Lingüística de corpus y lingüística del español

Guillermo Rojo

Universidad de Santiago de Compostela

1. Como muchos de ustedes saben, mi primera propuesta a la honrosa invitación de los organizadores de este congreso fue la de desarrollar un tema estrictamente gramatical, un tema en el que se pudieran combinar las cuestiones teóricas con la descripción de lo que se puede observar en español actual y también las líneas generales de la evolución experimentada por ese fenómeno a lo largo de los siglos. Sin embargo, la sugerencia de una buena amiga y colega de dedicar esta sesión a hacer una especie de panorama del estado actual de la lingüística de corpus y de lo que ha supuesto en la lingüística hispánica fue pareciéndome cada vez más interesante y seductora, de modo que me decidí a hacer la propuesta de cambio a los organizadores, que, en una nueva prueba de generosidad, no pusieron el menor inconveniente para aceptarla. Vaya pues, en primer lugar, mi agradecimiento por la invitación y también por la flexibilidad que han mostrado.

Creo que, en efecto, esta reunión internacional en la que está presente una muy nutrida representación de quienes nos dedicamos al estudio de las lenguas y muy especialmente del portugués y del español es una magnífica ocasión para que alguien que lleva ya unos cuantos años dedicado al trabajo en esta línea y tiene sobre sus espaldas la experiencia de haber coordinado el diseño, construcción y explotación de varios corpus desgrane ante ustedes algunas ideas acerca de un modo de entender y estudiar los fenómenos lingüísticos que, sin duda, ha supuesto un enorme cambio en nuestro modo de hacer y de pensar. Me propongo, pues, trazar las líneas básicas de la evolución experimentada por la lingüística de corpus, señalar sus logros fundamentales, esbozar la situación en que se encuentra actualmente y tratar de dibujar las grandes líneas de lo que espero que sea su futuro en los años inmediatos.

2. La historia de la lingüística de corpus es muy corta. El punto de arranque obligado es la aparición en 1964 del *Brown University Standard Corpus of Present-Day American English* (Brown Corpus), de Francis y Kučera, que es el primer corpus concebido y construido para residir en una computadora y ser explotado mediante programación

informática. Por tanto, tiene algo menos de medio siglo, más o menos la edad de ALFAL.

Una historia corta, pues. Corta y muy difícil en sus primeros tiempos, que coinciden casi exactamente con el primer período de expansión de la lingüística generativa, que en aquellos años mantenía, fundamentalmente como reacción a los planteamientos propios del distribucionalismo, una oposición radical a todo lo relacionado con hechos lingüísticos concretos, frecuencias, estadística, variación, comportamientos lingüísticos reales, etc. Las cosas han cambiado bastante y de algo de esto me ocuparé en el apartado siguiente. Me interesa ahora, en cambio, apuntar unas cuantas ideas rápidas que, como contribución a los orígenes de esta aproximación, nos permitirá entender mejor el modo en que se incardina actualmente en el mundo hispánico.

No se ha trabajado mucho el tema de los antecedentes de la lingüística de corpus. Probablemente, el factor fundamental de esta desatención se base en la idea, comprensiblemente extendida, de que la existencia de las computadoras es un factor ineludible, una condición *sine qua non* para la lingüística de corpus, de modo que no tendría sentido remontarse en el tiempo, ya que no puede haber nada que ofrezca interés más allá de lo que pueda implicar una conexión más o menos casual. Desde luego, muchos de los estudiantes de lingüística están convencidos de que las concordancias son una forma reciente de presentación de los materiales con los que vamos a trabajar, idea que deriva directamente del hecho de que se necesita una computadora, acceso a Internet, un corpus textual en formato electrónico, etc.

La realidad es, a mi modo de ver, bastante diferente. Nelson Francis, a quien podemos considerar, en principio, la persona más autorizada para buscar las raíces de la aproximación que él inauguró, señaló, hace ya algunos años (cf. Francis, 1992), los antecedentes de la que llamó, de forma realmente ingeniosa, 'lingüística a.C.', es decir, antes de la computadora. Como Francis parte de una definición de corpus en la que el empleo para propósitos de análisis lingüísticos es un elemento imprescindible, deja explícitamente a un lado conjuntos como, por ejemplo, el *Corpus Iuris Civilis* y se centra en tres grandes líneas clásicas: los corpus en lexicografía, dialectología y gramática. En la primera de ellas se refiere, como era de esperar, a los grandes proyectos lexicográficos del inglés: el diccionario de Johnson, el OED y el Merriam-Webster. Dedica luego mucha atención, a mi juicio un tanto sorprendente, a aquellos

estudios dialectológicos que, como el de Ellis, implicaron un enorme trabajo en la recogida de datos, mediante corresponsales más o menos expertos en muchos de los casos. Este tipo de reunión de materiales técnicos termina, siempre según Francis, en los atlas lingüísticos. Finalmente, en lo que respecta a corpus en estudios gramaticales, cita muy de pasada a autores como Jespersen, Kruisinga o Poutsma para centrarse finalmente en el *Survey of English Usage*, dirigido por Randolph Quirk y antecedente inmediato, como es bien sabido, del *Brown Corpus*. Algo no muy distinto se puede encontrar en Svartvik (2007), que repasa las mismas obras y dedica luego atención preferencial a la interesantísima relación que se produjo entre Francis y el equipo que colaboraba con Quirk en la construcción del SEU, equipo del que Svartvik formaba parte.

En mi opinión, ese panorama histórico es claramente mejorable en varios aspectos distintos. Creo, en primer lugar, que los acopios de ejemplos realizados con fines lexicográficos son un punto, de especial importancia sin duda, en una larga tradición que viene, en definitiva, del trabajo que los filólogos de la época alejandrina llevaron a cabo para fijar los textos homéricos con la mayor fidelidad posible, pasa luego por los esfuerzos renacentistas para recuperar la forma adecuada del latín y el griego, para lo cual hay que ir a los autores que pueden ser considerados como modelos y reunir las autoridades correspondientes. De ahí se salta a la recopilación de la obra de un autor especialmente importante, al que se aplica la misma técnica usada con los textos bíblicos y se conoce como concordancias. Son esas 'autoridades' las que están en el arranque de los usos lexicográficos, como muestra con claridad el primer diccionario de la RAE. Son autores seleccionados, obras elegidas las que constituyen el punto de partida de la documentación. En otras palabras, no se trata, como en el trabajo estrictamente técnico, de documentar los usos de una palabra, sino de comprobar si es utilizada o no y cuál es su uso en caso positivo, en un conjunto más o menos amplio, pero siempre seleccionado, de autores y obras.

No ofrece interés especial la remisión a estudios dialectológicos que hace Francis, puesto que, en el mejor de los casos, puede ser reducido a lo que se lleva a cabo en lexicografía y en gramática (o en fonética, si es lo adecuado). Olvida, en cambio, todo el trabajo relacionado con las listas de frecuencias, fundamentalmente léxicas, pero también de otros tipos, que se han realizado con diversos propósitos, enseñanza de lenguas en la mayor parte de las ocasiones. Elaborar una lista de frecuencias supone revisar y fichar el contenido de un conjunto más o menos amplio de obras para obtener

de ellas la información que interesa, que se mostrará luego debidamente procesada. Se construye un corpus y se procesa un corpus en el más estricto de las expresiones. Suele aludirse aquí, aunque Francis tampoco lo menciona, al sorprendente trabajo llevado a cabo por Käding a finales del siglo XIX. Con la ayuda de varios miles de colaboradores, procesó un corpus de unos doce millones de formas incluidas en textos alemanes con la finalidad de obtener las combinaciones de letras y sílabas más frecuentes y ayudar de esta forma a mejorar la taquigrafía de la época y su enseñanza.[1] Más cerca de nuestros objetivos y planteamientos actuales están las listas de frecuencias, léxicas o gramaticales, que se elaboran con regularidad para muy diversas lenguas, desde por lo menos los años veinte del siglo pasado. Se destaca habitualmente en este apartado la labor realizada por Thorndike, que publicó en 1921 la lista de palabras más frecuentes del inglés a partir del análisis de un corpus constituido por un total de 41 fuentes textuales y un total aproximado de 4,5 millones de palabras. Años más tarde, en 1944, colaboró con Lorge para analizar un corpus de 18 millones de formas y producir la lista de las 30 000 palabras más frecuentes del inglés.[2]

La segunda línea en la que creo que hay que completar la visión de Francis es mucho más general. Tanto él como los demás autores que hacen referencia a los antecedentes de la lingüística de corpus se quedan encerrados en el mundo anglosajón, en la tradición inglesa, sin aludir siquiera a otras tradiciones científicas y, probablemente, ignorando su existencia. Como no es posible dedicar aquí más tiempo a un tema que necesitaría mucha más atención, me limitaré a señalar, de pasada y solo para dejar constancia de ello, algunos antecedentes realmente destacados que se encuentran en la tradición hispánica.

Ya he aludido a los académicos redactores del llamado 'Diccionario de Autoridades', que papeletizan los ejemplos que consideran adecuados de autores y obras previamente seleccionados por su prestigio lingüístico y literario. Con un enfoque estrictamente lingüístico, la obra de Cuervo resulta modélica también en lo referente a la selección equilibrada de las obras y la utilización de los ejemplos que maneja en cada caso. En las listas de frecuencias hay que mencionar la obra de Juilland y Chang (1964), que se inscribe en un proyecto más general destinado, precisamente, a producir

---

[1] Para más detalles, vid. McEnery & Wilson (1996: 3), Kennedy (1998: 16) y la bibliografía que menciona. La obra de Käding, *Häufigkeitswörterbuch der deutschen Sprache*, fue publicada, de forma privada, en Berlín, en 1897-1898.

[2] Cf. Kennedy (1998: 16 y sigs.) y Berber Sardinha (2000: 325 y sigs.).

materiales de este tipo para diferentes lenguas románicas.[3] En cuanto a los estudios gramaticales, el reconocimiento de la excelencia del trabajo de Jespersen no puede impedirnos considerar la novedad que supuso el estudio de Cejador sobre la lengua de Cervantes, publicado antes que la obra de Jespersen. Como parte de un interesante proyecto aplicado a varias lenguas, Keniston publicó dos obras (1937a y 1937b) en las que facilita interesantísimos datos acerca de la frecuencia de una enorme variedad de estructuras sintácticas en la prosa española del XVI y luego en la de su época. Poca gente ha hecho un trabajo de ese tipo y, por tanto, no hay muchas personas que puedan hacerse cargo completamente de la naturaleza de las dificultades que presenta. He dejado para el último lugar, por factores de linealidad temporal, los miles y miles de fichas que escribió y con las que trabajó Salvador Fernández Ramírez. Muy posterior al de Jespersen, por supuesto, pero merecedor del más alto reconocimiento, como se ha hecho en los últimos años.[4]

Creo, pues, que, aceptando las ineludibles discrepancias derivadas de las servidumbres técnicas aplicables en cada caso y las diferencias debidas a los distintos objetivos fijados, no es difícil fijar una línea que relacione la construcción de corpus documentales como el *Corpus Iuris Civilis*, la elaboración de concordancias de los textos bíblicos destinadas a facilitar la localización de pasajes sobre contenidos determinados, la reunión de las obras de un autor considerado de especial relevancia en un cierto momento, la preparación de concordancias de esas obras, la extracción de ejemplos de uso de palabras o construcciones gramaticales de un conjunto de autores y obras previamente seleccionado, el análisis completo de obras para extraer la información estadística de interés en cada caso y el análisis, exhaustivo, de grandes masas de datos lingüísticos que llevamos a cabo actualmente en la orientación conocida como lingüística de corpus.

---

[3] A pesar del tiempo transcurrido desde su publicación, esta obra sigue siendo de referencia y manejo obligados, dado que trabajos más modernos como, por ejemplo, Alameda y Cuetos (1995) o Davies (2006) resultan menos útiles en algunos aspectos a pesar de estar basados en corpus mucho más amplios. El primero, extraído de un corpus de unos dos millones de palabras formado por fragmentos de 6100 palabras cada uno procedentes de un variado conjunto de textos contiene listas de formas, sílabas combinaciones de letras, etc., pero no está lematizado. El segundo, destinado a facilitar el aprendizaje del español como L2, está construido sobre un conjunto de unos veinte millones de formas y contiene únicamente los 5000 lemas (sin indicación de las formas asociadas) más frecuentes, con los índices de frecuencia correspondientes. El de Juilland y Chang resulta del análisis de un corpus de aproximadamente medio millón de formas. Contiene la lista de los 5024 lemas más 'frecuentes' (según el conjunto de factores utilizados por los autores: frecuencia, dispersión y uso) y las formas asociadas a ellos. El inventario inicial de lemas produjo alrededor de 20 000, que fueron disminuyendo en sucesivas operaciones de reducción. Para más detalles, cf. Juilland y Chang (1964: lxxiv-lxxvi).

[4] Una parte de este enorme fichero puede ser consultado en http://cvc.cervantes.es/obref/agle.

Naturalmente, hay también diferencias claras entre todos estos antecedentes y nuestra práctica actual, pero me parece del mayor interés señalar antes que las líneas de trabajo que he mencionado de pasada se agrupan en dos conjuntos bastante bien diferenciados y que el reconocimiento de esas dos aproximaciones distintas nos permitirá entender mejor la naturaleza de las diferencias que la lingüística de corpus muestra con respecto a todo lo anterior.

La primera de esas líneas es la constituida por las concordancias de obras (la Biblia, Shakespeare, etc.) y las listas de frecuencias de palabras, construcciones gramaticales, etc. Con las diferencias esperables entre estas dos variantes, se trata siempre de examinar de forma exhaustiva el contenido de unos textos determinados y volcar esa información de modo que pueda ser utilizada por otras personas. Es, pues, el análisis completo de un cierto número de textos realizado con la intención de conservar esa información de forma estable y duradera: las concordancias de la obra de Virgilio, por ejemplo, nos permiten saber si este autor utiliza o no una determinada palabra, con qué frecuencia lo hace, en qué obras, etc. y tener, por tanto, los materiales necesarios para saber qué significado(s) posee, etc. Lo mismo, *mutatis mutandis*, con respecto a una construcción gramatical, un conjunto de autores, etc. La diferencia fundamental con lo que se hace hoy en las fases iniciales del trabajo con un corpus consiste en que, gracias al formato electrónico, hoy manejamos los textos completos y podemos obtener con gran comodidad y rapidez la información que nos interesa en cada caso. La transición entre la formulación tradicional y la actual está, con toda claridad, en los primeros trabajos informatizados, en los que los procesos eran tan lentos y costosos que la información (las concordancias, por ejemplo) eran extraídas mediante computadora, pero los resultados se imprimían para poder hacer luego las consultas pertinentes.[5] La conexión está, pues, en el carácter exhaustivo del análisis que se practica y la diferencia radica en la necesidad de almacenar la información obtenida (que solo se podría repetir realizando de nuevo todo el trabajo anterior). Esa diferencia, crucial, se basa en la posibilidad de almacenar textos (e información si se desea) en formato electrónico, lo cual significa que podemos hacer en diferentes momentos las búsquedas que nos interesan.

---

[5] El pionero fue, como es bien sabido, Roberto Busa, que, ya en 1949, comenzó la labor de información necesaria para producir las concordancias de la obra de Tomás de Aquino. La producción de índices (listas de formas con indicación de su frecuencia y lugares en los que se localiza cada una de sus apariciones) representa una etapa intermedia que tiene la ventaja de reducir el volumen de la publicación a cambio de requerir la consulta de las obras examinadas para poder acceder al texto.

La segunda línea, que es la que habitualmente se ha considerado más próxima al trabajo con corpus, reúne cantidades significativas de ejemplos de uso de palabras, construcciones gramaticales, etc. con la intención de proporcionar los materiales necesarios para elaborar diccionarios, tratados gramaticales, monografías, etc. Esta aproximación, que es la característica de la que, por comodidad, voy a llamar lingüística descriptiva tradicional, hace, de forma casi inevitable, una selección inicial de las documentaciones, esto es, aplica un filtro más o menos fuerte, lo cual produce diversos problemas. En primer lugar, puede haber errores como resultado de las distracciones, cansancio, etc. de las personas que hacen la recogida inicial de datos. En segundo término, es esperable que, en las fases iniciales de la recopilación, se desconozcan los factores realmente relevantes del fenómeno en cuestión, con lo que no existe garantía de que la selección realizada sea, en todos sus estadios, la que realmente debería haberse hecho, es decir, la que permitiría realmente estudiar todas las caras de esa palabra o construcción. Por fin, como efecto de la actuación de un mecanismo natural, está perfectamente claro que estos procesos selectivos tienden a prestar mayor atención a los casos menos habituales, a aquellos que muestran aspectos marginales o simplemente menos representativos del fenómeno estudiado, con lo que la selección de materiales resultantes puede dar –al menos desde el punto de vista cuantitativo– una visión bastante diferente de la que realmente presenta el comportamiento lingüístico real.

La diferencia es, me parece, fundamental y proporciona el modo de entender y explicar las luces y sombras de la consideración de la lingüística basada en el análisis de corpus como la versión contemporánea de la lingüística descriptiva tradicional. La posibilidad de poder examinar, cuando es necesario, **todos** los casos pertinentes de un determinado fenómeno, de recuperar aquellos que se nos han podido escapar en un análisis preliminar, etc. es la base sobre la que se levantan las fuertes diferencias metodológicas existentes. En otras palabras, la posibilidad de alcanzar la *total accountability* a que se refieren, entre otros, Leech (1992: 112) y Quirk (1992: 467), esto es, la posibilidad de dar cuenta del comportamiento que los elementos estudiados muestran en todos los casos registrados en el corpus. Para decirlo de nuevo con palabras de Quirk, los tratados de gramática o los diccionarios basados en el uso real nos aseguran que todo lo que figura en ellos se da en la lengua, pero no pueden garantizar que contengan todo lo que se da en la lengua, ni siquiera lo que se encuentra en las

secuencias que contienen los ejemplos seleccionados.[6]

Naturalmente, esa *total accountability* es un objetivo que podría ser perseguido antes de la constitución de la lingüística de corpus en sentido estricto, bien por la vía de la recogida de todos los casos de un cierto elemento o fenómeno gramatical en un conjunto de textos que puede alcanzar una extensión notable (por ejemplo, todas las oraciones condicionales en un número amplio de textos de un cierto período), bien por la del fichado de todos los fenómenos –o, al menos, todos los de un cierto ámbito– en un conjunto relativamente reducido de textos.

La primera de estas dos líneas tiene escaso interés para nuestros propósitos, puesto que la diferencia con la aproximación más tradicional es puramente cuantitativa y está condenada por su propia naturaleza a darse en un ámbito reducido. La segunda, en cambio, muestra un interesante carácter de puente entre la lingüística descriptiva tradicional y la lingüística de corpus tal como la entendemos actualmente. El proyecto que materializa la transición es, como se ha puesto repetidamente de relieve, el *Survey of English Usage*. Iniciado bajo la dirección de Randolph Quirk en 1959, contiene en su forma final un millón de palabras (200 textos de 5000 formas cada uno) de inglés británico oral y escrito del período comprendido entre 1955 y 1985. Su versión inicial, no informatizada, consiste en miles y miles de fichas con anotaciones detalladas sobre los más diversos fenómenos gramaticales que sirvieron de base para la confección de *A Grammar of Contemporary English* (Quirk, Greenbaum & Svartvik, 1972) primero y *A Comprehensive Grammar of the English Language* (Quirk, Greenbaum, Leech & Svartvik, 1985) después, entre muchas otras. Aunque no fue concebido para estar contenido en una computadora, el SEU es el primer corpus construido de modo comparable a como se hace hoy.[7]

Además de este carácter evidente de elemento de engarce entre las dos etapas, al que volveré dentro de un momento, el SEU tiene otra característica muy marcada, fuertemente rupturista en la época: la atención prestada a la lengua hablada y, sobre todo, su consideración a un nivel semejante al otorgado a la lengua escrita, en la que

---

[6] Según Quirk, "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).

[7] Para más detalles sobre la historia del proyecto, cf. http://www.ucl.ac.uk/english-usage/about/history.htm y también los detalles facilitados por el propio Lord Quirk en http://www.ucl.ac.uk/english-usage/about/quirk.htm.

entran, además de materiales impresos, noticias de radio (leídas) y discursos. El SEU, por tanto, representa el punto de coincidencia de varias tendencias metodológicas diferentes entre las que interesa destacar la atención a la lengua realmente empleada (tanto oralmente como por escrito) y el registro completo de los fenómenos contenidos en los textos incorporados. No hay aquí selección de ejemplos, ni predominio de la lengua escrita, ni reducción del interés hacia la variedad estándar.

El salto hacia lo que entendemos habitualmente por lingüística de corpus se da muy poco tiempo después, cuando, en un proyecto en el que Quirk participa como asesor, se construye un corpus, concebido ya para ser introducido en una computadora de la época, formado por un millón de palabras (500 muestras de aproximadamente 2000 formas cada una) de inglés estadounidense, procedentes todas ellas de textos impresos y publicados en Estados Unidos en 1961: el *Brown University Standard Corpus of Present-Day American English*.

El *Brown Corpus*, basado en algunos rasgos estructuradores del SEU,[8] establece las que serán características definitorias de los primeros corpus textuales construidos ya para computadoras: dado que el tamaño no puede ser muy grande por las limitaciones de las máquinas de la época, la representatividad ha de conseguirse a base de reunir un conjunto relativamente amplio de muestras de tamaño relativamente reducido (nótese que, con respecto al SEU, la igualdad del total de formas se consigue aumentando el número de muestras y reduciendo el tamaño de cada una de ellas). Las demás diferencias con el SEU están igualmente claras: solo textos escritos y atención al inglés estándar.

Mi intención en esta primera parte es, como ya he señalado, complementar la presentación habitual de los antecedentes de la lingüística de corpus con lo que se puede observar cuando se considera desde la lingüística hispánica. Y, precisamente para cerrarla, me gustaría añadir aquí la referencia a un proyecto muy vinculado con ALFAL, que representa una línea hasta cierto punto paralela a la del SEU. Me refiero, naturalmente, al proyecto de *Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica*, propuesto por Juan M. Lope Blanch, en el simposio de Bloomington, que tuvo lugar precisamente en 1964. Por supuesto, procede de una orientación distinta, pero el proyecto resultante encaja bien en la línea del SEU: aquí se pretende construir un gran corpus (sin utilizar ese nombre) de

---

8 Para las conexiones entre ambos proyectos es de gran interés la visión de Svartvik (2007)

materiales orales en el que fuera posible estudiar los más diversos fenómenos y, sobre todo, contrastar las características que presentan en la lengua de diferentes ciudades del mundo hispánico. Quizá sea este marcado interés en la variación y la complejidad de recopilación que introduce lo que lo orienta hacia la producción de monografías especializadas y también el factor fundamental que impidió que se convirtiera en el paralelo al SEU (o al *Brown Corpus*) para el mundo hispánico. Ha habido que esperar hasta 1998 (cf. Samper, Hernández y Troya, 1998) para disponer de una muestra reducida, organizada al estilo sociolingüístico, de lo que ese corpus pudo haber representado. No me parece que sea casual ni irrelevante el hecho de que la continuación natural del SEU sea el *International Corpus of English* (ICE), dirigido por Sidney Greenbaum y constituido por la agrupación de un millón de formas procedentes de textos orales y escritos procedentes de cada uno de los países en los que el inglés funciona como lengua oficial o principal.

3. La búsqueda de antecedentes sirve también para poner de relieve las diferencias. La fundamental, a la que me he referido repetidamente, la que produce el corte metodológico, tiene una marcada dependencia de los computadores. Esa es la razón de que algunos hayamos insistido con tanta frecuencia en que el hacer figurar la alusión al formato electrónico en la definición de corpus lingüísticos no es una especie de guinda tecnológica que añadimos para dar mayor relieve a nuestro trabajo, sino un elemento imprescindible, puesto que, en efecto, la existencia del formato electrónico y la posibilidad de almacenamiento y explotación es lo que hace realmente posible esta aproximación al estudio de los fenómenos lingüísticos y le confiere sus características más marcadas.

Este hecho, que hoy nos parece tan marcadamente positivo, tuvo un carácter muy diferente para una buena parte de los lingüistas durante los primeros años de vida de la lingüística de corpus. El rechazo de la lingüística de orientación chomskyana hacia el trabajo con corpus –electrónicos o no– fue muy intenso y hunde sus raíces en las discrepancias básicas con la metodología dominante en los años anteriores. El interés en la construcción de corpus que pudieran ser analizados automáticamente (hasta cierto punto), la consideración de lo que se puede encontrar en un corpus como representativo de lo que sucede en la variedad lingüística de la que procede, el interés en los datos que se encuentran en los hechos lingüísticos reales, la cuantificación de los resultados, etc. chocaban de modo bastante violento con la perspectiva, que empezaba a imponerse en Estados Unidos, según la cual el objetivo es el conocimiento lingüístico del hablante-

oyente ideal, el método adecuado es la introspección, etc. Aunque, como era de esperar por otra parte, el pensamiento de Chomsky sobre todas estas cuestiones cambia de forma notable a lo largo de estos primeros años (cf. Karlsson: en prensa), la larga y compleja tradición existente en torno a estos puntos destaca sistemáticamente dos citas suyas que parecen representar la opinión, –fuertemente negativa– que el trabajo con corpus merece a la lingüística generativa en su primera fase. Según la primera de ellas, para Chomsky,

> [a]ny natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.[9]

Además, al negar la validez del modelo de estados finitos como forma de construir la gramática, incluso con el refinamiento de introducir en él las probabilidades de los diferentes elementos en cada estado, tal como había propuesto Hockett, rechaza Chomsky la utilidad de los aspectos estadísticos ya que, según dice,

> this is an irrelevant complication. It seems that probabilistic considerations have nothing to do with grammar, e.g. surely is not a matter of concern for the grammar of English that 'New York' is more probable than 'Nevada' in the context 'I come from__.' In general, the importance of probabilistic considerations seems to me to have been highly overrated in recent discussions of linguistic theory (Chomsky, 1962: 215, nota 10). [10]

Esta oposición se plasma en la conocida discusión, a comienzos de los sesenta, entre Robert J. Lees, entonces uno de los más decididos partidarios de la entonces novedosa gramática generativo-transformacional, y W. Nelson Francis. Cuando este último, a una pregunta acerca de sus actividades en aquel momento, respondió que se ocupaba de construir un corpus electrónico, Lees respondió indignado que era una lástima perder tiempo e invertir dinero en un recurso que cualquier hablante de inglés podía mejorar en un par de minutos de reflexión acerca de su lengua.

Visto desde hoy, es fácil llegar rápidamente a la conclusión de que Chomsky y

---

[9] Tomo la cita de Leech (1991: 8), que la remite a la pág. 159 de una comunicación que Chomsky presentó a la *3rd Texas Conference on Problems of Linguistic Analysis* celebrada en 1958. No he podido consultar las actas de esa conferencia (= Hill, 1962) y tampoco he conseguido localizar el texto mencionado en la reimpresión que figura en Fodor y Katz (1964: 211-245). La cita aparece en muchos otros lugares (cf., p.e., McEnery & Wilson, 1996: 8, con una acotación adicional entre corchetes), casi siempre con referencia a la conferencia de Texas.

[10] La tradición a la que aludo un poco más arriba no opone Nueva York a Nevada, sino a Dayton (Ohio) (cf., por ejemplo, McEnery & Wilson (1996: 8), pero el único texto que yo he podido encontrar es el que reproduzco aquí. Cf. Stefanowitsch (2005: nota 1) para más detalles sobre estas discrepancias textuales.

Lees estaban muy equivocados y que, por el contrario, Francis y Kučera habían iniciado el camino correcto. Afortunadamente, las cosas son bastante más complejas y la discusión anterior no podría tener lugar en nuestros días por varias razones, que apuntan en direcciones diversas. Parece claro, en primer lugar, que las reticencias de Chomsky y sus seguidores derivaba de la asimilación de los corpus electrónicos de los que se comenzaba a hablar en estos años con el corpus al que solían hacer referencia los distribucionalistas. Es, por supuesto, una asimilación errónea: se trata de dos tradiciones y dos metodologías totalmente distintas, como estaba claro ya desde el principio. Los corpus de los que se habla en cada caso son diferentes y, sobre todo, los objetivos establecidos son totalmente distintos. La lingüística de corpus no ha aceptado nunca que el trabajo esté terminado en el momento en que se pueda considerar que lo que hay en el corpus ha sido suficientemente descrito.[11]

En segundo término, la famosa afirmación de Chomsky acerca del sesgo inevitable de cualquier corpus puede ser aceptable únicamente cuando se trata de corpus de tamaño muy reducido, como nos parecen hoy los que se construían en aquellos años. Ese problema no se produce cuando trabajamos con un corpus de doscientos o trescientos millones de formas, bien construido y debidamente equilibrado.

La preocupación por construir corpus representativos y equilibrados, cuestión sobre la que tanto se ha escrito, es, por otro lado, una indicación muy clara de cómo se concibe realmente el corpus. Es cierto que, como acabo de señalar, hoy trabajamos habitualmente con conjuntos textuales que son cientos de veces mayores que el *Brown*

---

[11] Para más datos sobre la desconexión entre el corpus de los distribucionalistas y la concepción habitual en lingüística de corpus, vid., por ejemplo, Leech (1991) y Caravedo (1999: 38 y sigs.). La cuestión es muy compleja y no puedo ocuparme aquí de sus muchas implicaciones. Apuntaré simplemente que su comprensión completa requiere, al menos, la consideración de tres factores. En primer lugar, hay que tener en cuenta que los distribucionalistas hablan de construir, segmentar y describir un corpus obtenido de lenguas de las que no hay muchos más datos (al menos, no hay más datos a mano) que los que han sido incluidos en ese corpus. Puede vincularse este punto al hecho, evidente entre nosotros, de la tradición de trabajo con corpus o sus equivalentes que existe entre los especialistas en lenguas clásicas, por ejemplo, o en trabajos sobre etapas anteriores de lenguas, a cuyo conocimiento solo se puede acceder mediante el análisis de la documentación conservada. En segundo lugar, creo que podría discutirse también que la concepción del corpus entre los distribucionalistas fuese –al menos, sistemáticamente– del estilo rechazado por Chomsky. Por dar un ejemplo, en un texto de Hockett sobre el que llama la atención Leech (1991: 25-26, nota 1) se establece con toda claridad que el objetivo del lingüista estructural “is not simply to account for all utterances which comprise his corpus”, sino que "the analysis of the linguistic SCIENTIST is to be of such a nature that the linguist can account for utterances which are NOT in the corpus at a given time” (Hockett, 1948: 269; elementos destacados en el original). Por último, conviene tener en cuenta, en un ámbito de mayor generalidad, que la consideración del distribucionalismo habitual en la Europa de los años 60 y 70 se basó en un conocimiento muy superficial de las contribuciones de esta orientación y fue determinada más por la interpretación (y la crítica) que de ella hacía el generativismo que por el conocimiento directo de la producción científica de sus defensores.

Corpus, pero también lo es que todos los que nos movemos en ese terreno estamos convencidos de que un corpus, por muy grande que sea, no puede contener todo lo que es posible en una lengua, un estado de lengua o una variedad.

Considerando ahora la cuestión desde el otro lado, los años transcurridos desde las primeras formulaciones chomskyanas han dejado clara la necesidad de aceptar concepciones mucho más elaboradas, menos ingenuas, de, por ejemplo, la competencia lingüística de los hablantes, el carácter gramatical o agramatical de una secuencia o el papel de la introspección. La competencia es mucho más compleja y, sin duda, incluye componentes que los generativistas despreciaban en los primeros años. Como consecuencia de ello –o, si se prefiere, en paralelo–, está claro que la línea de atribuir carácter gramatical o agramatical a una secuencia, tarea que fue considerada la fundamental de la lingüística durante algunos años, resulta excesivamente simplista, dado que parece claro que no puede concebir de forma binaria, sino que existe un cierto número de zonas intermedias entre los extremos. En conexión con lo anterior, la introspección de la que se hablaba en los primeros tiempos no puede ser concebida como el resultado de los juicios que una persona hace en un momento determinado acerca de secuencias de su propia lengua. Tal como se hace en psicología cognitiva, hay que recoger, procesar y valorar los diferentes juicios que distintos hablantes de esa lengua hacen acerca de las secuencias sometidas a análisis, con lo cual obtenemos una panorámica infinitamente más compleja, repleta de posibilidades intermedias que se nos presentan sin solución de continuidad y en la que, por cierto, las valoraciones estadísticas juegan un papel fundamental .[12]

En otras palabras, la conversación entre los herederos intelectuales de Lees y Francis tendría lugar hoy en términos bastante diferentes de los originales, aunque me temo que no se produciría un acuerdo rápido y consistente. Ambas orientaciones (empirista y racionalista, para usar las denominaciones habituales, cómodas, aunque no del todo adecuadas) han evolucionado, se han hecho mucho más complejas y han experimentado numerosos procesos de fragmentación y reunificación, pero sigue habiendo diferencias entre ellas. A pesar de ello, me gustaría cerrar este punto recordando una cita de Charles Fillmore, protagonista él mismo y sujeto paciente de las luchas internas entre corrientes metodológicas, que hace ya bastantes años oponía la

---

[12] Vid., por ejemplo, Kepser & Reis (2005), el artículo de Sampson (2007) sobre la gramaticalidad y las reacciones publicadas en el mismo número de *CLLT* , Featherston (2005) sobre juicios de mayor o menor naturalidad de secuencias o bien, sobre el papel de los datos en la gramática generativa, Featherston (2007), acompañado también de réplicas.

perspectiva de la lingüística de sillón o de gabinete a la lingüística de campo y concluía acerca de su propia experiencia de trabajo con corpus:

> 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 observations 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).

4. Teniendo en cuenta todas las precauciones necesarias a partir de lo anterior, se puede aceptar que la lingüística de corpus, de orientación empirista, es la heredera, evolucionada, de la lingüística descriptiva, tal como se practicaba en el tercer cuarto del siglo XX. Es preciso resaltar que ese vínculo, que creo innegable, no puede hacernos olvidar las fuertes diferencias existentes entre la vieja lingüística descriptiva y la reciente lingüística de corpus, diferencias que surgen precisamente de la posibilidad de acceder de forma cómoda y rápida al contenido de miles y miles de textos, con todo lo que ello implica. No se trata solo del simple crecimiento en el número de textos o de ejemplos que podemos estudiar. Es, por el contrario, uno de esos casos en los que el crecimiento cuantitativo produce un salto cualitativo. En este terreno, gracias a las ventajas tecnológicas, se pasa del fichero de ejemplos seleccionados de una palabra o construcción, más o menos amplio, a la posibilidad de trabajar con todos los que se pueden localizar en un corpus de cientos de millones de formas.

Ahí radica, a mi modo de ver, el origen de las diferencias metodológicas que existen entre la lingüística descriptiva tradicional y la lingüística de corpus. El filtro, la selección de casos, suponen una cierta distorsión, que se produce inevitablemente por muy grande que sea el tamaño del fichero de papeletas. Un fichero de ejemplos no es un corpus, aunque eso no significa que no pueda intentarse su explotación como tal. [13] Como se deduce de lo anterior, el tamaño de los corpus es una cuestión fundamental, aunque no la única, puesto que un corpus enorme puede resultar poco menos que

[13] Por ejemplo, Mair (2004) ha estudiado la gramaticalización de ciertas construcciones del inglés usando como material de trabajo (esto es, como corpus) la versión informatizada de los ficheros del OED. Debe tenerse en cuenta que no existe para el inglés un corpus con las características del CORDE, de modo que el espléndido fichero del OED proporciona muchos más datos que los corpus diacrónicos existentes para esa lengua, aunque, por supuesto, no deja de presentar los inconvenientes de un fichero de este tipo. Por otro lado, el hecho de que el fichero esté informatizado permite hacer búsquedas sobre la totalidad del texto de las citas que contiene, con independencia del lema al que pretendiera adscribir cada una de ellas la persona que seleccionó el fragmento correspondiente.

inservible si no ha sido construido de modo que resulte representativo y esté equilibrado. La pregunta acerca del tamaño más adecuado de un corpus se responde habitualmente de modo muy claro: lo más grande posible. La cuestión es, sin embargo, un tanto más complicada. De un lado, sabemos que, frente a los fichados selectivos tradicionales, los corpus muestran sistemáticamente muchos casos de unos pocos elementos o fenómenos y pocos casos (o ninguno) de muchísimos otros. Cabe, por tanto, preguntarse acerca de la conveniencia de invertir recursos humanos y económicos en construir conjuntos textuales que, a partir de un determinado punto, podrían ofrecer muy escasas novedades y verse reducidos a acumular más y más casos de los mismos fenómenos. De otra parte –pero también como complemento a lo anterior– en los últimos años ha aparecido la posibilidad de trabajar directamente con las ingentes cantidades de textos de los más diversos estilos y procedencias que se pueden localizar directamente en Internet. La utilización de la *Web* como un gran corpus es una posibilidad que está recibiendo notable atención en estos últimos años y tiene partidarios decididos. Resulta discutible, por tanto, la conveniencia de construir corpus generales, que nunca van a alcanzar el tamaño que tiene la red.

La llamada 'ley de Pareto', conocida también como 'ley del 80 / 20' establece que en la mayoría de las distribuciones el 80% de los efectos se puede explicar con el 20% de las causas. Se aplica a la distribución de rentas de un país, la cuenta de resultados de una empresa, las ventas de libros, etc. En el caso de los corpus textuales, el reparto resulta mucho más radical, como muestran con toda claridad los datos que figuran en la tabla 1. Para alcanzar el 80% de todas las formas de la parte escrita del BNC es suficiente con sumar las frecuencias correspondientes a las 3000 formas más frecuentes de ese subcorpus, es decir, el 0,79% del total de formas distintas. Para alcanzar el mismo porcentaje del total, español y gallego requieren unas 5000 formas, lo cual supone el 0,68% en el primer caso y un 1,30% en el segundo (hay que tener en cuenta que es un corpus notablemente más pequeño). En números muy gruesos, pero significativos, se puede pensar que en torno al 1% de las formas distintas documentadas en un corpus es responsable de aproximadamente el 80% de las frecuencias totales.[14]

---

[14] Con los cambios esperables por la agrupación de formas, lo mismo sucede con los lemas. Como dato ilustrativo, la frecuencia acumulada de los mil lemas con los índices más altos en el recuento de Davies (2006) supone, aproximadamente, el 80% del total del corpus manejado para elaborar el diccionario. Es un fenómeno que se puede observar en todos los ámbitos. El corpus manejado para elaborar la BDS (www.usc.bds.es) proporciona documentación de un total de 3437 verbos en función de predicado de otras tantas cláusulas. Los 172 más frecuentes (es decir, el 5% del total) producen una frecuencia acumulada del 67% del total. En esa misma base de datos se comprueba que los cinco

Las *n* formas más frecuentes	BNC_90 (377 384 formas ortográficas distintas)	CREA_150 (737 799 formas ortográficas distintas)	*CORGA_24 (385 291 formas 76,51 ortográficas distintas)*
	suponen en conjunto el		
10	23,48%	28,65%	*23,29%*
25	32,30%	39,55%	*32,64%*
50	39,86%	43,94%	*38,15%*
100	46,66%	48,26%	*43,85%*
150	50,26%	51,08%	*47,24%*
500	61,35%	60,16%	*57,70%*
1000	68,77%	66,21%	*64,06%*
3000	**80,04%**	76,51%	*74,75%*
5000	86,02%	**81,22%**	***79,56%***
10 000	91,48%	87,08%	*85,61%*
15 000	93,94%	90,03%	*88,68%*
25 000	96,35%	93,17%	*92,01%*
50 000	*98,44%*	*96,35%*	*95,52%*

Tabla 1

Porcentaje que suponen las formas más frecuentes en distintos corpus textuales.
Fuentes: De la parte escrita del *British National Corpus* (BNC), lista de formas y frecuencias elaborada por Mike Scott (http://www.liv.ac.uk/~ms2928/homepage.html).
Del *Corpus de Referencia del Español Actual* (CREA) en la configuración que presenta en julio de 2008, con unos 152 millones de formas, www.rae.es.
Del *Corpus de Referencia do Galego Actual* (CORGA), en la configuración que presenta en julio de 2008, con unos 24 millones de formas, www.cirp.es (Centro *Ramón Piñeiro* para a investigación en Humanidades).
Elaboración propia en los tres casos.

Este hecho, conocido desde hace mucho tiempo, explica las expectativas pesimistas que existían acerca de las consecuencias del aumento del tamaño de los corpus en los primeros años de desarrollo de esta orientación. En 1967, todavía en plena infancia de la lingüística de corpus, John B. Carroll afirmó que la relación entre *types* y *tokens* presenta una distribución log-normal, de modo que el aumento en el número de formas distintas tendería a reducirse fuertemente a medida que aumentara el número de formas totales (es decir, el tamaño de los corpus). En palabras de Kučera (1992: 407), según Carroll, "the number of new lexical items as the size of the text increases gradually slows to a trickle, to reach, for example, just barely over 200,000 in a sample

esquemas sintácticos más frecuentes suponen, en conjunto, el 68,59% del total de las cláusulas fichadas (cf. Rojo, 2003).

of 100 million tokens". Como muestra la tabla 1, la predicción de Carroll era errónea, ya que la parte escrita del BNC, que consta de unos noventa millones de formas, contiene 377 384 formas distintas, esto es, casi el doble de las supuestas por Carroll. Mucho más ajustado parece el procedimiento propuesto por Sánchez y Cantos (1997), según el cual el número de formas diferentes en un corpus es equivalente a la raíz cuadrada del número total de formas multiplicada por una constante cuyo valor ha de ser extraído del análisis de muestras homogéneas, relativamente pequeñas, de textos del mismo tipo que los que van a componer el corpus total. En el caso del español, el valor de la constante calculada por ellos es de 56,17 para textos periodísticos y de 51,45 para textos de ficción. En la tabla 2 aparecen los datos correspondientes a diferentes segmentos del CREA –véase la tabla 4 para más detalles sobre su construcción–, con el resultado de aplicarles la fórmula de Sánchez y Cantos dando a la constante el valor de la media entre los dos establecidos por ellos (53,81), el número real de formas distintas y la desviación de la predicción con respecto a la realidad.

Tamaño del corpus	Formas distintas previstas	Formas distintas halladas	Porcentaje de desviación de la cifra supuesta con respecto a la real
1 602 351	68114,852	68 468	-0,51578488
3 172 859	95849,175	96 623	-0,80087056
6 885 997	141203,81	149 565	-5,5903405
13 838 517	200174,05	218 743	-8,48893624
27 798 451	283709,12	320 549	-11,4927465
53 319 062	392920,1	440 682	-10,8381788
117 070 367	582219,14	644 841	-9,7112087
147 180 549	652812,11	717 149	-8,97120282
152 558 294	664631,47	737 799	-9,91699986

Tabla 2. Comparación entre el número de formas distintas de un corpus de español contemporáneo según la hipótesis de Sánchez y Cantos (1999) y las que aparecen en diferentes segmentos del CREA. Cf. tabla 4 para más detalles sobre la organización de los diferentes segmentos. Elaboración propia.

Es fácil observar que aparecen desajustes de entidad notable a partir de un tamaño del corpus relativamente bajo. Por tanto, parece necesario ajustar el valor de la constante establecida por estos autores para adaptarla a un corpus con las características del CREA (textos de características y procedencias variadas, aunque pertenecientes todos ellos al español contemporáneo). Si damos el valor 60 a la constante, el resultado es el que muestra la tabla 3, mucho más ajustado.

Tamaño del corpus	Formas distintas previstas	Formas distintas halladas	Porcentaje de desviación de la cifra supuesta con respecto a la real
1602351	75950,402	68468	10,9283201
3172859	106875,13	96623	10,6104398
6885997	157447,1	149565	5,27001617
13838517	223200,94	218743	2,03798227
27798451	316345,42	320549	-1,3113695
53319062	438119,42	440682	-0,58150399
117070367	649194,36	644841	0,67510645
147180549	727907,95	717149	1,50023844
152558294	741086,94	737799	0,44564223

Tabla 3. Comparación entre el número de formas distintas de un corpus de español contemporáneo según la hipótesis de Sánchez y Cantos (1999) con el valor de la constante modificado a 60 y las que aparecen en diferentes segmentos del CREA. Cf. la tabla 4 para más detalles sobre la organización de los diferentes segmentos. Elaboración propia.

Según esta hipótesis, con resultados que parecen razonables, al menos hasta un tamaño de unos 150 millones de formas, el CORPES, que constará de trescientos millones de palabras (cf. infra), contendrá algo más de un millón de formas distintas.[15]

Además del interés general que posee cualquier mejora de nuestros conocimientos, los datos anteriores tienen toda la importancia que deriva del hecho de que permiten deducir, creo que con toda claridad, que es necesario seguir construyendo corpus y que su tamaño debe ser lo más grande que podamos conseguir siempre que ello signifique sacrificar otros aspectos más importantes. Visto de este modo, lo realmente interesante está en tratar de encontrar la causa de este incremento constante y de entidad no despreciable que muestra el número de formas distintas en relación con el aumento del tamaño de los corpus. La explicación, que estimo clarísima, está en otro hecho,

[15] Los datos numéricos correspondientes al CREA que presento aquí, y que condicionan el valor corregido de la constante que se propone en el texto, se compadecen mal con los procedentes de otros corpus del español de gran tamaño. Según los datos que proporciona Scott Sadowsky, responsable del *Corpus dinámico del castellano de Chile,* "[l]a LIFCACH [=*Lista de frecuencias de palabras del castellano de Chile*, G. R.] contiene 477 293 lemas, derivados de aproximadamente 4,5 millones de *types* extraídos de los 450 millones de palabras de texto corrido que contemplaba el CODICACH al momento de elaborar la LIFCACH" (http://www2.udec.cl/~ssadowsky/lifcach.html [comprobado en julio de 2008]). Por otro lado, un corpus de textos españoles tomados de Internet elaborado en la Universidad de Leeds arroja un total de 1 994 619 formas ortográficas distintas (*types*) procedentes de un total de 143 millones de formas (cf. http://corpus.leeds.ac.uk/frqc/internet-es-forms.num [comprobado en julio de 2008]). Para valorar adecuadamente estas diferencias, debe tenerse en cuenta que en mis recuentos del CREA no he tomado en consideración signos de puntuación ni cifras y, además, he reducido la diferencia entre mayúsculas y minúsculas, factores que, en cambio están presentes en los otros recuentos. Hacer algo similar en el CREA hace pasar de 737 799 formas distintas a 1 129 443, aumento más que considerable, pero que da una cifra todavía muy alejada de la que aparece en el corpus de Leeds, que tiene un tamaño similar al CREA. No tengo explicación para la discrepancia.

también conocido desde hace tiempo, pero que no ha sido puesto en relación con el anterior: la alta frecuencia de formas que tienen frecuencia igual a 1, es decir, los llamados, adaptando la expresión de los estudios clásicos, *hápax legómena*. En efecto, se sabe que un alto porcentaje de las formas que componen un texto aparece una sola vez y la llamada 'ley de Zipf' lo tiene debidamente en cuenta. Lo que sucede, a mi modo de ver, es que teníamos la impresión de que el porcentaje de hápax debería ir reduciéndose a medida que el corpus aumentara de tamaño. Dicho de otra forma, suponíamos que un alto número de formas de frecuencia igual a uno con un corpus de tamaño x pasarían a ser formas de frecuencia igual a dos con un corpus de tamaño 2x. Algo de eso sucede, sin duda, pero ocurre también que la entrada de nuevos textos produce la entrada de un número igualmente alto de formas con frecuencia igual a uno.

Para comprobar el modo en que este fenómeno tiene lugar, emprendí hace ya algún tiempo un experimento con los textos del CREA que ahora, con el corpus ya cerrado, puedo presentar en su configuración final. Los datos, completos, son los que aparecen en la tabla 4.

Datos de la parte escrita del CREA (situación en abril de 2008)							
			Formas diferentes			*Hápax*	
Núm. ficheros	MBytes	Núm. total de formas	Núm. formas diferentes	% formas dife-rentes	1 forma diferen-te cada	Total	% sobre formas diferentes
25	9,7	1 602 351	68 468	4,27	23,4	29 440	42,9
50	19,1	3 172 859	96 623	3,04	32,8	39 809	41,2
150	41,5	6 885 997	149 565	2,17	46,0	60 403	40,4
310	83,0	13 838 517	218 743	1,58	63,3	86 824	39,7
750	166,6	27 798 451	320 549	1,15	86,7	127 649	39,8
1500	318,6	53 319 062	440 682	0,82	121,0	179 607	40,7
3212	700,7	117 070 367	644 841	0,55	181,5	271 615	42,1
4188	905,7	147 180 549	717 149	0,49	205,2	303 924	42,4
5426	937,7	152 558 294	737 799	0,48	206,8	314 065	42,6

Tabla 3. Comparación del número de formas distintas, porcentaje que suponen sobre el total de formas, número de *hápax* y porcentaje que suponen sobre el total de formas distintas en diferentes segmentaciones del CREA. Fuente: www.rae.es. Elaboración propia. Los recuentos no toman en consideración signos de puntuación ni cifras y anulan la diferencia entre mayúsculas y minúsculas.

Los datos que aparecen en el cuadro fueron obtenidos mediante segmentaciones sucesivas del conjunto de ficheros que constituían la parte escrita del CREA en diciembre de 2002, abril de 2005 y abril de 2008. La primera parte del experimento se llevó a cabo con los 3212 textos que había en diciembre de 2002 (= CREA125),

tomando los veinticinco primeros ficheros (por orden alfabético), luego los 50 primeros (incluyendo, naturalmente, los veinticinco del primer paso), después los ciento cincuenta primeros, etc. hasta trabajar con el total de los 3212 existentes en diciembre de 2002. En la segunda fase, se añadió el recuento de los 4188 que había en abril de 2005. La tercera, cierre del CREA, en abril de 2008, incorporó ya los 5426 que componen la forma final del corpus. En los primeros saltos se buscaba duplicar el tamaño del segmento considerado en número de palabras. Los siguientes están condicionados por el volumen del corpus en cada momento.

La tabla muestra que el aumento del número de formas distintas tiene una curva de crecimiento considerablemente menor que la correspondiente al incremento del tamaño del corpus, que es lo esperable. De ahí que, por ejemplo, obtengamos una forma distinta cada 23,4 formas cuando el corpus tiene menos de dos millones, cada 121 cuando está alrededor de los cincuenta millones y, finalmente, cada 206 cuando rebasa ligeramente los ciento cincuenta millones. Lo llamativo, lo que justifica la construcción de corpus de estos tamaños es lo que muestra la columna de la derecha: el porcentaje de formas con frecuencia igual a uno con respecto al número de formas distintas se mantiene constante, alrededor del 40%, con independencia del tamaño del corpus. Si esa constante se mantiene, el CORPES debería tener unos 400 000 *hápax legómena*.

De lo anterior se deduce que es necesario seguir pensando en la construcción de grandes corpus textuales, puesto que se puede garantizar que cada nuevo bloque añadirá formas no documentadas en los anteriores. Pero la cuestión tiene más interés y relevancia. Las ya mencionadas leyes de Pareto y de Zipf indican que hay muchos elementos con una frecuencia muy baja, de modo que no podemos asegurar su presencia en un número muy significativo de casos si el corpus no tiene el tamaño suficiente para ello. Y, todavía más allá, solo los grandes corpus textuales pueden darnos seguridad relativa de que vamos a poder documentar también los usos o acepciones menos frecuentes de, por ejemplo, las palabras que tienen índices altos de aparición.

5. Veamos ahora el segundo de los factores mencionados. Gracias al incremento de la potencia de las computadoras y la reducción relativa de sus precios, el tamaño de los corpus ha crecido de forma rápida y continua. Renouf (2007: 28), que ha vivido directamente todo el proceso, diferencia, pensando siempre en el inglés, las siguientes épocas:

A partir de 1960: corpus de un millón de formas.
A partir de 1980: corpus de varios millones de formas.

A partir de 1990: *'Modern Diachronic' Corpus* (que se relaciona con la idea del 'monitor corpus' desarrollada por Sinclair.
A partir de 1998: la Web como corpus
A partir de 2005: computación distribuida (tecnología GRID).

En efecto, es generalmente aceptado que los corpus son instrumentos de gran utilidad –imprescindibles, en mi opinión– en el análisis lingüístico y también lo es que, a igualdad de los demás parámetros, son mejores y más útiles cuanto mayor sea su tamaño. Dada la ingente cantidad de textos existentes en la parte pública de Internet, ¿tiene sentido mantener las más que considerables inversiones necesarias para mantener los corpus existentes, ampliarlos y, en su caso, crear otros nuevos? En los últimos años se ha producido un movimiento muy intenso, conocido como *Web as Corpus*, que se mueve precisamente en la línea de utilizar todo lo existente en la red como un enorme corpus que está a disposición de quien quiera y sepa utilizarlo.[16]

A mi modo de ver, no son opciones excluyentes. Es evidente que cualquier persona, simplemente con acceso a Internet, puede conectarse a un buscador comercial y obtener en décimas de segundo información sobre los (miles o millones de) documentos en los que está contenida la palabra o expresión en la que está interesada. Dado el enorme volumen de datos y la continua adición y modificación de documentos, es altamente probable que algunas de estas búsquedas proporcionen resultados más interesantes que los que podríamos obtener de un corpus en aquellos casos en los que está implicado, por ejemplo, un elemento de escasa frecuencia o que se da solo en textos de tipos restringidos a Internet. Es verdad también que algunos de los buscadores más conocidos permiten refinar las búsquedas para que los resultados proporcionados procedan únicamente de textos escritos en una lengua determinada, de un cierto país o de un dominio específico. Las ventajas derivadas de la gran cantidad de datos y la posibilidad de refinamientos en las búsquedas han llevado al desarrollo de algunos proyectos que, como *WebCorp* o *KwicFinder*, utilizan buscadores comerciales como

---

[16] Hasta 2005, Google daba en su pantalla de entrada el número de páginas contenidas en sus índices: 8000 millones era la última cifra facilitada, aunque algunas estimaciones externas daban cifras bastante más altas. En un comunicado oficial emitido a finales de julio de 2008 (http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html), se afirma que sus ingenieros han llegado a la estimación de que la red contiene un billón (10^{12}) de URLs únicas (es decir, después de haber eliminado las que son simples copias o espejos de otras). Aunque ese comunicado no da el número de las que Google indexa actualmente, hay cálculos independientes que lo estimaban en 24 mil millones en 2005 (http://www.tnl.net/blog/2005/09/27/google-has-24-billion-items--index-considers-msn-search-nearest-competitor/) y en unos 40 mil millones en 2008 (http://www.techcrunch.com/2008/07/25/googles-misleading-blog-post-on-the-size-of-the-web/.

herramientas de búsqueda, pero presentan luego los resultados obtenidos del modo más habitual para el análisis lingüístico, esto es, en forma de concordancias.[17]

Los inconvenientes fundamentales del uso de la red como un corpus proceden, en mi opinión, de tres orígenes distintos. En primer lugar, la dependencia de buscadores comerciales, instrumentos deslumbrantemente útiles, pero concebidos con propósitos muy alejados de los que subyacen a una consulta lingüística. Esta dependencia se manifiesta en rasgos como las dificultades en el uso de expresiones regulares en la formulación de la búsqueda o, más en general, la sujeción a la disponibilidad de ciertas utilidades, facilidades, etc. El primer problema podría ser superado, aunque con algunos problemas, por programas de intermediación, como los ya mencionados *Webcorp* o *KwicFinder*, pero el segundo es determinante y no puede ser evitado.[18]

En segundo lugar, los materiales que hay en Internet a disposición pública (solo una parte de de los existentes) son muy numerosos y, en cierto sentido, mucho más ricos que los que se pueden encontrar en un corpus, por muy grande que sea. En otro sentido, sin embargo, son más pobres, precisamente por el carácter público que hace que no figuren textos cuyos propietarios no permiten que sean descargados, pero que no tienen inconveniente en que se procesen para extraer de ellos la información que termina en una línea o un párrafo que ilustra el uso de una palabra o una cuestión gramatical. Todo ello se materializa en la imposibilidad de acceder si se usan solo los textos públicos de Internet a una amplísima gama de materiales del más alto interés para el análisis lingüístico.

Por fin, Internet es, por su propia naturaleza, un conjunto permanentemente cambiante, lo cual constituye una enorme ventaja, pero es también un grave inconveniente para el análisis lingüístico, ya que esta inestabilidad consustancial hace que los resultados obtenidos cambien continuamente y no sea posible reproducir adecuadamente las búsquedas previas, ni siquiera las propias.[19]

---

[17] Sobre esta opción 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 (2007) 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).

[18] Por citar un caso, los usuarios de KwicFinder se quedaron (nos quedamos) sin esa herramienta de trabajo durante algún tiempo por el hecho de que el buscador comercial en el que se basaba (Alta Vista) anuló ciertas funcionalidades que el programa necesitaba.

[19] Sinclair (2005: 15) lo expresa con toda claridad:

> 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

Superada ya la etapa en la que había que invertir mucho tiempo y esfuerzos en transferir a formato electrónico los textos publicados previamente en papel, los costes de un corpus derivan fundamentalmente del esfuerzo necesario para codificar los textos (y transcribirlos si se trata de materiales orales). Ese esfuerzo, sin embargo, se ve sobradamente compensado cuando, en corpus como CREA, CORDE o CORPES, es posible recuperar los casos del fenómeno que nos interesa por áreas temáticas, tipos de texto, países, ámbitos temporales, etc. Esta posibilidad es, por otra parte, la consecuencia natural del diseño de un corpus, que ha sido construido de una forma y no de otra precisamente en función de permitir ciertos análisis lingüísticos, buscando la adecuación con respecto a lo que se pretende representar y el equilibrio necesario en los textos que lo constituyen. Es claro que no se puede recuperar automáticamente la información que no haya sido introducida previamente. Una operación aparentemente tan sencilla como la obtención de las frecuencias general y normalizada de una expresión en dos países, tipos de texto o épocas distintas requiere un trabajo previo de diseño y codificación que solo se puede dar en un corpus que ha sido construido precisamente con el propósito de permitir y facilitar análisis de este tipo. Naturalmente, algo parecido, referido a los aspectos relevantes en cada caso, se puede decir con respecto a todo lo relacionado con la información lingüística añadida. Solo en un corpus anotado o en una base de datos derivada es posible obtener los casos de un verbo cualquiera seguido de una preposición determinada o de cualquier preposición o los de una cláusula con un cierto predicado que contienen en su interior una completiva en subjuntivo en función de complemento directo. Y con la posibilidad, además, de reducir las búsquedas por países, años, etc.

En definitiva, la existencia de Internet y de las múltiples posibilidades que presenta no hace inútil ni superflua la construcción de grandes corpus textuales. En una caracterización rápida, los corpus tienen más sentido y llegan a hacerse imprescindibles en la medida en que la información que hay que manejar requiere mayor finura y capacidad de discriminación en la recuperación de los datos. De otra parte, la utilización de Internet para obtener algunos (o la totalidad) de los textos que van a formar el corpus constituye una enorme ventaja que abarata los costos de construcción y proporciona mayor riqueza al corpus. Como es lógico, entre esos textos están también los producidos

---

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.

precisamente para ser consultados en Internet.

6. La posibilidad generalizada de trabajar con corpus lingüísticos de extensión media o grande –con todo lo que ello significa de comodidad y rapidez en la recuperación de los casos pertinentes y de amplitud en los datos abarcados– ha cambiado radicalmente el panorama de los estudios lingüísticos. Ciertamente, el cambio se ha debido a la actuación simultánea de varias líneas de aproximación al estudio de los fenómenos lingüísticos, pero, a mi modo de ver, la lingüística de corpus ha tenido un protagonismo especial en este proceso.

Ese protagonismo ha sido incluso más destacado en la lingüística española, que presenta hoy una situación radicalmente distinta de la que mostraba hace veinticinco o treinta años. Desde una perspectiva global, con la simplificación que ello supone, esta especial intensidad se debe a que la lingüística de corpus se mueve en una dirección que da prioridad a aspectos habitualmente desatendidos –o menos atendidos– en la lingüística tradicional española, como la lengua realmente utilizada, las variedades distintas de la estándar en una consideración no fragmentaria, la lengua oral, los aspectos cuantitativos a partir de estadísticas que tengan en consideración todos los casos o el discurso, por citar únicamente los más destacados.[20]

Siempre en términos generales, la lingüística basada en el análisis de corpus comenzó a desarrollarse entre nosotros con un cierto retraso con respecto a lo realizado en otras lenguas, especialmente el inglés.[21] Gracias a un esfuerzo colectivo en el que en

---

[20] Según Gries (2006: 4), la mayor parte de los trabajos sobre lingüística de corpus comparten los principios 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.

[21] Téngase en cuenta que la primera versión completa del BNC se cerró en 1994 y que en 1995 aparecieron varios diccionarios para aprendices de inglés basados en el contenido de corpus de cien millones de formas o mayores. Sirva esa indicación para resaltar la importancia y novedad de proyectos para el español como, entre otros, el *Corpus oral de referencia de la lengua española* (integrado en la parte oral del CREA) o *Admyte* (http://www.admyte.com), dirigidos ambos por Francisco Marcos Marín, la participación española en *NERC* (cf. Alvar Ezquera y Villena Ponsoda,1994) o el corpus *CUMBRE* (cf. Sánchez, Sarmiento, Cantos y Simón, 1995).

muchos casos se han integrado las realizaciones de muchas personas y equipos distintos, la situación ha mejorado considerablemente y, al menos en ciertos aspectos, es ahora mismo perfectamente comparable (o incluso mejor) que la existente para inglés, francés, alemán, portugués o italiano. Limitándome, por cuestiones de espacio y también de adecuación al propósito general de esta ponencia, a los corpus de tamaño medio o grande que están a disposición pública, tenemos el *Corpus del español*, construido por Mark Davies, con cien millones de formas, parcialmente lematizado (www.corpusdelespanol.org), el *Corpus de referencia del español actual* (CREA: www.rae.es), con algo más de ciento cincuenta millones de formas procedentes de textos producidos entre 1975 y 2004 y el *Corpus diacrónico del español* (CORDE: www.rae.es), con cerca de trescientos millones de formas procedentes de textos desde los orígenes de la lengua hasta 1974. Precisamente como muestra de esa integración de esfuerzos a que acabo de hacer referencia, debo señalar que tanto la parte oral del CREA como la escrita contienen en su interior, además de textos transcritos directamente para el corpus, textos producidos en otros proyectos, debidamente reconvertidos al sistema de codificación del CREA y generosamente cedidos todos ellos por quienes los diseñaron, construyeron, transcribieron y codificaron (cf. www.rae.es). La suma de CREA y CORDE pone a disposición de los investigadores un conjunto de más de cuatrocientos millones de formas del español de todos los tiempos y todos los países. Más importante que el volumen es el hecho de que la información buscada puede ser recuperada de forma selectiva por tipo de texto, soporte, país, autor, obra o cualquier combinación de dos o más de estos rasgos. Constituyen, tanto en conjunto como por separado, una herramienta con la que no podíamos ni siquiera soñar hace veinticinco o treinta años y que la Real Academia Española mantiene a disposición pública desde 1998.

Como saben muchos de ustedes, los proyectos CREA y CORDE están ya cerrados. Las versiones del CREA posteriores a la actual cambiarán únicamente para corregir errores –inevitables en un recurso de estas características– y las del CORDE se limitarán a añadir en los próximos meses un conjunto adicional de unos treinta millones de formas constituidas por textos nuevos pendientes de revisión y sustitutos de textos ya incorporados, pero en ediciones mejorables. Habrá, no obstante, algunas novedades de interés general. Está prevista la aparición en la página electrónica de la RAE de una versión anotada del CREA, disponible ya, en fase de pruebas, en la red interna de la Academia. Está terminado también, a falta únicamente de la aplicación informática de

recuperación de datos, un corpus oral complementario que consta de un millón de formas con la interesantísima característica de tener alineados sonido y transcripción (cf. Sánchez Sánchez, 2005)

Por otra parte, en el Congreso de Academias de la Lengua Española celebrado en Medellín (Colombia) en marzo de 2007, se aprobó, a iniciativa de la RAE, el proyecto de construcción de un nuevo corpus, el *Corpus del Español del siglo XXI* (CORPES), que contendrá un total de 300 millones de formas procedentes de textos producidos en todos los países hispánicos entre los años 2000 y 2011.[22] El CORPES es, en cierto modo, la continuación del CREA (con unos años de solapamiento entre ellos), pero difiere de él en varios aspectos importantes. En primer lugar, el tamaño, que pasa ahora a 25 millones de formas para cada uno de los años comprendidos en el período abarcado. Esas cifras se hacen posibles gracias a una notable simplificación de la codificación de los textos –que no renuncia a nada importante, pero prescinde de todo aquello que la experiencia de todos estos años ha revelado inútil o poco rentable– y también al trabajo conjunto de varios equipos, situados en diferentes países y coordinados por un equipo central, que tiene su sede en la RAE. Además, la distribución de los textos ha cambiado en muchos aspectos. El más visible de ellos es, sin duda, el cambio en la asignación de pesos a las áreas geográficas. La relación entre América y España es ahora 70 / 30. En línea con la técnica adoptada en proyectos anteriores, el CORPES podrá integrar textos procedentes de otros corpus, con las adaptaciones oportunas en cada caso. Destaca en ese sentido el acuerdo alcanzado con los responsables del PRESEEA, que cederán los textos que encajen en la delimitación temporal del CORPES para contribuir a construir su parte oral. A comienzos del próximo mes de octubre será posible ya la consulta abierta de los primeros textos de este nuevo e importante corpus del español contemporáneo.

Existen actualmente muchos otros corpus del español,[23] de uso general en su mayor parte, y son muchos los grupos que, en paralelo, han diseñado y producido las herramientas necesarias para proceder a su anotación automática a diferentes niveles (morfológico, sintáctico, semántico y pragmático). Con todo ello será posible conocer mejor las características del español en todos sus variedades y registros y desarrollar los recursos que, en la parte aplicada, confieran a esta lengua el lugar que le corresponde en

---

[22] Se trata, pues, de un proyecto de la Asociación de Academias de la Lengua Española, por lo que cada una de las Academias integrantes ha designado a una persona responsable de él entre sus miembros. El CORPES es parcialmente financiado por el Santander.

[23] Parodi (2007b: 6-10) proporciona una selección actualizada.

la sociedad del conocimiento.

Referencias bibliográficas

Alameda, José Ramón y Cuetos, F. (1995): *Diccionario de frecuencias de las unidades lingüísticas del castellano*, Univ. de Oviedo, 1995, 2 vols.

Berber Sardinha, Tony (2000): "Lingüística de corpus: histórico e problemática", *D.E.L.T.A.*, 16/2, 2000, 323-367). Versión en PDF descargada de www.scielo.br. [Comprobado en julio de 2008].

Caravedo, Rocío (1999): *Lingüística del Corpus. Cuestiones teórico-metodológicas aplicadas al español* ((= *Gramática española. Enseñanza e investigación* I.6), Univ. de Salamanca, 1999.

Chomsky, Noam A. (1962): "A transformational approach to syntax" (comunicación presentada en la *3rd Texas Conference on Problems of Linguistic Analysis in English*, Univ. of Texas, Austin, 1958), en Hill (1962: 124-158). Cito por su reedición en Fodor & Katz , 1964, 211-245.

Davies, Mark (2006): *A Frequency Dictionary of Spanish. Core Vocabulary for Learners*, Nueva York / Londres: Routledge, 2006.

Facchinetti, Roberta: *Corpus linguistics 25 years on,* Amsterdam / Nueva York: Rodopi, 2007.

Featherston, Sam (2005): "The Decathlon Model of Empirical Syntax", en Kepser & Reis (2005), 187-208

Featherston, Sam (2007): "Data in generative grammar: The stick and the carrot", en *Theoretical Linguistics*, 33/3, 2007, 269-318.

Fillmore, Charles J. (1992): " 'Corpus linguistics' or 'Computer-aided armchair linguistics' ", en Svartvik (1992), 35-60.

Fletcher, William H. (2007): "Concordancing the Web: promise and problems, tools and techniques" en Hundt, Nesselhauf & Biewer (2007) 25-45.

Fodor, Jerry A. and J. J. Katz (eds.), *The structure of language. Readings in the Philosophy of Language*, Englewood Cliffs: Prentice-Hall, 1964.

Francis, W. Nelson (1992) : "Language corpora B.C. ", en Svartvik (1992), 17-31.

Gries, Stefan Th. (2006): "Introduction" a Gries, Stefan Th. & Anatol Stefanowitsch (eds.): *Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis*, Berlin: Mouton, 2006, 1-17.

Hill, Archibald A. (ed.) (1962): *Proceedings of the 3rd Texas Conference on Problems of Linguistic Analysis in English, 1958*, Univ. of Texas, Austin, 1962.

Hockett, Charles F. (1948): "A note on 'structure' ", *IJAL*, XIV, 1948, 269-271.

Hundt, Marianne, N. Nessehaulf & C. Biewer (eds.) (2007a): *Corpus Linguistics and the Web*, Amsterdam / Nueva York: Rodopi, 2007.

Hundt, Marianne, N. Nessehaulf & C. Biewer (2007b): "Corpus Linguistics and the Web", en Hundt, Nessehaulf & Biewer, 2007a, 1-5.

Juilland, Alphonse & E. Chang-Rodríguez (1964). *Frequency Dictionary of Spanish Words*. La Haya: Mouton, 1964.

Karlsson, Fred (en prensa): "Early generative linguistics and empirical methodology", en Kytö, Merja & A. Lüdeling (eds.): *Handbook on Corpus Linguistics*, Berlin / Nueva York: Mouton de Gruyter. Versión en PDF descargada de www.ling.helsinki.fi/~fkarlsson/earlygen.pdf [comprobado el 4/08/2008].

Keniston, Hayward (1937a): *The syntax of Castilian prose. The sixteenth century*, Chicago: The Univ. of Chicago Press, 1937.

Keniston, Hayward (1937b): *Spanish syntax list: A statistical study of grammatical usage in contemporary Spanish prose on the basis of range and frequency*, Nueva York: H. Holt & Co., 1937.

Kennedy, Graeme (1998): *An Introduction to Corpus Linguistics*, Londres / Nueva

York: Longman, 1998.
Kepser, Stephan & M. Reis (eds.) (2005): *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*, Berlin / Nueva York: Mouton de Gruyter, 2005.
Kilgarrif, Adam & G. Grefenstette (2007): "Introduction to the Special Issue on the Web as Corpus", *Computational Linguistics*, 29/3, 2003, 333-347.
Kučera, Henry (1992): "The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids", en Svartvik (1992), 401-420.
Leech, Geoffrey (1991): "The State of the Art in Corpus Linguistics", en Aijmer, K. & B. Altenberg (eds.): *English Corpus Linguistics. Studies in Honour of Jaan Jan Svartvik*, Londres: Longman, 1991, 8-29.
Leech, Geoffrey (1992): "Corpora and theories of linguistic performance", en Svartvik(1992), 105-147.
Lindquist, Hans and Christian Mair: *Corpus Approaches to Grammaticalization in English*, Amsterdam: John Benjamins, 2004.
Lüdeling, Anke, S. Evert & M. Baroni (2007): "Usign web data for linguistic purposes", en Hundt, Nessehaulf & Biewer, 2007a, 7-24.
McEnery, Tony & Andrew Wilson (1996): *Corpus Linguistics*, Edimburgo: Edinburgh Univ. Press, 1996; 2001[2].
Mair, Christian (2004) : "Corpus linguistics and grammaticalisation theory. Statistics, frequencies and beyond"), en Lindquist & Mair (2004), 121-150.
Parodi, Giovanni (2007a) (ed.): *Working with Spanish Corpora*, Londres: Continuum, 2007.
Parodi, Giovanni (2007b): 'Introduction' a Parodi (2007a), 1-10.
Quirk, Randolf, Sydney Greenbaum & Jan Svartvik (1972): *A Grammar of Contemporary English*¸ Harlow: Longman, 1972.
Quirk, Randolf, Sydney Greenbaum, Geoffrey Leech & Jan Svartvik (1985): *A Comprehensive Grammar of the English Language*, Londres: Longman, 1985.
Quirk, Randolf (1992): "On corpus principles and design", en Svartvik (1992), 457-469.
Renouf, Antoinette (2007): "Corpus development 25 years on: from super-corpus to cyber-corpus", en Facchinetti (2007), 27-49.
Renouf, Antoinette, Andrew Kehoe & Jayeeta Barnerjee (2007): "WebCorp: an integrated system for web text search", en Hundt, Nesselfauf & Biewer (2007), 47-67.
Rojo, Guillermo (2003): "La frecuencia de los esquemas sintácticos clausales en español", en Moreno Fernández, Francisco, Francisco Gimeno Menéndez, José Antonio Samper, M.ª Luz Gutiérrez Araus, María Vaquero y César Hernández (coords.): *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales*, Arco/Libros: Madrid, 2003, vol. I, 413-424.
Samper, José Antonio, Clara E. Hernández Cabrera y Magnolia Troya Déniz (eds.) (1998): *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*, edición en CD preparada por José Antonio Samper Padilla, Clara Eugenia Hernández Cabrera y Magnolia Troya Déniz. Univ. de Las Palmas de Gran Canaria / Asociación de Lingüística y Filología de la América Latina, edición, 1998.
Sampson, Geoffrey R. (2007): "Grammar without grammaticality", en *Corpus Linguistics and Linguistic Theory*, 3/1, 2007, 1-32.
Sánchez, Aquilino, Ramón Sarmiento, Pascual Cantos y José Simón (1995): *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*, Madrid: SGEL, 1995.
Sánchez, Aquilino y Pascual Cantos (1997): "Predictability of Word Forms (types) and

Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the *Cumbre* Corpus: An 8-Millon-Word Corpus of Contemporary Spanish", *IJCL*, 2/2, 1997, 259-280.

Sánchez Sánchez, Mercedes (2005): "El corpus de referencia del español actual (CREA): el CREA oral", *Oralia*, 8, 2005, 37-56.

Sinclair, John (2005): "Corpus and Text. Basic Principles", en Wynne (2005), 1-16.

Stefanowitsch, Anatol (2005): "New York, Dayton (Ohio), and the Raw Frequency Fallacy", en *Corpus Linguistics and Linguistic Theory*, 1/2, 2005, 295-301.

Svartvik, Jan (2007): "Corpus linguistics 25+ years on", en Facchinetti (2007), 11-25.

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.

Wynne, Martin (2005): *Developing Linguistic Corpora. A Guide to Good Practice*, Oxford: Oxbow Books, 2005.

Direcciones electrónicas mencionadas

ADMYTE:
http://www.admyte.com

AGLE:
http://cvc.cervantes.es/obref/agle/

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

BNC (listas de frecuencias):
http://www.liv.ac.uk/~ms2928/homepage.html

CORGA:
http://www.cirp.es

Corpus del español:
http://www.corpusdelespanol.org

Corpus dinámico del castellano de Chile (listas de frecuencias):
http://www2.udec.cl/~ssadowsky/lifcach.html

CREA, CORDE y CORPES
http://www.rae.es

Google (número de páginas indexadas y tamaño de la red):
http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
http://www.tnl.net/blog/2005/09/27/google-has-24-billion-items--index-considers-msn-search-nearest-competitor/)
http://www.techcrunch.com/2008/07/25/googles-misleading-blog-post-on-the-size-of-the-web/.

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