Congreso. (http://gramatica.usc.es/~grojo/En_prensa/Rojo_Corpus_diacronicos.pdf)
- Rojo, Guillermo y Mercedes Sánchez (2010): *El español en Internet*. Madrid / Barcelona: Fundación Telefónica / Ariel, 2010.
- Schulte, Kim (2009): "Using non annotated diachronic corpora: benefits, methods and limitations", en Enrique-Arias, Andrés (ed.): *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*. Madrid / Frankfurt: Iberoamericana / Vervuert, 2009, 167-180.
- Sinclair, John (2005): "Corpus and Text. Basic Principles", en Wynne (2005), 1-16.
- Stefanowitsch, Anatol and Stefan Th. Gries (2009): "[Corpora and grammar](#)", en Lüdeling, Anke & Merja Kytö (eds.), *Corpus linguistics: an international handbook*. Berlin & New York: Mouton de Gruyter, 2009, vol. 2, 933-951.
- Svartvik, Jan (ed.) (1992): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82 (= Trends in Linguistics. Studies and Monographs, 65)*. Berlin: Mouton - de Gruyter, 1992.
- Svensén, Bo (1993): *Practical Lexicography*. Oxford: Oxford Univ. Press, 1993; es adaptación de la edición original sueca de 1987.

- Volk, Martin (2002): "Using the Web as Corpus for Linguistic Research", en Pajusalu, Renate & Tiit Hennoste (eds.): *Tähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*. Publications of the Department of General Linguistics, 3. Univ. of Tartu, 2002. (http://www.ling.su.se/DaLi/volk/papers/Oim_Festschrift_2002.pdf; comprobado 3/5/2010).
- Wynne, Martin (2005): *Developing Linguistic Corpora. A Guide to Good Practice*. Oxford: Oxbow Books, 2005.