El futuro de los corpus de referencia[1]

The future of reference corpora

Guillermo Rojo
Universidade de Santiago de Compostela (Santiago de Compostela, España) / Real Academia Española (Madrid, España)
guillermo.rojo@usc.es
https://orcid.org/0000-0002-1771-6561

Recibido el 19/10/2023, aceptado el 4/2/2024, publicado el 18/10/2024

Cómo citar este artículo
Rojo, Guillermo 2024. El futuro de los corpus de referencia. *Studia linguistica romanica* 2024.12, 18-33. https://doi.org/10.25364/19.2024.12.2.

Resumen
Los corpus de referencia, situados en medio de un espectro en cuyos márgenes se encuentran los corpus masivos de un lado y los especializados de otro, constituyen un recurso ampliamente utilizado en las investigaciones lingüísticas. En este trabajo se intenta señalar sus características más destacadas e identificar los factores que muestran sus ventajas e inconvenientes frente a los otros tipos de corpus, resaltar aquellos aspectos en los que su influencia ha sido más notable en la lingüística española y, por último, esbozar las que podrían ser las grandes líneas de su evolución en los próximos años.

Abstract
Reference corpora, situated in the middle of a spectrum with massive corpora on one end and specialized corpora on the other, constitute a widely used resource in linguistic research. In this paper we aim to point out their most prominent characteristics, identify the factors that show their advantages and disadvantages compared to other types of corpora, and highlight the aspects in which their influence has been most notable in Spanish linguistics. Finally, we outline the potential main lines of their evolution in the coming years.

1 El texto de este artículo es una reelaboración de la ponencia pronunciada en la sección *Corpus históricos electrónicos de la lengua española. Retos y direcciones futuras* del XXIII Congreso de la Asociación alemana de hispanistas, celebrado en Graz del 22 al 25 de febrero de 2023. Deseo expresar mi agradecimiento a Concepción Company Company y Virginia Bertolotti (coordinadoras de la sección) por la invitación a participar en él y a los revisores anónimos por las sugerencias formuladas sobre una versión previa.

Índice

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

1 Introducción

[1] Incluso sin tener en cuenta los variados antecedentes de los corpus textuales (Rojo 2015), los cambios que han tenido lugar en la historia de lo que hoy consideramos lingüística de corpus (LC) son realmente llamativos e impresionantes. En efecto, desde la concepción y finalización del *Brown corpus*, a mediados de los años 60 del siglo pasado, hasta la actualidad, el panorama se ha modificado radicalmente. Esto explica que a quienes tenemos ya una cierta edad y hemos vivido como experiencia propia todos esos desarrollos nos resulta un tanto difícil de entender la falta de comprensión y valoración que, de forma sin duda justificada, muestran a veces colegas que, por ser (mucho) más jóvenes, tienen una perspectiva limitada a lo sucedido en los 15 o 20 últimos años.

[2] Como es lógico dadas sus características, la evolución de la LC es en gran parte consecuencia de la evolución de las computadoras. Aunque el propio Gordon Moore, formulador de la ley epónima, considera que la vigencia de esta ley cesará en los próximos años (por razones físicas de espacio), lo cierto es que, hasta el momento, los microprocesadores siguen duplicando su capacidad cada poco tiempo. Como consecuencia de ello, las computadoras tienen cada vez más velocidad, más potencia y mayor capacidad, al tiempo que van reduciendo su precio. El resultado de todo ello es que el tamaño (un millón de palabras) y la codificación (casi inexistente) del *Brown corpus* se nos aparecen hoy como algo enternecedor al lado de los cientos o miles de millones que alcanzan el *CdE Web/Dialects*, el *esTenTen* o los inmensos agregados de textos que se construyen y utilizan en tareas vinculadas al componente de procesamiento de lenguaje natural en proyectos de inteligencia artificial.

[3] Sin mengua de la importancia crucial de este factor, la comprensión correcta de la LC requiere que tomemos en cuenta también aspectos procedentes de dos ámbitos distintos. Por una parte, la evolución del mundo de la computación en aspectos diferentes de la velocidad y la capacidad. En este ámbito hay que destacar, en primer lugar, la aparición de las computadoras personales. Estos aparatos pusieron una parte importante de las posibilidades de los recursos informáticos al alcance de personas sin formación especial en ciencias o técnicas computacionales, con lo que la utilización de las computadoras se extendió a la generalidad de las ciencias, la lingüística entre ellas. El Apple II a partir de 1977 y la computadora personal IBM a partir de 1981 fueron las primeras máquinas producidas a gran escala y constituyeron éxitos de ventas a pesar de los elevadísimos precios que tenían en aquel momento[2]. En segundo lugar, la aparición de internet y, como una de las consecuencias de su existencia, el arranque de la red mundial (la web). Gra-

2 Según la *Wikipedia* (s. v. *Apple II*), una buena parte del éxito del Apple II se debe a la inclusión de la *VisiCalc*, una hoja de cálculo de propósito general que refleja lo que acabo de indicar acerca de la generalización del uso y su origen en la falta de necesidad de conocimientos informáticos especiales para el manejo de estas máquinas.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

cias a internet, el manejo y explotación de corpus se puede hacer sin necesidad de desplazamiento físico hasta el lugar en que se encuentra la máquina que contiene el corpus o la duplicación del recurso en otra computadora. La historia de la explotación del *BNC* y los sistemas utilizados para su difusión y consulta pueden dar una buena idea de la evolución experimentada en este punto. A mis estudiantes les resultaba casi imposible de entender que unos índices léxicos de la obra poética de Quevedo, que se produjeron mediante tratamiento informático en nuestro departamento de la Universidade de Santiago de Compostela a comienzos de los años 90, se difundieran exclusivamente en forma de libro. Sin embargo, muy poco tiempo después, la primera publicación del *CREA* se hizo ya con el modelo cliente-servidor y sin más requisitos para el cliente que el uso de un navegador actualizado. Son las ventajas que en algunos casos tiene llegar con cierto retraso a algunos desarrollos.

[4] Por otro lado, la lingüística ha cambiado considerablemente desde mediados de los años 60 y ha desaparecido aquella especie de prohibición que, por influencia de los postulados del primer generativismo, pesaba sobre los corpus, los datos de carácter estadístico, etc. Aunque no toda la historia de la LC es explicable con ese factor, sí es cierto que en los Estados Unidos esta actitud y su repercusión sobre la financiación de proyectos y avances en las carreras personales tuvieron una gran influencia. Desde hace ya bastantes años, la situación es claramente distinta y el enfoque empírico, basado en el análisis de datos externos, sobre los que se construyen las hipótesis acerca de cómo funciona el sistema lingüístico es, creo yo, el más utilizado. Me parecen muy ilustrativas las palabras de Sampson (2011: 197), para quien

> corpus linguists are just people who study language and languages in an empirical, scientific manner, using whatever sources of empirical data are available; at the present time it happens that, for many aspects of language, the most useful data sources are often electronic corpora. I work a lot with corpora, but I think of myself as a linguist, not a 'corpus linguist'. If some aspect of language is better studied using other tools, I will use those.

A consecuencia de todo este conjunto de factores, los corpus han experimentado una asombrosa evolución en aspectos relacionados con el tamaño, el grado de codificación, la anotación añadida o las posibilidades de explotación, como trataré de mostrar en los apartados siguientes.

2 Corpus masivos (oportunistas) y corpus especializados

[5] En la faceta que nos interesa aquí debemos diferenciar varios tipos de corpus. En una zona exterior (quizá incluso marginal) se encuentra la aproximación conocida como *web as corpus*. Ya Sinclair (2005: 15) rechazaba la consideración de la web como un corpus ingente basándose en su falta de diseño, su carácter continuamente cambiante y la carencia de codificación de los textos. Sin em-

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

bargo, debe decirse que también ha tenido defensores, como por ejemplo Kilgarriff (2013).

[6] Dejando a un lado esta opción, los corpus de referencia, con el *CREA* y el *CORPES XXI* como ejemplos prototípicos para el español, están situados en una zona intermedia entre los conjuntos textuales que se encuentran a ambos lados. En el primero de ellos debemos situar los corpus masivos, formados por miles o incluso cientos de miles de millones de palabras. En la zona más externa se encuentran los conjuntos inmensos que se están utilizando hoy en día para aplicaciones vinculadas a la inteligencia artificial, la traducción automática, etc. para la identificación de modelos y patrones lingüísticos que se puedan emplear, por ejemplo, en la generación de textos. En el ámbito del español, el corpus vinculado al proyecto *MarIA* es el más importante en la actualidad[3]. Se ha construido sobre las enormes masas de datos que se pueden descargar de la red (con los filtros necesarios) utilizando los recursos proporcionados por la Biblioteca Nacional de España. Son corpus que tienen, sin duda, una gran utilidad, pero no para la investigación lingüística en su sentido habitual, sino en aplicaciones del estilo de las mencionadas. No tienen diseño, no hay selección de los materiales más allá de la que pueden proporcionar identificadores de lenguas y algunos filtros de carácter formal, con lo que los datos resultantes no pueden alcanzar el grado de refinamiento, la granularidad que se necesita en la investigación lingüística.

[7] Mucho más cerca de nuestro ámbito de trabajo están los corpus, también masivos, del estilo del *CdE Web/Dialects*, *CdE NOW*, *esTenTen* o el *CEA*[4]. Con independencia de su tamaño, la característica común a todos radica en que están formados por textos descargados de la red (de nuevo, con la posible aplicación de reconocedores de lengua y filtros de eliminación de materiales inservibles) o bien de grandes recursos construidos previamente, como la *Wikipedia*, materiales procedentes de instituciones como parlamentos o asambleas, mensajes de *Twitter*, etc. A estos corpus se les pueden aplicar las herramientas de anotación disponibles (para análisis morfosintáctico, por ejemplo), de modo que, como ocurre con *esTenTen*, admiten la incorporación de capas adicionales del mayor interés para la investigación, al estilo del *Sketch engine*.

[8] El punto flaco de estos corpus está, sin duda, en la codificación. Su enorme tamaño y la procedencia de los materiales hacen imposible alcanzar un nivel

3 «Para ser exactos, MarIA se ha entrenado con 135.733.450.668 de palabras procedentes de millones de páginas web que recolecta la Biblioteca Nacional y que ocupan un total de 570 gigabytes de información. Para estos mismos entrenamientos, se ha utilizado el superordenador MareNostrum del Centro Nacional de Supercomputación de Barcelona y ha sido necesaria una potencia de cálculo de 9,7 trillones de operaciones (969 exaflops)» (Ministerio para la transformación digital y de la función pública 2022). Para su descripción técnica, ver Gutiérrez-Fandiño et al. (2022).

4 Con datos actualizados en septiembre de 2023, el *CdE Web/Dialects* tiene 2000 millones de formas, el *CdE NOW* 7300 millones, el *esTenTen* del 2018 unos 17000 millones y el *CEA* 540 millones.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

elevado en la consideración de, por ejemplo, tipos de texto, tan fundamental en la investigación de los fenómenos lingüísticos. El *esTenTen* utiliza el dominio de alto nivel (.es, .com, etc.) para la clasificación de los resultados, con lo que la tipología se convierte en una extraña mezcla de documentos procedentes de servidores situados en un determinado país (.es o .uy, por ejemplo) y documentos de páginas dedicadas a negocios, salud, ciencia, deportes, etc. En *CdE Web/Dialects* y *CdE NOW* se utiliza el dominio del país en el que reside físicamente el servidor y se supone que ese rasgo proporciona también el país al que debe ser asignado el texto. En algunos casos, sería posible incluir, de forma automática, indicaciones acerca del área temática. Por ejemplo, usando los datos incluidos en la cabecera de las noticias periodísticas. Esto, sin embargo, resulta complicado por la falta de uniformidad, de modo que no se hace porque obligaría a entrar en el sistema de codificación concreta utilizado por cada publicación.

[9] Corpus de este tipo son los que Mair (2006: 355) ha calificado como «big and messy». No hay en ellos diseño, sino que se trata de conseguir todo lo posible entre los materiales a los que hay acceso en la red. Eso supone un insalvable problema de falta de variedad en los textos que entran, agravada por las dificultades para caracterizarlos tipológicamente. De ahí que hayan recibido la denominación de «corpus oportunistas» (cf. Llisterri Boix & Torruella Casañas 1999: 56).

[10] Al otro lado del espectro tenemos corpus pequeños y muy especializados. Se trata de corpus constituidos por textos que tienen unas características muy marcadas: textos orales, producciones de aprendices de una segunda lengua, de personas que viven en zonas rurales, de intercambios que tienen lugar en ciertas situaciones (entre pacientes y personal sanitario, por ejemplo), cartas particulares, etc. En este grupo entran también corpus consistentes en textos de cierto carácter, época o territorio, como el *CORDIAM*, el *CHARTA*, el *CODEA+ 2022*, el *Post scriptum*, el *ODE* y algunos otros de características generales similares.

[11] Corpus de este tipo son los que Mair (2006: 356) ha caracterizado como «small and tidy». Resultan muy útiles para ciertos propósitos, con una codificación muy detallada y cuidada, con materiales tratados de modo uniforme y, en muchos casos, inéditos hasta su inclusión en el recurso. Además, suelen presentar la posibilidad de añadir capas con utilidades específicas, como, entre otras, diferentes transcripciones (ediciones paleográficas y críticas, por ejemplo), o de incluir una imagen del manuscrito o alineación de textos en varias lenguas, como en el corpus *Biblia medieval*. El problema está en su tamaño reducido y, con mucha frecuencia, el carácter uniforme de los textos recogidos, lo cual hace que la posibilidad de proyección de los resultados obtenidos sea limitada.

[12] En este grupo hay que tener muy en cuenta los factores que determinan su diseño y, por tanto, su construcción. Por ejemplo, hay corpus orales, como el *PRESEEA*, el *COSER* o el *ESLORA*. Encontramos también corpus orientados hacia estudios diacrónicos (como *CORDIAM*, *CHARTA*, *Biblia medieval*, *CODEA+*

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

2022, *Post scriptum*, *ODE*, etc.) y corpus formados por producciones, orales o escritas, de estudiantes de una segunda lengua, como el *CAES* o el *CEDEL2*, o centrados en la lengua de cierto segmento de la población, como el *COLA*. Como se ve, el carácter muy especializado de estos corpus hace que sea necesario tomar en cuenta su especificidad para hacer las precisiones adicionales necesarias.

3 Los corpus de referencia

[13] Entre ambos extremos se encuentran los que llamamos corpus de referencia (CR). Son corpus de propósito general y, por tanto, necesitan un diseño que asegure la representatividad y el equilibrio necesarios para garantizar la fiabilidad de los resultados obtenidos de los múltiples y variados subcorpus que pueden contener. El modelo clásico es, sin duda, el *BNC*, finalizado en la primera mitad de los años 90 y constituido por cien millones de formas del inglés británico, con un 10 % de textos de carácter oral. Su volumen se queda corto hoy, pero el modelo sigue vigente en todos los demás aspectos.

[14] Desde una perspectiva general, la utilidad básica de los CR reside en que tienen el tamaño y el grado de codificación necesarios para proporcionar resultados que muestren no ya los valores generales de los fenómenos lingüísticos, sino su distribución según los parámetros utilizados en el diseño del corpus. Es decir, país, época, medio, tipo de texto, área temática, etc. Gracias a ello, pueden ser utilizados para comprobar el grado en que la frecuencia de un elemento o fenómeno varía en diferentes tipos de texto. El problema está en que añadir esa codificación, semejante a la que se utiliza en los corpus pequeños, exige la intervención de seres humanos. Con una visión más detallada:

1. Los CR se caracterizan por partir de un diseño concreto, no sometido al peso de los materiales con presencia mayoritaria en la red o sean de algún tipo especial (discursos parlamentarios, mensajes de *Twitter*, etc.).
2. Dado lo anterior, en un corpus de referencia es inevitable la actuación de profesionales (no solo lingüistas), primero en la selección de los textos y luego en su codificación. Por citar algún aspecto llamativo, este rasgo puede llevar a investigaciones acerca del origen del autor de una noticia periodística o de una persona que ha intervenido en una tertulia radiofónica.
3. Los dos puntos anteriores producen la posibilidad de trabajar con subcorpus que pueden resultar altamente especificados (por ejemplo, noticias de prensa sobre economía publicadas en periódicos colombianos entre 2010 y 2013) y comparar los resultados con los de otros subcorpus para tratar de establecer las condiciones en las que un determinado fenómeno varía o cambia.
4. Si, como es habitual, los CR no admiten la descarga de los textos completos para proteger los derechos de autores, editores literarios y editores comerciales, es posible incluir una novela recién publicada o el guion de una película que se acaba de estrenar. Con ello, la variedad y actualidad de los textos de los que se pueden obtener datos se amplía extraordinariamente.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

Para decirlo de forma rápida, los CR intentan conjugar lo más útil de los otros tipos de corpus que hemos analizado. Por una parte, pretenden lograr un tamaño de cientos, quizá incluso miles de millones de formas. Por otra, añaden a los textos una codificación cuidadosa y detallada, acorde con el diseño establecido y los objetivos perseguidos. Adaptando la expresión de Mair (2006), aspiran a ser corpus *big and tidy*. Es un objetivo perfectamente alcanzable, pero, por supuesto, implica costes importantes y, además, una financiación continuada si se pretende que estén actualizados, como es el caso del *CORPES XXI*.

4 El efecto de los corpus de referencia

[15] Ya he aludido al cambio en el panorama general de los estudios lingüísticos y la consiguiente expansión del manejo de datos externos como fase inicial de cualquier trabajo de investigación. Incluso en terrenos en los que la introspección no resultaba una vía practicable, como, por ejemplo, en los estudios históricos, la difusión del uso de corpus textuales ha tenido el efecto beneficioso de poner al alcance de cualquier persona interesada un amplio conjunto de datos que no le han exigido una enorme inversión previa de tiempo para su recolección. Hilpert & Mair (2015: 199) destacan el papel de los corpus en las investigaciones diacrónicas indicando que «[they] illustrate how the use of corpus data allows researchers to go beyond the mere statement that a grammatical change happened, and to address the questions of when and how something happened». En el caso concreto de la lingüística hispánica, creo que también es importante señalar que los corpus han roto la limitación de trabajar con un conjunto reducido de textos, casi siempre pertenecientes al canon, sin tener en cuenta que la historia de la lengua se manifiesta también (y a veces lo hace de forma preferente) en textos que no son fundamentales para la comprensión de la historia literaria.

[16] *Mutatis mutandis*, algo muy parecido se podría decir de los corpus no diacrónicos con respecto a la variación. La existencia de corpus de referencia correspondientes a todo el español contemporáneo ha hecho posible trabajar con comodidad en la comprensión de los fenómenos lingüísticos y la forma en que se manifiestan en las diferentes variedades. Se trata, por tanto, de enfocar adecuadamente la variación, que ha dejado de ser considerada como un problema con el que había que convivir para convertirse en un aspecto básico en la propia concepción del sistema lingüístico.

[17] Creo que, en términos generales, el impacto de la LC (y, en particular, de los CR) sobre la investigación lingüística ha sido amplio y muy importante. Es posible que continúen todavía las discusiones acerca de si la LC es una teoría, una (sub)disciplina, una metodología, las tres cosas o ninguna de ellas. No tiene mayor importancia porque la diferencia no produce ningún efecto en la práctica. La LC constituye

> una aproximación al estudio de los hechos lingüísticos de orientación empírica y basada en el análisis detallado de gran cantidad de datos (los corpus), con lo que queda

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

> patente su oposición tanto a la lingüística racionalista como a la descriptiva tradicional. (Rojo 2021: 48)

Hay en esta caracterización dos aspectos generales que conviene destacar. El primero de ellos consiste en la facilidad con que cualquier persona interesada puede acceder a grandes masas de datos, habitualmente bien organizados y codificados, sin necesidad de tener que invertir enormes cantidades de tiempo en la recogida de un conjunto de datos infinitamente más reducido y menos variado. No se trata de una cuestión puramente cuantitativa, sino que produce un salto cualitativo, como señalaré de nuevo posteriormente: la consideración de cualquier fenómeno lingüístico proporcionada por la posibilidad de estudiarlo en un corpus de 400 o 500 millones de formas es muy diferente de la que nos ofrecían las técnicas tradicionales.

[18] En segundo lugar, el trabajo en LC supone (o debería suponer) moverse en la línea de lograr la explicabilidad total, es decir, la «total accountability» a la que se refería Labov (1972: 108) para la sociolingüística, introducida por Quirk (1992) para la LC y repetidamente destacada como principio fundamental por autores como Leech (1992, 2015). Se trata de analizar todos los casos de un determinado fenómeno o elemento, incluidos los contextos en los que podría darse, pero no se da. Es evidente que este precepto no puede tener una interpretación literal cuando la consulta nos devuelve, por ejemplo, 20000 casos de un cierto fenómeno. Es relativamente sencillo descargar una buena parte de la revisión en herramientas computacionales que clasifiquen, agrupen, extraigan rasgos comunes, etc., al menos de algunos de los aspectos que hay que estudiar (cf. Rojo 2023a). La forma de trabajar en lexicografía electrónica es un ejemplo claro de cómo desenvolverse con cifras de este tamaño y lograr resultados válidos en un plazo razonable.

[19] Con la intención de estructurar la influencia de la LC en la lingüística contemporánea, Leech (2015) se refiere a hallazgos inesperados, a la inclusión de áreas desatendidas por la lingüística anterior y al estudio de la lengua oral. Aunque su perspectiva se circunscribe voluntariamente a la lingüística inglesa, constituye una buena guía para intentar algo semejante en la lingüística española.

[20] Entre los descubrimientos inesperados, destaca Leech (2015) el realizado por Hudson (1994) de que el 37 % de las formas de un corpus son sustantivos (en una acepción muy amplia, que incluye nombres propios y comunes y también pronombres personales). Leech (2015) destaca la importancia del hecho de que este porcentaje conjunto es estable y se obtiene tanto en la lengua oral como en la escrita. Con independencia de la valoración que otorguemos a esta cifra y su reorganización según diferentes tipos de texto, lo realmente importante aquí es que la posibilidad de extraer los datos de conjuntos formados por cientos de millones de formas nos proporciona una visión muy distinta de la forma en que está organizada la estructura de las lenguas, tanto en el léxico como en la gramática. Por citar

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

un caso claro, la limitación del volumen en los corpus construidos y la imposibilidad en la práctica de publicar (en forma impresa) los datos de todos los elementos identificados producía en los diccionarios de frecuencias una visión basada fundamentalmente en el comportamiento y las características de los elementos con frecuencias relativamente altas o muy altas. La ampliación de los corpus produce cambios relevantes incluso en aspectos tan aparentemente neutros como la distribución de las conjugaciones verbales en español (Rojo 2006).

[21] En la segunda de las líneas enumeradas en [19] hace referencia Leech (2015) al trabajo en áreas que no han despertado interés en los estudios más orientados hacia cuestiones teóricas por su supuesta falta de relevancia en ese aspecto. Menciona concretamente el estudio de los elementos adverbiales (en general) y los marcadores del discurso. No estoy seguro de en qué medida esto se puede aplicar a la lingüística española. Lo que sí veo claro es que durante muchos años una buena parte de los trabajos se refería a fenómenos cuyo carácter daba mayor importancia a las cuestiones teóricas y metodológicas, con un cierto descuido hacia la descripción completa de los fenómenos en sí mismos (favorecida, probablemente, por el uso de la introspección como procedimiento fundamental o incluso único). Por los factores ya apuntados, la LC procede de modo distinto y pretende trabajar con todos los casos del fenómeno que se está estudiando.

[22] Es evidente también la influencia de la LC en los estudios sobre la lengua oral. En lo que se refiere al español, hemos pasado en poco tiempo de investigaciones basadas en observaciones personales aisladas o textos escritos en los que supuestamente se reproduce la lengua oral (*El Jarama*, *Historias del Kronen* y textos similares, obras de teatro, etc.) a disponer del conjunto de corpus amplios, diversificados en cuanto a orientaciones y con un grado bastante alto de anotación y codificación. *PRESEEA*, *Val.Es.Co.*, *AMERESCO*, *COSER* o *ESLORA* son algunos de los corpus que tienen a su disposición quienes deseen trabajar sobre la lengua oral.

[23] Hay todavía otro aspecto que me gustaría resaltar. La posibilidad de trabajar con corpus de gran tamaño y, al tiempo, bien codificados permite renovar algunos recursos de utilidad bastante limitada cuando se basan en los recursos tradicionales. Un caso claro, en mi opinión, es el constituido por los diccionarios de frecuencias léxicas. No se puede negar su utilidad, pero también es cierto que se limitan a elementos léxicos, las divisiones son muy generales y no se pueden entrecruzar. La posibilidad de utilizar un corpus de buen tamaño, debidamente anotado y codificado permite llegar al concepto de diccionario de frecuencias dinámico que permita recuperar, por ejemplo, el inventario de cierto tipo de formas verbales y su distribución por países, áreas temáticas, tipo de texto, etc. (cf. Rojo 2023b). No es algo diferente de lo que se hace habitualmente en la consulta de un corpus, pero la novedad radica en la posibilidad de convertirlo en un recurso general, manejable con carácter independiente, y, sobre todo, en su carácter dinámico,

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

lo cual le proporciona la flexibilidad derivada del hecho de que no está precalculado y, por tanto, se ajusta a lo que se necesita en cada caso.

[24] No se puede ocultar que la utilización de los materiales incluidos en un corpus resulta más o menos sencilla según el tipo de fenómeno de que se trate en cada caso. Incluso se ha aludido en bastantes ocasiones a que los corpus permiten la recopilación de lo que está en los textos, pero no, en cambio, de aquello que no se da, como podría ser, por ejemplo, ausencia de sujeto en una cláusula. Es evidente que las recuperaciones son más sencillas y rápidas cuando lo que hay que buscar son cadenas de caracteres gráficos, al estilo de los buscadores comerciales. A partir de ahí, las aplicaciones de consulta tienen que utilizar la vía de los metacaracteres ('comodines') y las expresiones regulares para dar potencia y carácter general a las búsquedas. Pero, sobre todo, la cuestión está en el tipo de anotación que se añade a los textos. Recuperar los ejemplos que contienen todas las formas pertenecientes al paradigma de un verbo se puede hacer con metacaracteres (sobre todo si se trata de un verbo regular), pero devolverá también formas que responden al patrón utilizado y no pertenecen al verbo. Para poder utilizar elementos abstractos en las búsquedas (del tipo *formas del futuro simple de indicativo de un verbo concreto o de todos los verbos*) es preciso que esa información haya sido añadida mediante un proceso de anotación morfosintáctica y lematización. Lo mismo se aplica a otros rasgos de nivel más elevado. No es posible recuperar información sobre aspectos sintácticos si previamente no se ha aplicado un análisis sintáctico automático o bien se ha almacenado en una base de datos el resultado del análisis manual, al estilo de lo que se puede conseguir en la *BDS* o en la base *ADESSE*[5]. La cuestión no es, por tanto, cuáles son las capacidades de los corpus textuales, sino qué tipo de información debe figurar en un corpus para que se pueda obtener un cierto tipo de resultados.

[25] Me parece evidente, por tanto, que los efectos de la LC sobre los logros de la investigación acerca de la lengua española son de importancia innegable y han cambiado radicalmente el panorama en este campo. Hay en este punto, sin embargo, un aspecto menos positivo sobre el que debo decir algo. En la etapa anterior a la LC, lo habitual era que en cada investigación fuera necesario dedicar tiempo, mucho tiempo, a la recogida de datos, realizada de forma manual. Esto exigía la lectura de los textos utilizados, al menos en el grado necesario para localizar y entender el fragmento seleccionado tomando en cuenta su contexto. La facilidad en el acceso a grandes masas de datos permite trabajarlos de forma conjunta, quizá solo en el aspecto puramente cuantitativo, sin tener en cuenta su valoración individual. Sé, por supuesto, que esto es imposible cuando la consulta nos devuelve 10000 casos de un fenómeno, pero es necesario, al menos, revisar los criterios utilizados en la codificación de los corpus con los que se trabaje y examinar

5 Por ejemplo, cláusulas que respondan al esquema de sujeto-predicado-complemento directo en las que el sujeto sea una frase nominal de carácter no animado y el complemento directo una cláusula completiva con el verbo en subjuntivo.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

con cuidado aquellos ejemplos que choquen con lo esperable o presenten características muy especiales. Se trata, en definitiva, de recuperar la parte más positiva e interesante del enfoque filológico que era preciso para enfrentarse con cada texto antes de la existencia de los corpus.

[26] En un aspecto conectado, se ha resaltado con mucha frecuencia la importancia del enfoque cuantitativo en la LC. En efecto, la frecuencia con que un elemento o un fenómeno se presenta en general y en los diferentes subcorpus que pueden ser establecidos en su interior es de la mayor relevancia. Y, como es natural, manejar esa faceta requiere el uso de técnicas y herramientas especializadas. Algunos autores han insistido en que la formación de quienes pretendan dedicarse profesionalmente a la lingüística debería incluir algún curso de estadística. En paralelo, se han multiplicado los libros y los cursos sobre estadística para lingüistas, muchos de ellos vinculados al uso del lenguaje de programación R. Esa necesidad es evidente: además de hacer recuentos, necesitamos poder valorar e interpretar los resultados obtenidos.

[27] En algunos casos, sin embargo, se está produciendo una distorsión en los objetivos. En un estudio reciente de Larsson, Egbert & Biber (2022) se compara la atención (en realidad, el espacio) dedicada a los aspectos netamente lingüísticos y a las discusiones de carácter estadístico en trabajos publicados en diferentes revistas en los años 2009 y 2019. El resultado muestra claramente que en los artículos más recientes la atención dedicada a las discusiones acerca de las pruebas estadísticas aplicadas se ha incrementado considerablemente y supera en muchos casos la referida a las cuestiones netamente lingüísticas que se están investigando. Con sus propias palabras, «our results showed that the greater the focus on statistical reporting, the more likely it is for language and linguistic analysis to get backgrounded» (Larsson, Egbert & Biber 2022: 153). Por supuesto, el trabajo en LC requiere dedicar cierta atención a los aspectos estadísticos, entender las pruebas aplicadas y poder valorar sus resultados desde el punto de vista lingüístico. En algunos casos, las discusiones exclusivamente estadísticas son inevitables. Por ejemplo, en los últimos tiempos se ha discutido bastante acerca de la fiabilidad del conocido índice de dispersión de Juilland (Juilland & Chang 1964). Eso es necesario y nos aclara a quienes no tenemos formación estadística especializada cuál es el índice que debemos aplicar preferentemente si trabajamos en este terreno, pero no significa que esas discusiones tengan que aparecer en todo trabajo dedicado a las frecuencias léxicas. Para decirlo todo, debo indicar que esta atención excesiva a lo puramente cuantitativo se produjo también en épocas anteriores, probablemente como consecuencia de la difusión de las computadoras y las facilidades consiguientes para hacer cálculos sin el esfuerzo que suponían anteriormente.

5 El futuro de los corpus de referencia

[28] Las líneas previsibles para los CR en los próximos años se sitúan, a mi modo de ver, en los aspectos siguientes. En la línea principal, se producirá el in-

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

cremento del tamaño (quizá hasta llegar a tamaños de mil millones de formas) y también una codificación más detallada en, por ejemplo, la tipología de los textos, para indicar, supongamos, no solo que se trata de un texto procedente de la prensa, sino si es una noticia, un reportaje, un editorial, la carta de un lector, etc. También será preciso refinar la codificación interna de los textos para, entre otros aspectos, marcar las citas y establecer normas específicas para su procesamiento. Además, es importante mejorar los procesos de anotación morfosintáctica y lematización.

[29] Con carácter especial, es necesario aumentar considerablemente el peso de los textos orales. El modelo del *BNC* establece un 10 % del total y ese es también el fijado para el *CORPES XXI*, pero en este último se está todavía muy lejos de alcanzarlo. Es comprensible, porque son textos más difíciles de conseguir y, sobre todo, su transcripción es muy trabajosa a pesar de las mejoras que se pueden derivar de sistemas de transcripción automática. Un punto importante en esta línea es la complementar el texto transcrito con la posibilidad de acceder al audio. El *CORPES XXI* incorpora esta posibilidad desde hace ya algún tiempo. No se trata de poder oír la transcripción de toda la entrevista, sino de recuperar únicamente la zona correspondiente al texto que aparece en la concordancia, como se hace con todos los textos de, por ejemplo, *ESLORA*.

[30] La inclusión del audio como un elemento constitutivo del conjunto de posibilidades de recuperación se conecta con un factor más general, que es la posibilidad de dar a los corpus un carácter multimodal. En efecto, para muchos textos orales la posibilidad de incorporar el audio supone un avance importante, pero todavía nos deja sin información crucial acerca del modo en que se producen los intercambios, que solo puede ser recuperado mediante la inclusión del vídeo. Se trata, pues, de construir corpus en los que estén alineados texto transcrito, audio y vídeo, de modo que la recuperación se haga a través del texto, pero exista la posibilidad de examinar también el sonido y la imagen. En otra dirección, es importante pensar en corpus en los que, en niveles superpuestos y conectados, podamos disponer de información fónica (si es un texto oral), ortográfica, anotación morfosintáctica, sintáctica, semántica y pragmática. Es evidente que todo esto requiere el desarrollo de herramientas potentes, pero precisa, sobre todo, el conocimiento técnico que permita producir las aplicaciones de análisis necesarias para ello.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

Abreviaturas y referencias bibliográficas

ADESSE = José M. García-Miguel (ed.) 2002-2023. *ADESSE. Base de datos de verbos, alternancias de diátesis y esquemas sintáctico-semánticos del español*. http://adesse.uvigo.es.

AMERESCO = Marta Albelda Marco, María Estellés Arguedas (eds.) 2010-. *AMERESCO. América y España. Español coloquial*. https://esvaratenuacion.es/ameresco.

BDS = Guillermo Rojo (ed.) 2021. *Base de datos sintácticos del español actual*. http://www.bds.usc.es.

Biblia medieval = Andrés Enrique-Arias (ed.) 2008-. *Biblia medieval*. http://bibliamedieval.es.

BNC = BNC Consortium (ed.) 2007. *British national corpus*. www.natcorp.ox.ac.uk.

Brown corpus = W. Nelson Francis, Henry Kučera (eds.). *Brown corpus. The standard corpus of present-day edited American English*. https://varieng.helsinki.fi/CoRD/corpora/BROWN.

CAES = Guillermo Rojo, Ignacio Palacios (eds.) 2022. *Corpus de aprendices de español L2*. http://galvan.usc.es/caes.

CdE NOW = Mark Davies (ed.) 2019. *Corpus del español NOW (News on the web)*. https://www.corpusdelespanol.org/now.

CdE Web/Dialects = Mark Davies (ed.) 2016. *Corpus del español (Web/Dialects)*. https://www.corpusdelespanol.org/web-dial.

CEA = Carlos Subirats, Marc Ortega (eds.) 2012. *Corpus del español actual*. http://spanishfn.org/tools/cea/spanish.

CEDEL2 = Cristóbal Lozano (ed.) 2020-. *Corpus escrito de español L2*. http://cedel2.learnercorpora.com.

CHARTA = Pedro Sánchez-Prieto Borja (ed.) 2011-. *Corpus hispánico y americano en la red*. https://www.redcharta.es.

CODEA+ 2022 = Pedro Sánchez-Prieto Borja (ed.) 2022. *CODEA+ 2022. Corpus de documentos españoles anteriores a 1900*. https://www.corpuscodea.es.

COLA = Annette Myre Jørgensen (ed.) 2023. *Corpus oral de lenguaje adolescente*. https://blogg.hiof.no/colam-esp.

CORDIAM = Concepción Company Company, Virginia Bertolotti (eds.) 2016-. *Corpus diacrónico y diatópico del español de América*. https://www.cordiam.org/.

CORPES XXI = Real Academia Española (ed.) 2023-. *Corpus del español del siglo XXI*. https://www.rae.es/corpes/.

COSER = Inés Fernández-Ordóñez (ed.) 2005-. *Corpus oral y sonoro del español rural*. http://www.corpusrural.es.

CR = Corpus de referencia.

CREA = Real Academia Española (ed.) 2008. *Corpus de referencia del español actual*. https://corpus.rae.es/creanet.html.

ESLORA = Victoria Vázquez Rozas (ed.) 2007-. *Corpus para el estudio del español oral*. http://eslora.usc.es.

esTenTen = Lexical computing (ed.) 2011-2021. *Spanish web corpus*. https://www.sketchengine.eu/estenten-spanish-corpus.

Gutiérrez Fandiño et al. 2022 = Asier Gutiérrez Fandiño et al. 2022. MarIA: Spanish language models. *Procesamiento del lenguaje natural* 68, 39-60. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405/3820.

Hilpert & Mair 2015 = Martin Hilpert, Christian Mair 2015. Grammatical change. Douglas Biber, Randi Reppen (eds.). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 180-200.

Hudson 1994 = Richard Hudson 1994. About 37 % of word-tokens are nouns. *Language* 70, 331-339.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

Juilland & Chang 1964 = Alphonse G. Juilland, Eugenio Chang-Rodríguez 1964. *Frequency dictionary of Spanish words*. Mouton.

Kilgarriff 2013 = Adam Kilgarriff 2013. Using corpora and the web as data sources for dictionaries. Howard Jackson (ed.). *The Bloomsbury companion to lexicography*. Bloomsbury, 77-96.

Labov 1972 = William Labov 1972. *Sociolinguistic patterns*. University of Philadelphia Press [hay traducción española de José Miguel Marinas Herreras: *Modelos sociolingüísticos*. Cátedra. 1983].

Larsson, Egbert & Biber 2022 = Tove Larsson, Jesse Egbert, Douglas Biber 2022. On the status of statistical reporting *versus* linguistic description in corpus linguistics: a ten-year perspective. *Corpora* 17.1, 137-157.

LC = Lingüística de corpus.

Leech 1992 = Geoffrey Leech 1992. Corpora and theories of linguistic performance. Jan Svartvik (ed.). *Directions in corpus linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. De Gruyter Mouton, 105-122.

Leech 2015 = Geoffrey Leech 2015. Descriptive grammar. Douglas Biber, Randi Reppen (eds.). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press, 146-160.

Llisterri Boix & Torruella Casañas 1999 = Joaquim Llisterri Boix, Joan Torruella Casañas 1999. Diseño de corpus textuales y orales. José M. Blecua, Gloria Clavería, Carlos Sánchez, Joan Torruella (eds.). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Universidad Autónoma de Barcelona, 45-77.

Mair 2006 = Christian Mair 2006. Tracking ongoing grammatical change and recent diversification in present-day standard English: the complementary role of small and large corpora. Antoinette Renouf, Andrew Kehoe (eds.). *The changing face of corpus linguistics*. Rodopi, 355-376.

MarIA = Asier Gutiérrez Fandiño et al. 2021-. *MarIA. Spanish language models*. https://github.com/PlanTL-GOB-ES/lm-spanish.

Ministerio para la transformación digital y de la función pública 2022 = Ministerio para la transformación digital y de la función pública 2022. *Así es MarIA, la primera inteligencia artificial de la lengua española*. https://datos.gob.es/es/blog/asi-es-maria-la-primera-inteligencia-artificial-de-la-lengua-espanola.

ODE = Miguel Calderón Campos, María T. García Godoy 2019-. *Oralia diacrónica del español*. http://corpora.ugr.es/ode.

Post Scriptum = Centro de lingüística da Universidade de Lisboa (ed.) 2014. *Post Scriptum. Archivo digital de escritura cotidiana en Portugal y España en la Edad moderna*. http://teitok.clul.ul.pt/postscriptum/index.php.

PRESEEA = Francisco Moreno Fernández, Ana Cestero Mancera (eds.). *Proyecto para el estudio sociolingüístico del español de España y de América*. https://preseea.uah.es.

Quirk 1992 = Randolph Quirk 1992. On corpus principles and design. Jan Svartvik (ed.). *Directions in corpus linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. De Gruyter Mouton, 457-469.

Rojo 2006 = Guillermo Rojo 2006. Sobre las frecuencias verbales en español. Mercedes Sedano, Adriana Bolívar, Martha Shiro (eds.). *Haciendo lingüística. Homenaje a Paola Bentivoglio*. Universidad Central de Venezuela, 309-324.

Rojo 2015 = Guillermo Rojo 2015. Sobre los antecedentes de la lingüística de corpus. *Studium grammaticae. Homenaje al Profesor José Antonio Martínez*. Universidad de Oviedo, 675-689.

Rojo 2021 = Guillermo Rojo 2021. *Introducción a la lingüística de corpus en español*. Routledge.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815

Rojo 2023a = Guillermo Rojo 2023a. *Análisis informatizado de textos*. Universidade de Santiago de Compostela.

Rojo 2023b = Guillermo Rojo 2023b. Hacia un nuevo concepto de diccionario de frecuencias. Dolores Corbella, Josefa Dorta, Rafael Padrón (eds.). *Perspectives en linguistique et philologie romanes (I et II)*. ÉLiPhi, 45-63.

Sampson 2011 = Geoffrey Sampson 2011. A two-way exchange between syntax and corpora. Vander Viana, Sonia Zyngier, Geoff Barnbrook (eds.). *Perspectives on corpus linguistics*. Benjamins, 197-211.

Sinclair 2005 = John Sinclair 2005. Corpus and text. Basic principles. Martin Wynne (ed.). *Developing linguistic corpora. A guide to good practice*. Oxbow Books, 1-16.

Sketch engine = Lexical computing 2003-. *Sketch engine. Corpus query and management system*. https://www.sketchengine.eu.

Val.Es.Co. = Salvador Pons Bordería (ed.) 2024. *Corpus Val.Es.Co. 3.0.* http://www.valesco.es.

Wikipedia = Wikimedia Foundation (ed.) 2024. *Wikipedia*. https://en.wikipedia.org/wiki/Apple_II.

DOI: https://doi.org/10.25364/19.2024.12.2 ISSN: 2663-9815