

# Sobre la Lingüística basada en el análisis de corpus

Guillermo Rojo

Universidade de Santiago de Compostela

## 1. Introducción

Permítanme que comience mi ponencia en estas jornadas expresando públicamente mi agradecimiento y mi satisfacción por encontrarme hoy entre ustedes. El agradecimiento se dirige a los organizadores de la reunión por haber tenido la gentileza de invitarme y haberme asignado la función de trazar para todos los asistentes un panorama rápido del estado actual de la Lingüística de corpus entre nosotros. Sé perfectamente que son muchas las personas que han colaborado en la organización del evento y mi reconocimiento se dirige a todas ellas, pero quisiera destacar especialmente, y espero que nadie me lo tome a mal, a Miriam Urkia entre todas ellas.

La satisfacción procede del hecho de que hoy, a finales del año 2002, un lingüista como yo, integrado originalmente en la tradición estructuralista europea, resistente a los cantos de sirena del enfoque chomskiano, pero consciente de lo mucho válido que hay en él, pueda en esta ocasión volver la vista atrás y contemplar con satisfacción el panorama de lo que hemos hecho hasta ahora entre todos, del ritmo de trabajo que hemos seguido y de los logros que hemos alcanzado. Nada más lejos de mi ánimo y mi carácter que el absurdo triunfalismo de quienes consideran que todo está muy bien y que las cosas no podían haber ido mejor, pero creo, honradamente, que los lingüistas que nos hemos movido en este terreno hemos logrado llevar a cabo, en no mucho tiempo, una labor que ha situado las infraestructuras materiales para el estudio de nuestras lenguas en un nivel difícilmente imaginable hace diez años. Mi objetivo en los minutos siguientes es ofrecerles un balance de lo realizado y el análisis de algunos de los problemas que se nos plantean o nos plantean con mayor frecuencia en este trabajo.

## 2. La Lingüística de corpus: del Infierno al Paraíso

Aunque siempre es posible remontarse muy lejos en busca de antecedentes más o menos remotos, la historia de la Lingüística basada en el análisis de corpus es bastante corta si, como creo que hay que hacer, vinculamos su existencia a la de conjuntos más o menos amplios de textos en formato electrónico. El primer corpus de estas características es, como sabemos todos, el *Brown Corpus*, construido hace ahora solo cuarenta años, y constituido por 500 muestras de unas 2000 palabras cada una.

Los primeros tiempos son realmente difíciles. El trabajo con corpus electrónicos no solo resultaba claramente marginal en relación con la corriente principal de la lingüística estadounidense de la época —distribucionalismo en franca decadencia sustituido por la gramática generativo-transformacional, en alza—, sino que, a causa de algunos factores secundarios, recibió todo el peso de la oposición de Chomsky al empleo de corpus. En efecto, en un texto fuertemente crítico hacia los objetivos y hábitos científicos de los distribucionalistas, Chomsky juzgaba totalmente inadecuado considerar que el objetivo del trabajo realizado por el lingüista pudiera consistir en dar cuenta de las secuencias que integran un corpus. El objetivo debe ser la lengua, esto es, las secuencias posibles en la lengua, las reglas que hacen posibles esas secuencias, la gramática, el conocimiento lingüístico que existe en el hablante-oyente ideal. Y la forma adecuada de acceder a ese conocimiento es, por supuesto, la introspección. 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, rechaza Chomsky la utilidad de los aspectos estadísticos que, según dice "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). Esta visión explica, me parece, la anécdota correspondiente a la conversación entre Robert J. Lees, en aquel momento uno de los más decididos partidarios de la entonces novedosa gramática generativo-transformacional, y el recientemente fallecido 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.

Casi todos los estudios que tratan sobre Lingüística de corpus se refieren al error de Chomsky al emitir ese juicio y, en la mayor parte de los casos, se limitan a señalar que los corpus que hoy manejamos tienen un tamaño considerablemente mayor (cientos de veces mayor) que el *Brown Corpus* y que, en consecuencia, aquella objeción carece de relevancia porque el objeto al cual se aplicaba era distinto. Y en cuanto a la airada reacción de Lees, el comentario suele referirse a que, con toda claridad, Francis estaba en el buen camino y que Lees no fue capaz de verlo. Sin embargo, creo que en estas críticas hay elementos de diferentes tipos y que entenderlas adecuadamente nos permitirá comprender mejor la evolución seguida por la Lingüística de corpus y la situación que presenta en la actualidad.

La primera cuestión es, sin duda, el hecho de que Chomsky y sus seguidores no estaban pensando, al formular sus reservas, en lo que conocemos actualmente como Lingüística de corpus, ni siquiera en sus formulaciones iniciales. Tras 'corpus', 'análisis de corpus' y expresiones de este tipo, los chomskianos ven la forma de trabajar de los distribucionalistas. Carece de sentido, por tanto, aplicar automáticamente las críticas, tanto para aceptarlas como para rechazarlas, puesto que su objeto era una forma de trabajar bastante diferente de la que se lleva a cabo en el *Brown Corpus* y sus continuaciones. En este sentido, es totalmente adecuada la afirmación de Leech (1991: 8), que estima que no hay continuidad entre la noción de corpus entre los distribucionalistas y la que manejan ya los primeros integrantes de lo que luego se llamaría 'Lingüística de corpus'. Y no es irrelevante, me parece, recordar que estos desajustes entre el objeto inicial de las críticas chomskianas y las corrientes a las que fueron aplicadas se dieron también, incluso con mayor gravedad, cuando la gramática generativo-transformacional se difundió por Europa y todo lo que decían del distribucionalismo fue aplicado de modo irreflexivo a todos los estructuralismos europeos.

Un paso más allá, parece que todos estamos de acuerdo en que el objetivo de un lingüista no puede ser simplemente describir o explicar lo que hay en un corpus, por muy grande que sea. Es cierto que hoy trabajamos con conjuntos textuales que son cientos de veces mayores que los corpus que Chomsky tenía ante sus ojos, pero también lo es que estamos convencidos de que un corpus, por grande que sea, no puede contener todo lo que es posible en una lengua. Concebimos los corpus como muestras del estado de lengua al que pertenecen, de ahí que pongamos tanta atención a la composición de los corpus y nos fijemos tanto en su carácter representativo y equilibrado. También aquí está clara, me parece, la ruptura a que hace referencia Leech.

Me gustaría dejar claro, antes de abandonar este punto, que es importante enfocar de modo no ingenuo la cuestión del carácter infinito de las secuencias posibles en una lengua o en un estado de lengua determinado. Como ya señaló Gross hace más de 25 años, hacer estimaciones a base de considerar que son posibles secuencias como *El niño comió el bocadillo*, *La niña comió el bocadillo*, *El niño comió los bocadillos*, *La niña comió los bocadillos*, *Los niños comieron el bocadillo*, *Los niños comieron los bocadillos*, *El niño come el bocadillo*, etc. resulta de muy escaso interés. Las estructuras lingüísticas (en fonología, morfología y sintaxis) no son infinitas y, como veremos en un punto posterior, presentan características estadísticas realmente interesantes. No se trata, por tanto, de que aceptemos que los corpus que construimos o manejamos son limitados no solo por el hecho de que estemos seguros de que no contienen todas las combinaciones de elementos léxicos posibles, sino en un sentido mucho más fuerte, referido a los esquemas sintácticos, a las combinaciones morfológicas, a lo auténticamente constitutivo de una lengua, en definitiva.

Naturalmente, del acuerdo en que el objeto del trabajo lingüístico debe ser la totalidad de lo que es posible en una lengua (y no el conjunto de secuencias que la casualidad histórica ha incluido en un conjunto de mayor o menor tamaño) no se deduce forzosamente que la introspección sea el único procedimiento mediante el cual resulte posible acceder a los datos relevantes para la investigación y la comprensión de un determinado fenómeno lingüístico. La cuestión ha sido suficientemente debatida y hoy parece estar claro que, si bien los corpus no proporcionan todo lo que un lingüista necesita ni la caracterización estadística de los fenómenos es un elemento de consistencia universal por sí mismo, la pura introspección, aislada de los datos procedentes de los usos reales y debidamente documentados, está irremediabilmente abocada a discusiones estériles sobre secuencias marginales o incluso imposibles, dejando sin explicación en muchos casos lo que realmente ocurre en una lengua o un determinado estado de una lengua. La competencia lingüística que tienen los hablantes es algo infinitamente más complejo de lo que se incorporaba a los primeros modelos chomskianos, de modo que el análisis de las secuencias producidas en condiciones reales resulta un elemento imprescindible del trabajo de los lingüistas. Es necesario, pues, como sucede en todas las ciencias, pasar continuamente de la teoría a los datos —a los datos reales, objetivos, no solo a los que el lingüista desdoblado en hablante pretende incorporar—, en aplicación estricta y bien entendida del método hipotético-deductivo. En este sentido resulta significativa la evolución de lingüistas como Fillmore, que se confiesa un "lingüista de sillón" que se beneficia continuamente del análisis de los datos procedentes de los corpus, datos que no habría podido tener en cuenta de otro modo (cf. Fillmore, 1992: 35). Más cerca de nosotros, Ignacio Bosque ha experimentado un proceso similar.

Me gustaría aprovechar este momento para fijar las diferencias entre la metodología que se practica en el trabajo con corpus y los modos de hacer tradicionales en la orientación empirista en Lingüística. Creo que es importante fijarse en ello porque, con cierta frecuencia, se piensa que la oposición entre las aproximaciones racionalista y empirista no se dan en ciertos terrenos, como, por

ejemplo, la lingüística histórica o la dialectología, disciplinas en las que, por imposibilidad de explorar la competencia propia, siempre ha sido necesario moverse con datos obtenidos de secuencias reales. Eso es cierto, sin duda, pero la verdad es que existen diferencias muy fuertes entre ambas aproximaciones. La distinción básica estriba en las diferentes formas de recolectar los materiales, que produce luego importantes divergencias. El lingüista descriptivo tradicional revisa los textos que le interesan —en cantidades mayores o menores, según las circunstancias— y marca, anota o hace papeletas de los casos correspondientes del fenómeno o fenómenos que está estudiando. En buena parte de los proyectos, la recogida, selección y clasificación de los materiales son realizadas por las mismas personas que luego van a estudiar el fenómeno en cuestión, de modo que se cuenta siempre con las ventajas que proporciona la actuación continuada de expertos a lo largo de todo el proceso comprendido entre la recogida inicial de los materiales y el estudio final resultante. El lingüista que trabaja con corpus, en cambio, lanza una búsqueda sobre un conjunto de textos que habitualmente han sido seleccionados y procesados por otras personas y recibe, como resultado de una búsqueda 'ciega' llevada a cabo por la máquina, todos los casos —a veces en cantidades realmente aplastantes— que responden formalmente a lo que ha solicitado.

Dicho de otra forma, la aproximación tradicional sitúa, de forma casi inevitable, un filtro previo a las documentaciones, lo cual produce varios problemas. En primer lugar, puede haber errores como resultado de las distracciones, cansancio, etc. de las personas que hacen la selección inicial. En segundo término, es lógico que, en las fases preliminares de la investigación, se desconozcan los factores realmente relevantes del fenómeno en cuestión, con lo que no existe garantía de que la selección realizada sea la que realmente debería haberse hecho, la que permitiría estudiar todas las caras de ese fenómeno. 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 caras marginales o simplemente menos representativas del fenómeno en cuestión, con lo que la selección de materiales resultante puede dar —al menos desde el punto de vista cuantitativo— una visión bastante diferente de la que realmente muestra el comportamiento lingüístico real. Naturalmente, no todos los que participan en un proceso de recogida de materiales tienen los mismos conocimientos sobre la totalidad de los aspectos de los diferentes fenómenos implicados ni comparten las mismas ideas acerca de lo que es realmente relevante en cada caso. Como anécdota ilustrativa de las características de la aproximación tradicional puede servir la relatada por John Simpson, editor del *New Oxford English Dictionary*. Cuando dos de sus lectores más expertos se ocuparon, por error, del mismo libro, las coincidencias entre lo resaltado por ambos para que fuera papeletizado no superaron el 25% de los casos, lo cual significa, en una primera aproximación, que los diccionarios resultantes de la selección realizada por cada uno de ellos serían bastante diferentes (¿un 75%?) entre sí<sup>1</sup>.

Por supuesto, no hay nada en el procedimiento tradicional que impida hacer una recogida exhaustiva de los materiales disponibles, pero no ha sido esa la práctica habitual y, como es lógico, solo sería posible para ciertos fenómenos y con un conjunto reducido de ejemplos. Dado que, como ya he indicado, la selección realizada tiende, además, a sobrevalorar los casos menos habituales, la imagen resultante de la actuación del filtro puede resultar bastante distorsionada con respecto a la que en realidad presenta el fenómeno en cuestión.

La ventaja, pues, del modo habitual de trabajo con los corpus radica precisamente en que nos permite —con unos costes en tiempo razonablemente

<sup>1</sup> Tomo la anécdota de Quirk (1992: 467). Naturalmente, lo que insinúo en el texto tampoco responde plenamente a la realidad. Lo esperable es que, como en todos los procesos generales, las diferencias entre los comportamientos individuales tiendan a compensarse entre sí, con lo que el resultado final presentará un aspecto mucho más homogéneo.

bajos— tomar en consideración todos los datos pertinentes. Es la única forma de lograr una perspectiva realmente completa, de alcanzar la meta de la 'total accountability' a que se refieren, entre otros, Leech (1992, pág. 112) y Quirk (1992, pág. 467), esto es, a la posibilidad de dar cuenta del comportamiento que los elementos estudiados muestran en todos los casos registrados en el corpus. Es, me parece evidente, un cambio bastante fuerte con respecto a lo que supone el modo de operar tradicional tal como se practicaba habitualmente, es decir, sin perseguir la exhaustividad. Para decirlo de nuevo con palabras de Quirk, los manuales basados en el uso real (Jespersen, etc.) nos garantizan 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 propias fichas que contienen los casos seleccionados<sup>2</sup>.

Es necesario, por tanto, considerar que la Lingüística basada en el análisis de corpus, que ha sido presentada repetidamente como la forma que adopta la lingüística descriptiva a finales del siglo xx y comienzos del xxi, se opone tanto a la aproximación racionalista como a la aproximación empirista tradicional. Y, como se ha señalado también repetidamente —cf., por ejemplo, Halliday (1993: 1)—, no se opone, en cambio, a la lingüística teórica.

Para poner fin a esta parte, de carácter general, digamos que los juicios negativos de Chomsky, basados en el sesgo inevitable de los corpus y la irrelevancia de los factores estadísticos para el estudio de la gramática de una lengua, son claramente erróneos. Teniendo siempre en cuenta que sabemos que no es posible encontrarlo todo en un corpus, es evidente que los sesgos son evitables si la muestra —el corpus— está adecuadamente construida. En los últimos años se han desarrollado corpus para estudiar lenguas en todas sus variedades y registros. En los estudios correspondientes se muestra con toda claridad el modo en que los fenómenos presentan frecuencias y características muy diferentes que, por supuesto, están en correlación con los parámetros generales. La gramática variacionista, que ha cambiado incluso el concepto de regla lingüística, es buena prueba de ello. Finalmente, es claro también que la utilización de elementos estadísticos en gramática no consiste en saber que *Yo he nacido en Madrid* es más frecuente que *Yo he nacido en Santiago*. La consideración estadística de los fenómenos lingüísticos es algo mucho más profundo e interesante, como ha quedado de manifiesto en el trabajo realizado en los últimos años.

En conclusión, el anatema chomskiano, que sin duda ha ejercido efectos negativos sobre el desarrollo de la lingüística basada en el análisis de corpus, especialmente en sus primeros tiempos, estaba dirigido a una metodología claramente diferenciada de la que nosotros y nuestros predecesores hemos puesto en práctica y, por tanto, no es aplicable a nuestro trabajo o bien estaba mal enfocado, con lo que tampoco debe ser tenido en cuenta.

### 3. Los grandes corpus: características y problemas

Resulta innegable, me parece, que una buena parte de la fundamentación teórica de la Lingüística de corpus y de los resultados que ha obtenido en los últimos años se debe en buena parte a que la evolución de las computadoras nos ha permitido construir corpus de tamaños progresivamente mayores, de modo que hoy resulta perfectamente aceptable y no demasiado ambicioso pensar, con ciertos matices, en corpus de mil millones de formas, consultables en tiempo real o casi

<sup>2</sup> 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).

real. Me ocuparé, en el resto de la sesión, de las características y consecuencias más destacadas de este cambio de enorme importancia y de los problemas que plantea. La idea fundamental que guía este desarrollo consiste en que estar convencidos de la adecuación de la Lingüística basada en el análisis de corpus no puede cegarnos a la aceptación de que esta metodología presenta también dificultades y problemas, puesto que ese reconocimiento constituye la única forma correcta de superar los inconvenientes.

Algunos de los que se apuntan como inconvenientes de la utilización de corpus lingüísticos son, sin embargo, realmente seudoproblemas, como sucede con los elementos básicos de la crítica de orientación chomskiana examinados en la sección precedente.

También constituye un seudoproblema, en mi opinión, la idea, bastante extendida entre nosotros, según la cual el análisis de un corpus supone que, en las formas muy frecuentes, el estudioso se ve desbordado por la cantidad de casos que tiene que analizar. Utilizando el ejemplo de siempre, si la preposición *de* tiene en español una frecuencia situada en torno al 6%, el trabajo con un corpus de cien millones de formas producirá nada menos que seis millones de ejemplos, que debería examinar quien pretendiera, por ejemplo, redactar la entrada correspondiente a esta palabra para un diccionario. Naturalmente, no solo no se trata de tener que examinar todos y cada uno de los seis millones de casos, sino de actuar con la garantía —relativa, al menos— de que nuestro análisis no deja fuera nada verdaderamente importante. En realidad, la cuestión se invierte con facilidad si disponemos de un corpus debidamente anotado y usamos programas informáticos que permitan que la inevitable selección de casos con que hay que trabajar en formas de frecuencia muy alta resulte también representativa de todos los contextos sintácticos (así como semánticos si disponemos de esa posibilidad) en que aparece este elemento. No tendremos, pues, una selección decidida de forma puramente aleatoria, sino realizada en función de lo que creemos pertinente en cada caso. Al tiempo, el tener la posibilidad de disponer de todos los ejemplos de la preposición *de* en un corpus de cien millones de formas nos permite elaborar, por ejemplo, listas de todos los verbos documentados en el corpus que llevan un complemento introducido por esta preposición. Si para el estudio de la preposición es evidente que no necesitamos conocer *todos* los casos, también lo es que un estudio sintáctico de los verbos que se construyen con complemento preposicional requiere, en cambio, esa información. La supuesta imposibilidad de trabajar con grandes masas de datos es, pues, otro seudoproblema que podemos evitar con relativa facilidad si nuestro corpus está adecuadamente construido y anotado<sup>3</sup>.

Los problemas que se plantean realmente en el análisis de corpus proceden de la propia composición estadística de los textos lingüísticos, que muestran sistemáticamente una constitución bastante distinta de la que suponen tanto los hablantes como los lingüistas no especialmente interesados en cuestiones estadísticas. Por comenzar por algo llamativo, véase lo que muestra el cuadro 1 acerca de los porcentajes que suponían las formas más frecuentes de la parte escrita del *Corpus de Referencia del Español Actual* (CREA), construido por la Real Academia Española, cuando constaba de 117 millones de formas aproximadamente.

| Porcentajes acumulados de las formas más frecuentes del CREA 117 (644 841 formas distintas) |                        |
|---|------------------------|
| Las <i>n</i> formas más frecuentes  | suponen en conjunto el |
| n = 10  | 28,55 %                |

<sup>3</sup> La selección puramente aleatoria no es, sin duda, el mejor procedimiento, pero ha servido en la preparación de algunos magníficos diccionarios y puede servir todavía, en caso de que no haya otras posibilidades, a condición de que sea debidamente complementada con otras técnicas.

|         |         |
|---------|---------|
| n = 25  | 39,43 % |
| n = 50  | 43,87 % |
| n = 100 | 48,24 % |
| n = 150 | 51,09 % |

Cuadro 1

Fuente: Corpus de Referencia del Español Actual (Real Academia Española).  
Elaboración propia.

Esto es, con solo 150 formas<sup>4</sup> tenemos el 50% (una de cada dos palabras ortográficas) de un texto corriente en español. Por supuesto, de aquí no se deduce que conociendo solo esas 150 formas podamos entender el 50% de lo que dice un texto, puesto que la lista está constituida casi exclusivamente por artículos, preposiciones, conjunciones, pronombres, etc. y algunas, muy pocas, formas verbales (es, por ejemplo) y nominales.

Naturalmente, no se trata de un fenómeno exclusivo del español, como puede comprobarse en el cuadro número 2, que incorpora los datos de otros dos corpus. Como puede observarse, los porcentajes que suponen los diferentes bloques de formas más frecuentes que he tomado en consideración están siempre bastante próximos y, además, sus diferencias responden casi siempre a lo que nos hace esperar nuestro conocimiento de las características tanto morfológicas como ortográficas de las tres lenguas estudiadas. Aunque no se puede dar todavía demasiado valor a este dato porque el CORGA tenía aquí un tamaño muy inferior al de los otros dos, el porcentaje del gallego está siempre por debajo del de las otras dos lenguas. De otro lado, los porcentajes del español son superiores a los del inglés en los 150 primeros elementos. A partir de ese punto, los del inglés son siempre mayores que los del español.

| Las n formas más frecuentes | BNC_90<br>(377 384 formas<br>ortográficas distintas) | CREA_117<br>(644 841 formas<br>ortográficas distintas) | CORGA_12<br>(268 164 formas<br>ortográficas distintas) |
|-----------------------------|--|--|--|
|                             | suponen en conjunto el                               |  |  |
| 10                          | 23,48%   | 28,55%   | 22,88  |
| 25                          | 32,30%   | 39,43%   | 31,97  |
| 50                          | 39,86%   | 43,87%   | 37,72  |
| 100                         | 46,66%   | 48,24%   | 43,66  |
| 150                         | 50,26%   | 51,09%   | 47,32  |
| 500                         | 61,35%   | 60,21%   | 58,00  |
| 1 000                       | 68,77%   | 66,23%   | 64,30  |
| 5 000                       | 86,02%   | 81,14%   | 79,65  |
| 10 000                      | 91,48%   | 86,97%   | 85,68  |
| 15 000                      | 93,94%   | 89,95%   | 88,76  |
| 25 000                      | 96,35%   | 93,11%   | 92,13  |
| 50 000                      | 98,44%   | 96,33%   | 95,58  |

Cuadro 2

Porcentaje sobre el total 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>).

*Corpus de Referencia del Español Actual* (Real Academia Española, [www.rae.es](http://www.rae.es))

*Corpus de Referencia do Galego Actual* (CORGA) (*Centro Ramón Piñeiro para a investigación en Humanidades*, [www.cirp.es](http://www.cirp.es)).

Elaboración propia en los tres casos.

<sup>4</sup> Se trata aquí, por supuesto, de palabras ortográficas, de modo que *llegué, llegaré, llegaría, del* y *diciéndoselo* son todas ellas formas diferentes. Al tiempo, *sobre* cuenta como una única forma con independencia de que sea sustantivo, verbo o preposición.

Lo mismo que con las formas ortográficas sucede con los lemas, tanto en general como cuando son considerados desde perspectivas específicas. El primer prototipo del CREA anotado, constituido por un millón de formas, contenía un total de 23 391 lemas. Pues bien, los 69 lemas de frecuencia igual o superior al 0,1% suponían el 53,03% del total de este subcorpus. Como era de esperar, el número de lemas necesario para alcanzar el 50% de un texto corriente es menor que el número de formas requerido para lograr el mismo resultado, lo cual resulta perfectamente lógico: la lematización de los determinantes, pronombres y algunos otros elementos pertenecientes a clases cerradas, así como la de los verbos más frecuentes, explica perfectamente la llamativa reducción que podemos observar. Los datos obtenidos en el estudio de este primer prototipo del CREA anotado confirmaban los que mostraba el *Corpus del español mexicano contemporáneo* (CEMC), sobre el que se aplicó una lematización coincidente solo en parte con la lograda en el prototipo del CREA<sup>5</sup>. A pesar de las diferencias en el proceso, los 69 elementos de frecuencia más elevada del CEMC suponen el 50,18%<sup>6</sup> del total, cifra perfectamente congruente con la obtenida para el CREA. Puede afirmarse, pues, que, prescindiendo ahora de los problemas que supone el sistema de adscripción de formas a lemas, un subconjunto de unos 70 elementos da cuenta de aproximadamente el 50% de un texto corriente en español contemporáneo.

Podemos ahora confirmar la bondad de los datos procedentes de estos corpus de tamaño relativamente pequeño con los obtenidos del proceso de anotación automática del primer gran bloque del CREA, constituido por unos 40 millones de formas de textos escritos que producen 95 699 lemas una vez fundidos los nombres propios. También aquí, los setenta lemas que tienen frecuencia igual o superior al 0,1% suponen en conjunto el 50,79% del total.

Las muy diferentes características de la lengua hacen que los datos que derivan del corpus estadístico vasco (el *Egungo Euskararen Bilketa-lan Sistematikoa*, EEBS) sean bastante distintos. He podido disponer, de nuevo gracias a la amabilidad de Miriam Urkia, de la lista de los 200 lemas más abundantes y sus frecuencias respectivas. Sin contar nombres propios ni verbos auxiliares, los 200 primeros lemas alcanzan en conjunto únicamente el 37,36% del total. La impresión obtenida de las cifras se ve reforzada al analizar con más detalle los datos: frente a lo que sucede en las lenguas románicas que he podido examinar, donde artículos y preposiciones suponen porcentajes muy altos entre los lemas que ocupan los primeros puestos, en el corpus vasco aparecen siete verbos entre los veinte más frecuentes. No obstante, sin intención alguna de reducir la importancia de estas diferencias, creo que el factor de fondo es el mismo en todos los casos.

Algo muy parecido sucede si nos fijamos en una clase de palabras concreta. Los datos procedentes de la *Base de datos sintácticos del español actual* (BDS), que nuestro grupo de la USC ha puesto ya a disposición de todos los investigadores<sup>7</sup>, muestran que, sobre un conjunto de 3450 verbos documentados en un corpus de casi millón y medio de formas, los 32 más frecuentes (algo menos del 1% de los registrados) suponen un porcentaje conjunto ligeramente superior al 50%.

Esta clara característica de los textos, que he sintetizado en diversas ocasiones diciendo que están siempre formados por unas pocas formas que aparecen mucho y muchas formas que aparecen poco o muy poco, se da, como hemos podido observar, en todas las dimensiones: en todas las lenguas, con las formas ortográficas, con los lemas o en el interior de una clase de palabras determinada. En

<sup>5</sup> Debe tenerse en cuenta que el proceso de reducción de formas a lemas supone inevitablemente interpretación y organización de los datos en un nivel relativamente abstracto, de modo que los resultados oscilan en función de las decisiones adoptadas. Por ejemplo, las formas *yo*, *me* y *mi* son lematizadas a un único elemento o a varios en diferentes proyectos. Discrepancias similares aparecen en la consideración de formas compuestas, perífrasis verbales, etc.

<sup>6</sup> Cf. Ham (1979: 54-55). Según estas mismas tablas, para alcanzar el 53,03% del total del corpus hacen falta los 94 primeros elementos del CEMC.

<sup>7</sup> Puede consultarse en <http://www.bds.usc.es>.



términos generales, esta característica estadística de los conjuntos textuales se traduce en el hecho, conocido y sufrido por cualquiera que se haya enfrentado con un proyecto de este tipo, de que lo que podemos llamar la 'rentabilidad' de un corpus —esto es, el porcentaje de aparición de formas nuevas— disminuye de forma realmente drástica a medida que el tamaño del corpus aumenta. El cuadro 3 muestra, en su parte izquierda, lo que se puede observar haciendo cortes en el CREA, aumentando el tamaño a partir del factor puramente aleatorio del orden alfabético de los nombres de los ficheros que contienen los textos. Como se puede apreciar con facilidad, con un corpus formado por algo más de 1 600 000 palabras encontramos una forma ortográfica distinta (*type*) cada 23,4 formas ortográficas del texto (*tokens*). Es una relación interesante, pero que no se puede aplicar sin más para proyectar lo que podemos obtener en caso de continuar añadiendo textos. Si duplicamos el tamaño del corpus, nuestro inventario de formas experimenta únicamente un aumento de un 41%, lo cual significa que ahora encontramos una forma distinta cada 33. Lógicamente, esa tendencia se mantiene. Con cincuenta millones, los *types* suponen ya menos del 1% del total y con 117 millones —el tamaño de la parte escrita del CREA sin los textos misceláneos al cierre de su segunda etapa— están ya por debajo del 0,6%, esto es, que obtenemos una nueva forma aproximadamente cada 180 palabras ortográficas introducidas en el corpus. Como se ve, la 'rentabilidad' disminuye considerablemente a medida que el corpus aumenta de tamaño.

Los números que acabamos de examinar no constituyen, por supuesto, una sorpresa, aunque siempre resulta interesante el poder poner cantidades reales a lo que nuestra intuición nos indica<sup>8</sup>. Quiero decir que era un fenómeno perfectamente conocido antes de poder disponer de corpus de tamaño superior a los cien millones de formas. El estudio del vocabulario de cualquier texto muestra inmediatamente que el porcentaje de formas distintas disminuye con fuerza desde muy pronto. Un experimento rápido que he hecho con algunos textos españoles que tenía a mi disposición muestra que más allá de las primeras ocho mil formas, el porcentaje de formas nuevas que aparecen en un texto cae de forma muy llamativa con independencia del tipo de texto.

El cuadro 3, que muestra los datos obtenidos en el análisis de un corpus real, deja ver lo inadecuada que resultaría la predicción del número de formas distintas en un corpus de español de unos cien millones de formas a partir de lo que se encuentra en los primeros dos o tres millones. Suponer el mantenimiento constante de la relación entre el número total y el número de formas distintas tal como se establece en los primeros dos millones (alrededor de un 4%) nos llevaría a calcular un total de 2 500 000 formas diferentes en un corpus de cien millones, cantidad que resulta muy superior a las aproximadamente 600 000 que podemos estimar a partir de lo que figura en el cuadro 3. Sin embargo, esta importante reducción del número de formas distintas a medida que aumenta el tamaño del corpus resulta muy inferior a la que podría predecirse si se hiciera la proyección de la tasa de descenso comprobada en los primeros tramos del corpus. Dicho con otras palabras, si se hace la suposición de que la reducción en los porcentajes de entrada de formas nuevas que se observa de las primeras cien mil formas al primer millón se va a mantener desde un millón a cien millones de formas.

---

<sup>8</sup> Sin embargo, se dice que un conocido sociólogo afirma que cuando necesita datos aproximados se sirve de las estadísticas, mientras que cuando necesita datos exactos utiliza su intuición.

| Datos de la parte escrita del CREA (situación en diciembre de 2000) |        |             |            |         |             |         |               |
|---|--------|-------------|------------|---------|-------------|---------|---------------|
| Núm. ficheros   | MBytes | Núm. tokens | Types      |         |             | Hápax   |               |
|   |        |             | Núm. types | % types | 1 type cada | Número  | % sobre types |
| 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          |

Cuadro 3

Datos obtenidos del conjunto de ficheros que constituían el CREA en diciembre de 2000. El experimento se ha hecho tomando primero los veinticinco primeros ficheros (por orden alfabético), luego los 50 primeros (comprendiendo, naturalmente, los veinticinco del primer paso) después los ciento cincuenta primeros, etc. hasta trabajar con todos los que había en aquel momento. En cada salto se ha buscado duplicar el tamaño del segmento previo en número de palabras. No se han tomado en cuenta las cifras ni, por supuesto, los signos ortográficos. Téngase en cuenta que no incluye los algo más de setecientos ficheros correspondientes a textos de carácter misceláneo.

Algo así hizo John B. Carroll y, afortunadamente, se equivocó. En 1967, cuando los corpus reales eran todavía muy pequeños, afirmó que la relación entre *types* y *tokens* tendería a disminuir considerablemente a medida que aumentaba el tamaño del corpus. Por tanto, tal como lo presenta Kučera (1992: 407), para 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 of 100 million tokens". Ahora sabemos que, afortunadamente, los noventa millones de formas que constituyen la parte escrita del BNC contienen 377 384 formas distintas<sup>9</sup>, casi el doble de las calculadas por Carroll.

Ahora sabemos que Carroll estaba equivocado y, lo que es realmente importante, conocemos también la causa de su error, que queda de manifiesto en la parte derecha del cuadro 3, donde se muestra una faceta diferente de las peculiaridades que presenta la configuración estadística de los textos. Aunque pueda resultar sorprendente para el profano, es algo relativamente bien conocido para quienes se han ocupado alguna vez de estas cuestiones que el número de formas que aparecen solo una vez en un texto (los *hápax legómena*) es bastante alto para lo que cabría suponer inicialmente: puede situarse en torno al 50% de las formas diferentes (*types*) que integran un texto cualquiera. Mucho menos conocido, en cambio, es el hecho de que el porcentaje de *hápax* parece ser una constante que se mantiene con independencia del tamaño del corpus. Al menos, eso es lo que se deduce de los resultados que he obtenido realizando diversos cortes de carácter acumulativo en el CREA. No tengo todavía datos que permitan extender a otros corpus este fenómeno que se observa en el CREA, pero todo indica que se trata de un fenómeno general, poco difundido hasta el momento: el porcentaje de *hápax* en un corpus es un factor constante cuya importancia depende, entre otros aspectos, de las características morfológicas y ortográficas de la lengua en cuestión. En otras palabras, en inglés contemporáneo debería presentar un aire no muy distinto del que tiene en español, pero tendría que ser inferior en unos cuantos puntos.

El peso de las formas únicas, sin duda inesperado, es lo que explica el fracaso en la predicción de Carroll, que era demasiado simple. Más ajustado que el de Carroll parece el procedimiento propuesto por Sánchez y Cantos, según el cual el número de formas diferentes es igual a la raíz cuadrada del número total de formas multiplicada por una constante cuyo valor ha de ser extraído del análisis de

<sup>9</sup> De nuevo según los datos que figuran en la lista de formas obtenida por Mike Scott. Cf. Cuadro 2.

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. Si aplicamos la media de estos dos valores al conjunto de 117 millones de formas del CREA que he utilizado ya en varias ocasiones a lo largo de esta ponencia, resultan 582 219, cantidad no muy alejada de las 644 841 que contiene en realidad (sin cifras ni signos de puntuación, por supuesto). Es una desviación situada en torno al 10%, importante sin duda, pero que podría ser válida como valor mínimo esperable, al menos para hacer una predicción de grandes números. Si la proyectamos a un corpus de 500 millones, nos dice que encontraremos alrededor de 1 200 000 formas distintas.

#### 4. Los corpus como fuentes de datos

He hecho referencia anteriormente a la oposición entre la aproximación tradicional y la basada en el análisis de corpus al estudio de los fenómenos lingüísticos. Además de todas las discusiones que podamos llevar a cabo en un ámbito puramente teórico, disponemos ya de datos reales acerca de las diferencias en los resultados obtenidos, al menos en algunos campos.

A mi modo de ver, resultan de enorme importancia los datos obtenidos en el proceso de confección del *Corpus textual informatizat de la llengua catalana* que nos han facilitado sus constructores y que figuran resumidos y parcialmente reelaborados en los cuadros 4 y 5. Hay en ellos dos aspectos que debemos examinar por separado. En primer lugar, siempre teniendo en cuenta que estamos hablando ya de lemas —más bien de metalemas<sup>10</sup>—, debe observarse la distribución de los elementos que figuran en este corpus, que contiene unos 54 millones de formas procedentes de textos escritos a lo largo de unos ciento cuarenta años de historia del catalán. La división de los textos en un conjunto de textos literarios y otro de textos no literarios, produce el hecho, realmente llamativo, de que el total de los lemas que tienen representación en ambos subconjuntos (con independencia de la mayor o frecuencia) no alcanza ni siquiera el 40%.

| <i>Corpus textual informatizat de la llengua catalana</i> (CTILC) (54 millones de formas) |         |         |
|---|---------|---------|
| Lemas comunes a partes literaria y no literaria   | 56 123  | 37,76 % |
| Lemas exclusivos de la parte literaria  | 29 926  | 20,14 % |
| Lemas exclusivos de la parte no literaria   | 62 577  | 42,10 % |
| Número total de lemas en el CTILC   | 148 626 | 100,00  |

Cuadro 4  
Fuentes: Rafel (1996), Soler (1998a, 1998b).

El factor realmente relevante es, sin duda, lo que esta diferencia significa. Para decirlo rápidamente, la imagen del catalán que se puede obtener a través de los textos que forman el bloque literario es distinta de la que aparece si trabajamos con el otro subconjunto. No creo que se pueda pensar en atribuir estas diferencias exclusivamente a las características diferenciales de cada uno de esos dos registros, que sería la explicación fácil y relativamente consoladora. El fenómeno es mucho más grave, más de fondo, y puede resumirse en la idea de que la distribución de formas y lemas en un conjunto textual está fuertemente determinada por el número y variedad de los textos que lo componen.

<sup>10</sup> Utilizo este término para indicar que todas las variantes ortográficas y dialectales han sido subsumidas en un elemento único.

Si, por lo que podemos ver en este momento, lo anterior es inevitable y nos refuerza en el convencimiento de que los corpus solo pueden ser considerados como muestras en mayor o menor grado representativas de un estado de lengua, el cuadro 5 nos enfrenta con las diferencias en la visión de las lenguas que se puede obtener cuando se contrasta la metodología tradicional con la basada en el análisis de corpus. Algo menos del 33% de los lemas resultantes de la fusión del corpus y el inventario contenido en las obras lexicográficas de referencia para el catalán están en ambos conjuntos. El resto está únicamente en los diccionarios (16,83 %) o únicamente en el corpus (el 50,76%).

| <i>Corpus textual informatizat de la llengua catalana (CTILC)</i>    |         |         |
|--|---------|---------|
| Número total de lemas en diccionarios anteriores (Fabra + Enc. Cat.) | 87.992  |         |
| Lemas comunes a corpus y diccionarios                                | 57.916  | 32,41 % |
| Lemas del corpus que no aparecen en los diccionarios                 | 90.708  | 50,76 % |
| Lemas de diccs. que no están documentados en el corpus               | 30.080  | 16,83 % |
| Total de lemas registrados en fuentes textuales y lexicográficas     | 178.704 | 100,00  |

Cuadro 5

Fuentes: Rafel (1996), Soler (1998a, 1998b).

La desesperación que las cifras pueden producir se reduce si tenemos en cuenta algunos factores diferenciales que, de forma irremediable, pesan sobre las dos perspectivas. El primero de ellos está relacionado con la inclusión en los diccionarios de la parte del léxico más relacionada con los procedimientos gramaticales, especialmente la derivación. En efecto, pensando en lo que sucede en español, los lexicógrafos se plantean continuamente y dan diferentes soluciones a si los diccionarios deben contener todos los casos de adverbios en *-mente* o de palabras formadas con prefijos como *anti-*, *pre-*, etc. o con sufijos como *-ción*, etc. No es infrecuente, al menos en diccionarios no especiales, que la solución sea la no inclusión de aquellos de estos elementos cuyo significado sea claramente derivable de la suma de los correspondientes al primitivo y al afijo, con lo cual se ahorran páginas que no aportarían demasiada información. En el examen de lo que hay en un corpus parece claro que cualquiera de estos elementos debe recibir análisis y, por tanto, da lugar a un lema, con lo que sin duda se produce una considerable cantidad de entradas que no encuentran correspondencia en los diccionarios porque la configuración de su macroestructura los ha excluido deliberadamente.

Surgen también diferencias, sin duda numéricamente importantes, en un terreno un tanto más complicado de examinar. Lo que se entiende por *lema* en un diccionario, la unidad en la que se reúne la información referente a cada una de las palabras que contiene, es el resultado de una serie de convenciones que tienen que ver con la tradición lexicográfica en que se inscribe el diccionario y otras convenciones propias de cada obra, incluyendo aquellas que tienen que ver con el hecho de que han sido pensadas para ser publicadas en papel y leídas por seres humanos. Por poner un ejemplo que permita a todo el mundo saber a qué me refiero, piénsese en lo que sucede con las diferentes clases de palabras a que puede pertenecer un cierto elemento. La diferencia entre usos verbales y sustantivos no da lugar generalmente a entradas distintas en la lexicografía inglesa, pero sí lo hace en la española. Lógicamente, un lexicón computacional, en el que los elementos deben llevar asociada toda la información morfológica y sintáctica pertinente, tiene que estar organizado de otro modo, con lo que las discrepancias son inevitables y eso produce cierto 'ruido' en el reconocimiento de las identidades y las diferencias.

Lo que acabo de señalar y algunos otros factores que debo dejar a un lado permiten explicar una parte, probablemente cuantitativamente importante, de las diferencias que hemos detectado entre los inventarios léxicos obtenidos por los

procedimientos tradicionales y los que derivan del análisis de corpus de tamaño mediano o grande. Sin embargo, hay otro porcentaje que deriva directamente del hecho de que la documentación obtenida por el método tradicional es fragmentaria, depende de los conocimientos de las personas que hacen la recogida inicial y tienden a favorecer lo que se considera —siempre desde los conocimientos de quienes seleccionan— extraño, poco habitual. El análisis de lo que contiene un corpus lingüístico, por el contrario, proporciona enormes cantidades de materiales repetidos, que pueden aplastar al estudioso si no dispone de medios para seleccionar adecuadamente esa información, y tiene, en cambio, bastantes dificultades para documentar adecuadamente los fenómenos que tienen baja o muy baja frecuencia.

En realidad, es fácil ver que no podría ser de otra forma. La estructura estadística de los corpus supone sistemáticamente la existencia de unos pocos elementos que aparecen mucho y muchos elementos que aparecen poco, muy poco ... o incluso nada, al menos en corpus de menos de algunos cientos de millones de formas. Para terminar esta parte e ir enfocando ya el final de la comunicación, me gustaría mostrarles lo que he podido obtener de la distribución de frecuencias en componentes distintos del léxico. Resultará ilustrativo, espero, de lo que significa la distribución de las frecuencias en un corpus el cuadro 6, en el que he situado las frecuencias de los fonemas y archifonemas del español que obtuve hace algunos años haciendo la transcripción fonológica automática en un corpus textual formado por un total de casi 3 650 000 realizaciones de fonemas. El resultado, como se puede apreciar en el cuadro, es que con únicamente diez fonemas o archifonemas alcanzamos el 75% del total de las realizaciones fónicas contenidas en el corpus. El fonema menos frecuente, /j/ (dejo fuera los archifonemas) tiene una tasa de aparición equivalente al 0,19%, es decir, algo menos de una vez cada quinientas realizaciones de fonemas, más o menos una vez cada ciento diez palabras. No sería estadísticamente extraño encontrar un texto de mil palabras —natural, no construido con el propósito deliberado de evitarlo— que no contuviera ninguna realización de este fonema. Ciertamente, mil palabras son pocas, pero si se piensa en lo reducido del inventario de fonemas del español se verá que la proporción no es tan irrelevante como pudiera suponerse inicialmente.

|     | Porcentajes | %acumulado |
|-----|-------------|------------|
| /a/ | 13,46       | 13,46      |
| /e/ | 13,46       | 26,92      |
| /o/ | 9,55        | 36,47      |
| /s/ | 7,55        | 44,02      |
| /i/ | 7,51        | 51,53      |
| /l/ | 5,12        | 56,65      |
| /N/ | 5,10        | 61,75      |
| /d/ | 4,72        | 66,47      |
| /t/ | 4,31        | 70,78      |
| /k/ | 3,81        | 74,59      |
| /P/ | 3,66        | 78,25      |
| /u/ | 3,15        | 81,40      |
| /b/ | 2,65        | 84,05      |
| /p/ | 2,59        | 86,64      |
| /m/ | 2,56        | 89,20      |
| /n/ | 2,39        | 91,59      |
| /R/ | 2,11        | 93,70      |
| /Ø/ | 1,69        | 95,39      |
| /g/ | 0,87        | 96,26      |
| /x/ | 0,73        | 96,99      |
| /r/ | 0,73        | 97,72      |
| /f/ | 0,68        | 98,40      |
| /s/ | 0,38        | 98,78      |
| /c/ | 0,27        | 99,05      |
| /D/ | 0,25        | 99,30      |
| /G/ | 0,22        | 99,52      |
| /j/ | 0,21        | 99,73      |
| /ʃ/ | 0,19        | 99,92      |
| /B/ | 0,08        | 100,00     |

Cuadro 6  
Frecuencia de fonemas y archifonemas en los textos de ARTHUS  
Fuente: Rojo (1991)

Para terminar en un ámbito un tanto más abstracto y elaborado que los anteriores, véase el cuadro número 7, que muestra el resultado de analizar los esquemas sintácticos que aparecen en las casi 160 000 cláusulas simples analizadas en la BDS. Teniendo en cuenta que el concepto de esquema sintáctico que hemos considerado solo tiene en cuenta la voz (activa, media, pronominal) y los elementos argumentales (no hay, por tanto, complementos circunstanciales ni predicativos no regidos por el predicado) no será inútil señalar que los 3 550 verbos existentes en la BDS documentan en total 145 esquemas sintácticos diferentes. El número obtenido proporciona, me parece, un buen elemento de juicio para situar la discusión mencionada al principio acerca del carácter finito o infinito de las secuencias existentes en una lengua. De ellos, únicamente 113 alcanzan una frecuencia igual o superior a cinco casos y el cuadro 7 muestra todos los esquemas que tienen un porcentaje igual o superior al 1%. Son únicamente 15 y los cinco primeros suponen en conjunto el 68,59%, lo cual significa que siete de cada diez cláusulas presenta uno de estos esquemas. Estos quince suman el 92,98% del total de las cláusulas del corpus, de modo que los ciento treinta restantes solo suponen el 7%.

| Construcción      | Esquema funcional                                 | Frecuencia del esquema sintáctico | Porcentaje del esquema sobre el total de ejemplos | Número de verbos que registran el esquema | Porcentaje sobre el total de verbos |
|-------------------|---|-----------------------------------|---|---|-------------------------------------|
| activa            | Sujeto-Predicado-CDirecto                         | 64.067                            | 40,04   | 2419                                      | 70,16                               |
| activa            | Sujeto-Predicado                                  | 16.857                            | 10,53   | 1140                                      | 33,06                               |
| pronominal        | Sujeto-Predicado                                  | 10.050                            | 6,28  | 1369                                      | 39,70                               |
| <i>activa</i>     | Sujeto-Predicado-Predicativo de Sujeto            | 10.010                            | 6,26  | 66  | 1,91                                |
| activa            | Sujeto-Predicado-CDirecto-CIndirecto              | 8.855                             | 5,53  | 584                                       | 16,94                               |
| activa            | Sujeto-Predicado-CAverbial                        | 6.750                             | 4,22  | 198                                       | 5,74                                |
| activa            | Sujeto-Predicado-Suplemento                       | 4.983                             | 3,11  | 340                                       | 9,86                                |
| <i>pronominal</i> | Sujeto-Predicado-Suplemento                       | 4.777                             | 2,99  | 491                                       | 14,24                               |
| activa            | Sujeto-Predicado-CIndirecto                       | 4.345                             | 2,72  | 222                                       | 6,44                                |
| activa            | Sujeto-Predicado-CDirecto-Predicativo de CDirecto | 3.921                             | 2,45  | 98  | 2,84                                |
| activa            | Sujeto-Predicado-CDirecto-CAverbial               | 3.088                             | 1,93  | 204                                       | 5,92                                |
| pronominal        | Sujeto-Predicado-CAverbial                        | 2.925                             | 1,83  | 218                                       | 6,32                                |
| pronominal        | Sujeto-Predicado-Predicativo de Sujeto            | 2.815                             | 1,76  | 104                                       | 3,02                                |
| <i>pronominal</i> | Sujeto-Predicado-CDirecto                         | 2.770                             | 1,73  | 384                                       | 11,14                               |
| <i>activa</i>     | Sujeto-Predicado-CDirecto-Suplemento              | 1.808                             | 1,13  | 288                                       | 8,35                                |

Cuadro 7

Esquemas sintácticos con un porcentaje de aparición en la BDS superior al 1%. Datos provisionales ordenados según la frecuencia del esquema sintáctico. En cursiva, los esquemas en los que la discrepancia entre las dos perspectivas reflejadas en el cuadro parece más fuerte.

Fuente: Rojo (2003)

## 5. Final

Va siendo hora ya de poner punto final a esta exposición. Ahora, con datos reales procedentes de algunos de los corpus de tamaño medio construidos en los últimos años, podemos evaluar lo que hemos conseguido y lo que cabe esperar de esta línea de trabajo en los próximos años.

Creo haber despejado cualquier duda acerca de la necesidad de disponer de corpus textuales lo más grandes, variados, equilibrados y representativos que sea posible, debidamente codificados y anotados en los diferentes niveles de análisis lingüístico. Es cierto que la rentabilidad de los corpus, medida en la producción de formas distintas, disminuye fuertemente a medida que el corpus aumenta su tamaño, pero también lo es que todo indica que el porcentaje de formas únicas se mantiene con independencia del tamaño del corpus, lo cual significa que siempre habrá crecimiento significativo. Puesto que la mayor parte de los elementos lingüísticos —sea cual sea el componente con el que se trabaje— presenta frecuencias bajas o muy bajas, la construcción de grandes corpus de referencia es el único modo razonable de obtener la documentación real que los lingüistas y todos los que trabajan en las llamadas 'industrias de la lengua', incluyendo la enseñanza de idiomas, necesitan para documentar adecuadamente los fenómenos de los que deben ocuparse.

Existe, además, otro factor relacionado con las frecuencias y los tamaños al que acabo de hacer alusión. En todo lo anterior he practicado una fuerte simplificación al referirme a formas ortográficas e incluso a lemas. El estudio de las características que un elemento presenta en un estado de lengua determinado necesita las documentaciones necesarias para la captación adecuada de cada una de sus posibilidades. Las acepciones de un verbo, por ejemplo, o los esquemas sintácticos con los que puede aparecer no tienen la misma frecuencia. Conseguir el número de casos adecuado para estudiar las acepciones o los esquemas más

frecuentes exige corpus de tamaños muy superiores a los que se requerirían si pudiéramos quedarnos únicamente con los más habituales.

Está claro también, aunque apenas haya hecho referencia a ello por falta de tiempo, que hoy no es posible pensar en un corpus lingüístico como el simple resultado de transferir a formato electrónico lo que previamente ha sido publicado en papel. Es necesario codificar esos textos para que la recuperación de la información pueda llevarse a cabo adecuadamente y de acuerdo con las necesidades de los expertos y, por supuesto, añadirle, como mínimo, la que llamamos habitualmente 'información morfosintáctica'. Es la única forma en que un corpus resulta de interés para los gramáticos y es el punto de partida de todos los estratos y aplicaciones posteriores.

Esto que acabo de indicar suscita dos cuestiones adicionales a las que me voy a referir únicamente de pasada. En primer lugar, es evidente que el trabajo de construcción de corpus nos ha enseñado bastantes cosas sobre la estructura de las lenguas que antes ignorábamos, sobre todo a partir del punto en que comienza la aplicación de métodos computacionales en el análisis de los textos. En ese sentido, el papel que la construcción de corpus ha jugado en el desarrollo de herramientas y recursos lingüísticos es fundamental. El deseo de anotar textos de forma automática nos ha hecho desarrollar métodos de anotación, métodos de desambiguación, nos ha llevado a construir gramáticas formales capaces de atribuir automáticamente a una secuencia la estructura sintáctica que le corresponde, etc. Y todo ello vuelve sobre los corpus, construyendo así una especie de espiral paradójica en la que cada vez nos elevamos más y, no obstante, podemos contemplar los fenómenos lingüísticos con más detalle y finura.

Evidentemente, todo eso requiere dotaciones económicas importantes, lo mismo que sucede, en general, con todo lo que es el desarrollo de infraestructuras. Pero, como ocurre con las infraestructuras de cualquier tipo, su rentabilidad —no solo científica en este caso— está asegurada si están bien construidas. Está claro, además, que los medios de que disponemos en este momento nos permiten pensar en corpus que sigan el modelo que empezamos a llamar de capas concéntricas: un núcleo central, relativamente pequeño, muy dotado de información adicional, una segunda capa, de mayor tamaño, con menos información adicional, luego una tercera y así tantas capas como sea oportuno hasta llegar a la consideración de todo lo que hay en Internet en una cierta lengua como un gran corpus del cual es posible obtener información valiosa con respecto a ciertos fenómenos.



## Referencias bibliográficas

- Chomsky, Noam A. (1962): Comunicación presentada en la *3rd Texas Conference on Problems of Linguistic Analysis in English*, Univ. of Texas, Austin. Cito por su reedición en Fodor, Jerry and Katz (eds.), *The structure of language. Readings in the Philosophy of Language*, 211-245. Englewood Cliffs, Prentice-Hall, 1964.
- Fillmore, Charles J. (1992): " 'Corpus linguistics' or 'Computer-aided armchair linguistics' ", en Svartvik (1992), 35-60.
- Halliday, M. A. K. (1993): "Quantitative Studies and Probabilities in Grammar", en Hoey, Michael (ed.): *Data, Description, Discourse. Papers on the English Language in honour of John McH Sinclair*, Londres, Harper-Collins, 1993, 1- 23.
- Ham Chande, Roberto: "Del 1 al 100 en Lexicografía", en Lara, Luis Fernando, Roberto Ham Chande y M.<sup>a</sup> Isabel García Hidalgo: *Investigaciones lingüísticas en Lexicografía*, México, El Colegio de México, 1979, 41-82.
- Kučera, Henry (1992): "The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids", en Svartvik, Jan (ed.) (1992), 401-420.
- Leech, Geoffrey (1991): "The state of the art in corpus linguistics", en Aijmer, Karin & Bengt Altenberg (eds.): *English Corpus Linguistics. Studies in Honour of Jan Svartvik*, Londres, Longman, 1991, 8-29.
- Leech, Geoffrey (1992): "Corpora and theories of linguistics performance", en Svartvik(1992), 105-147.
- Quirk, Randolph (1992): "On corpus principles and design", en Svartvik (1992), 457-469.
- Rafel i Fontanals, Joaquim (1996): "Diccionarios y corpus textuales. Perspectivas para el catalán", en García, Constantino, Isabel González Fernández, Manuel González González (eds.), *Actas do Simposio de Lexicografía actual. Elaboración de diccionarios* (= anexo 3 de *Cadernos da Lingua*), Real Academia Galega, A Coruña, s/a (pero 1996), 157-196.
- Rojo, Guillermo (1991): "Frecuencia de fonemas en español actual", en *Homenaxe ó Profesor Constantino García*, Universidade de Santiago de Compostela, 1991, 451-467.
- Rojo, Guillermo (2003): "La frecuencia de los esquemas sintácticos clausales en español", en *Homenaje a Humberto López Morales*, vol. III, 385-396. En prensa.
- Soler i Bou, Joan (1998a): "Los corpus textuales en lengua catalana", ponencia en el curso *Proyectos actuales en procesamiento del lenguaje natural*, organizado por la Fundación Duques de Soria, Soria, 13 a 17 de julio de 1998.
- Soler i Bou, Joan(1998b): Comunicación personal (25/8/98).
- Svartvik, Jan (ed.) (1992): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (= Trends in Linguistics. Studies and Monographs, 65), Berlín, Mouton - de Gruyter, 1992.