ACTAS DEL VIII CONGRESO INTERNACIONAL DE HISTORIA DE LA LENGUA ESPAÑOLA

Separata

Meubook

Actas del VIII Congreso Internacional de Historia de la Lengua Española

Santiago de Compostela, 14-18 de septiembre de 2009

Editadas por

EMILIO MONTERO CARTELLE

Secretaria de edición

CARMEN MANZANO ROVIRA

Separata

Edita: Meubook
ISBN: 978-849940469-19 Obra completa
ISBN: 978-84-940469-2-6 Volumen I
D.L. C 1628-2012
Unidixital S.L.

EL PAPEL DE LOS CORPUS EN EL ESTUDIO DE LA HISTORIA DEL ESPAÑOL

GUILLERMO ROJO
Universidad de Santiago de Compostela

1. Dado que soy el único integrante de esta mesa redonda a quien no cabe considerar especialista en historia de la lengua y teniendo en cuenta además que el *Corpus diacrónico del español*, un recurso suficientemente bien conocido para cuantos asisten a este Congreso, no ha tenido modificaciones sustanciales en los últimos tiempos, he decidido ampliar mi intención inicial y adoptar un planteamiento más genérico como punto de partida. Por tanto, dedicaré la primera parte a la enumeración —rápida, como es preceptivo en una intervención de este tipo— de los aspectos fundamentales por los que, a mi modo de ver, se puede considerar que la lingüística de corpus ha cambiado de modo radical la forma de acometer el estudio de la evolución del español en todos sus aspectos y componentes y está empezando a modificar también nuestra idea de cómo se produjo esa evolución en muchos aspectos. En la segunda, trataré de mostrar la medida en que el *CORDE* responde a las finalidades con que trabajan habitualmente quienes lo utilizan.

2. Desde la publicación —hace ya casi cincuenta años— del libro de Kuhn (1962) acerca de la estructura de las revoluciones científicas, se ha hecho habitual emplear la organización conceptual resultante de las ideas iniciales de Kuhn, matizadas y corregidas luego por otros muchos autores y también por él mismo, para diseñar el marco general en el que concebimos la evolución de las ciencias. Un factor presente con mucha frecuencia es el vínculo del concepto de revolución, de cambio radical en la historia de las disciplinas científicas, con aspectos conceptuales: de la concepción geocéntrica a la heliocéntrica, de la física newtoniana a la einsteniana primero y a la cuántica después, del distribucionalismo a la gramática generativo-transformacional, para ir acercándonos a nuestro terreno. La idea general de las revoluciones científicas es, en efecto, la de un gran cambio conceptual cuyo efecto más visible es, en palabras de Dyson, "explicar cosas antiguas de nuevas maneras" (Dyson 1997: 50). Sin embargo, para seguir con la concepción de este autor, son mucho más frecuentes, aunque, sin duda, menos llamativas, las revoluciones que él denomina 'instrumentales' (*tool-driven revolutions*), esto es, las derivadas de un cambio en las herramientas, como, por ejemplo, el que se produjo a partir del momento en que Galileo apuntó al cielo con el predecesor del telescopio que él mismo había fabricado

y se encontró con un panorama mucho más rico y notablemente más complejo del que hasta entonces se podía divisar a simple vista. "El efecto de una revolución impulsada por herramientas —sigue diciendo Dyson— es descubrir cosas nuevas que tienen que ser explicadas" (*ibídem*), aunque no es forzoso hacerlo desde una construcción teórica diferente.

3. A mi modo de ver, la lingüística basada en el análisis de corpus (LC) constituye una auténtica revolución instrumental, un gran cambio gracias al cual cualquier lingüista o filólogo puede disponer hoy, con rapidez y comodidad, de un conjunto de datos con el que era imposible incluso soñar hace tan solo unos pocos años. Como consecuencia de ello, este modo de hacer lingüística se diferencia tanto de los planteamientos racionalistas, propios de la lingüística basada en la orientación chomskyana, como de los que estamos acostumbrados a ver en la lingüística descriptiva tradicional. Y, por otro lado, permite entender los problemas que han tenido tanto quienes han pretendido hacer un hueco a la LC entre las disciplinas lingüísticas como entre quienes la han presentado como una metodología diferente, opuesta a la funcional, la generativa, etc. La LC es, como consecuencia de esa revolución instrumental, un modo especial y distinto de estudiar y explicar los hechos lingüísticos.

En efecto, es frecuente que, al hacer la oposición entre las aproximaciones que por comodidad llamamos racionalista y empírica y llegar a la conexión con la evolución de las disciplinas y subdisciplinas lingüísticas, se diga que hay que tener en cuenta que el objeto propio de algunas de ellas impide o, cuando menos, dificulta considerablemente ver cómo se concreta esta oposición. Es, claro está, el caso de la historia de las lenguas o el estudio de épocas anteriores de los sistemas lingüísticos, por la sencilla razón de que los investigadores no pueden apelar a la introspección, no pueden actuar como el hablante-oyente ideal que lo sabe todo acerca de su lengua y la investigación ha de hacerse a partir de los datos que la casualidad histórica ha conservado y puesto a nuestra disposición.

Todo eso es cierto, pero no lo es menos que las diferencias con el modo tradicional de llevar a cabo los estudios diacrónicos son muy fuertes y, en línea con lo anterior, más importantes para lo que aquí nos interesa, puesto que la otra oposición no tiene alcance real.

4. En línea con lo que supone una revolución debida a las herramientas, a los instrumentos de análisis, las diferencias entre la lingüística de corpus y la lingüística descriptiva tradicional no están tanto en los grandes conceptos organizadores como en la reunión y análisis de los datos. El método tradicional practica habitualmente una técnica de recogida de casos que supone un filtro selectivo, de modo que se basa en pocos textos, selecciona ejemplos de forma que prima inevitablemente lo diferencial, lo llamativo, aquello que a juicio de quien los reúne supone algo interesante (es decir, que depende de su conocimiento del fenómeno en el momento mismo en que hace la selección, etc.). Ese procedimiento, que sin duda tiene ventajas, encierra también algunos inconvenientes, que podemos resumir en la referencia a pocos ejemplos, seleccionados

sobre criterios inseguros, procedentes de pocos textos, casi siempre de un mismo tipo (literarios, por ejemplo)[1].

VENTAJAS DEL MÉTODO TRADICIONAL:

- Pocos textos, habitualmente bien seleccionados para conseguir la representatividad necesaria con respecto al fenómeno estudiado.
- Pocos ejemplos, habitualmente bien seleccionados.
- Inclusión de los parámetros necesarios para el estudio del fenómeno en cuestión.
- Inclusión de toda la información adicional aportada por el conocimiento lingüístico (lematización, análisis gramatical o léxico, etc.).

INCONVENIENTES DEL MÉTODO TRADICIONAL:

- Pocos textos, en muchos casos procedentes de la misma tradición textual, seleccionados en ocasiones con criterios derivados de factores estéticos, literarios, etc.
- Pocos ejemplos, en muchos casos seleccionados con criterios parciales, lo cual distorsiona la visión de la historia y la representatividad estadística.

El trabajo con corpus tiene también, como es lógico, inconvenientes y ventajas.

INCONVENIENTES DEL TRABAJO CON CORPUS:

- De forma dependiente del grado de anotación y la información asociada a los textos, la recuperación de la información deseada puede resultar complicada o excesivamente trabajosa.
- En el caso de fenómenos frecuentes o muy frecuentes, la gran cantidad de ejemplos documentados puede resultar excesiva.

1 En la línea de la que es, sin duda alguna, la mejor tradición de la lexicografía hispánica basada en el análisis de los textos, resultan muy significativas de las características de la aproximación tradicional las palabras de Félix Restrepo (1945: 430) al hacer la presentación de la continuación del *Diccionario de construcción y régimen*. Algunos artículos habían sido totalmente redactados ya por Cuervo. En otros

> no faltaba sino la redacción o encabezamiento de los distintos apartes en que esos ejemplos habían de figurar, redacción que pude hacer en unas horas de trabajo. Pero un estudio más detenido del acervo de papeletas legado por Cuervo a nuestra patria, estudio hecho en asocio del colaborador técnico del Instituto, D. Pedro Urbano González de la Calle, nos ha llevado a la convicción de que los ejemplos que en estas papeletas figuran son, en la mayoría de los casos, insuficientes para la redacción de los respectivos artículos.
>
> Hicimos, D. Pedro Urbano y yo, un considerable esfuerzo para clasificar y redactar la partícula *en*, para la cual dejó Cuervo cerca de 600 ejemplos, y pudimos comprobar que estos ejemplos son pocos si esta partícula, una de las más usadas en castellano, ha de guardar proporción con las ya publicadas *a* y *de*, en cada una de las cuales los ejemplos pasan de 1000. Quedaba pues la partícula *en* muy inferior a las anteriores y por consiguiente no tenemos ninguna garantía de que todos los usos de ella estén representados en los ejemplos coleccionados por Cuervo. En realidad, en la clasificación adoptada por nosotros hay usos importantes que apenas están autorizados por uno o por muy pocos ejemplos, cuando es práctica general de Cuervo apoyar con ejemplos abundantes cada uno de los usos de las palabras que estudia.

- En el caso de fenómenos o elementos poco frecuentes, la ausencia de determinados textos en la documentación incorporada puede dar lugar a vacíos que, en cambio, corresponden a datos relativamente bien conocidos por la investigación tradicional.

Ventajas del trabajo con corpus:

- Si el corpus está adecuadamente estructurado y codificado, la recuperación de la información deseada puede hacerse, incluso en los casos correspondientes a fenómenos muy frecuentes, segmentada por épocas, países, tipos de textos, autores, etc.
- En las mismas condiciones, obtener la frecuencia relativa, que es siempre la realmente interesante, proporciona la base necesaria para hacer estimaciones estadísticas adecuadas.
- El formato electrónico permite aspirar a analizar todos los casos registrados —o un porcentaje adecuado de ellos— sin que la selección, el filtro realizado en una fase anterior de la investigación sesgue nuestra visión de los datos.

5. En otras palabras, de modo perfectamente comprensible cuando se tienen en cuenta las circunstancias habituales, la investigación en la lingüística descriptiva tradicional suele desarrollarse mediante el análisis de todos los casos del fenómeno que se quiere estudiar en un conjunto reducido de textos o bien mediante una selección realizada con criterios variables de un conjunto más amplio de textos. En esta aproximación es inevitable que lo infrecuente, lo inesperado, lo llamativo, esté sobrerrepresentado. Es bien conocido que, en un proyecto modélico, representativo de la mejor línea de la lexicografía tradicional, Murray tuvo que pedir a sus colaboradores que no olvidasen enviar también ejemplos de palabras corrientes en sus significados habituales[2].

2 El apartado 7 de las *Directions to readers for the Dictionary* indica:

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

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

> If Readers will kindly remember that the Dictionary is to contain *all* English words ordinary and extraordinary, that it is to give, if possible, one quotation in each century for every sense or construction of every word, and that it is these quotations that we ask them to supply by their reading, they will at once see why we ask them to give us, not only all the *extraordinary* words or constructions in their books, but also as many *good*, *apt*, *pithy* quotations for ordinary words as their time and patience permit. The quotations for common words must come from *some* books; they ought to come from *all* books; and this can be realised only by each Reader sending some. The only difference is, that as quotations for *rare* words and rare uses of words are difficult to get, they ought to be seized at once, wherever they occur, and whether good of bad, for they may be the only ones; whereas quotations for common words in their common sense and construction need only be made when they are *good*, that is when the Reader can say, 'This is a capital quotation for, say, *heaven*, or *half*, or *hug*, or *handful*; it illustrates the meaning or use of the word; it is a suitable instance for the Dictionary'. Other things being equal, also, the shortest quotations, provided they are complete, are the best (Murray, 1879a).

Murray conoce bien el terreno, precisamente porque, al hacerse cargo de los materiales recogidos con anterioridad, ha tenido ocasión de observar el efecto de la falta de indicaciones en este punto:

> In my own opinion, the *Bases of Comparison* formerly issued by the Society were a mistake, and detrimental to the work which they were designed to serve. Their most obvious result, to one who exa-

El trabajo con corpus, en cambio, permite aspirar al análisis de todos los casos contenidos en él (que es siempre una muestra reducida de los textos posibles). Más bien, obliga a intentar un examen exhaustivo de la documentación existente, no condicionada por decisiones previas acerca de qué es relevante y qué no lo es[3]. No se trata, por tanto, de la superficial distinción entre trabajar sobre las fichas previamente seleccionadas o los ejemplos obtenidos, también mediante filtro, de un corpus, sino de una marcada diferencia en el modo de enfrentarse con los datos. Quirk lo expresa con total claridad cuando señala que es posible que los gramáticos (o los lexicógrafos) usen un corpus

> as a convenient source for "good examples" to put in their grammar. But that is not where the value or the challenge of a corpus will lie. If we ignore the value and evade the challenge of total accountability, our use of a corpus will be no advance on Jespersen's use of his voluminous *collections* of slips or Murray's use of those file boxes bursting with marked-up quotations for the *OED*. Such scholars certainly ensured that everything in their published volumes was firmly anchored in textual reality, but not that everything in their samples of textual reality was reflected in those published volumes (Quirk 1992: 467).

No se me escapa que la exhaustividad en el análisis de los casos de un determinado fenómeno es un objetivo al que se puede aspirar también con la metodología tradicional. No es, sin embargo, algo frecuente y, dada la naturaleza del proceso y los medios utilizables, resulta claro que se puede optar por examinar todos los casos presentes en un grupo reducido de textos o una (pequeña) parte de los casos que aparecen en un grupo amplio de textos, pero no es factible analizar todos los ejemplos existentes en muchos textos si se trabaja con fenómenos o elementos con frecuencias normales. Los corpus, construidos con propósitos generales, no reducidos a una determinada investigación, lo hacen posible.

6. Creo que de la contraposición trazada en los párrafos anteriores, que intento hacer de modo imparcial, aunque sin deseo de ocultar mis evidentes adscripciones metodológicas, se deducen interesantes consecuencias. En primer lugar, está claro que tanto el método tradicional como el basado en corpus son trabajos de segundo nivel, quiero decir, trabajos que se hacen posibles a partir de la

> mines the material, is, that while rare, curious, and odd words, are well represented, ordinary words are often most meagrely present; and the editor or his assistants have to search for precious hours for examples of common words, which readers passed by because they happened to find them put down in their 'Basis', as occurring in the Bible or in Burke. Thus of *Abusion*, we found in the slips about 50 instances: of *Abuse* not five, and we had to spend much time in tracing out the early occurrence of this word which readers had omitted to record. This is why we have asked every reader to give as many *common words* as he conveniently *can*: I had almost asked that rare and odd words should be omitted,—as apparently we have them all— and only common words noted henceforth (Murray, 1879b, 571-572; cursivas en el original).

3 Se trata, como he señalado en otras ocasiones, de lo que Leech (1992) y Quirk (1992), entre otros, han denominado la 'total accountability'. En palabras de Leech, los datos presentes en un corpus

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

consulta y el análisis de textos que han tenido que ser editados previamente. En efecto, quienes pretenden investigar sobre la forma en que presentan o evolucionan los fenómenos léxicos o gramaticales en etapas anteriores de las lenguas dependen de la posibilidad de analizar materiales que hayan sido editados con anterioridad, con todo lo que ello significa de sujeción a los textos publicados, a los criterios manejados por los editores, a sus hipótesis de reconstrucción, etc.[4]

Es este un terreno en el que yo no tengo casi nada que decir, y menos en un contexto como este, de modo que me limitaré a la indicación somera de algunos aspectos de interés a la hora de trabajar con corpus históricos. En primer lugar, aunque al hablar de estas cuestiones a todos se nos vengan a la cabeza textos de épocas anteriores, el problema de la fijación se da también, y no solo de forma anecdótica, en los autores contemporáneos. Por citar únicamente algunos casos bien conocidos, en los últimos tiempos Antonio Muñoz Molina, Juan Marsé y Gabriel García Márquez han publicado nuevas versiones, que presentan como definitivas, de, respectivamente, *El jinete polaco, Últimas tardes con Teresa* y *Cien años de soledad.*

En segundo lugar, se discute con mucha frecuencia acerca de las grafías, la conveniencia o inconveniencia del respeto al original, la modernización, la conservación de las diferencias gráficas que suponemos de importancia fonológica, etc. Sin embargo, el problema de la fijación de los textos es infinitamente más amplio y el centrar la atención en el problema de las grafías se explica únicamente por la dependencia que la recuperación de la información tiene con respecto a la forma gráfica.

La construcción de corpus es, como he indicado antes, un trabajo de segundo nivel en este sentido y depende totalmente de lo que se haya hecho previamente sobre los textos. No podría ser de otro modo. No tiene sentido pedir a quien codifica una versión electrónica de *La vida es sueño* para incorporar ese texto a un corpus mucho más amplio que solucione los múltiples problemas textuales (no ya de grafía) que los especialistas siguen discutiendo (cf. Iglesias Feijoo 2005). La exigencia en ese punto debe reducirse a la obligación de decir qué edición ha utilizado y a que haya la mayor coherencia posible entre los diversos materiales. Pero ese carácter de 'segundo nivel' se da también en muchas otras zonas del trabajo con corpus. Piénsese, por ejemplo, en todo lo relacionado con la adición de información lingüística, clasificación de palabras, categorías gramaticales, etc. Por supuesto, tampoco se puede pedir a quienes trabajan en lingüística computacional que resuelvan, en programas que asignan etiquetas automáticamente, problemas que los gramáticos no acaban de solucionar.

7. En la versión que se hizo pública en abril de 2005, que es la última de la que seré responsable, el *CORDE* consta de algo más de 260 millones de formas procedentes de textos de todas las épocas, todos los tipos y todos los países que configuran el mundo hispánico, desde los primeros textos hasta 1974, momento en el que enlaza con el *CREA*. Esa más que considerable masa de documentación está distribuida según los diferentes parámetros que se utilizan habitualmente

4 Naturalmente, todo esto se refiere a los corpus generales y no se puede aplicar a corpus especializados en cierto tipo de textos, que, como muchos de los mencionados en otras intervenciones que tuvieron lugar en esta misma sesión, afrontan la edición previa de los textos que van a incorporar al corpus.

en el diseño de corpus de este tipo (época, país, tipo de texto, área temática, etc.) con la intención de que los investigadores puedan obtener los datos necesarios para ver cómo se comporta un elemento o se produce un fenómeno con relación a las diferentes variables que quieren tomar en cuenta. El proyecto, llevado a cabo por la Real Academia Española y que contó con financiación parcial del Ministerio de Educación y Ciencia primero y del de Ciencia y Tecnología después, supuso la conversión a formato electrónico y la codificación de varios miles de textos[5], pero tuvo que enfrentarse también con el problema de que no había edición de muchos textos que seguramente son de gran interés para el estudio de la evolución del español o de que las ediciones existentes no ofrecían siempre las garantías exigibles. En algunos casos, esa necesidad se pudo remediar mediante encargos directos de transcripción y codificación de documentos inéditos, pero los costes de una acción como esta son prohibitivos más allá de acciones muy específicas. En cualquier caso, ese es un problema general, que afecta a quienes construyen corpus o trabajan sobre ellos en la misma medida que a todos los que investigan en la historia del español[6].

Aun a riesgo de parecer poco objetivo, debo señalar que la existencia de un corpus de estas dimensiones con textos que no pertenecen a los últimos años de la historia de la lengua, donde se dan todas las facilidades que proporciona la enorme cantidad de materiales digitalizados, constituye un hecho de especial relevancia para la investigación y un rasgo positivo para las instituciones que han financiado su construcción. Para situarlo adecuadamente, téngase en cuenta que en inglés, que es la lengua con mayor cantidad de recursos electrónicos, la que nos sirve a todos de referencia, quienes trabajan en la historia de la lengua deben manejarse con corpus como el de Helsinki (1,6 millones de formas procedentes de textos escritos entre c. 730 y 1710), *ARCHER* (*A Representative Corpus of Historical English Registers*, con 1,8 millones de formas de textos situados entre 1650 y 1999), *NEET* (*Network of Early Eighteenth-century English Texts*, que tiene unos 3 millones de formas del siglo XVIII), entre otros de tamaños similares. Es evidente que no son adecuados para el estudio de fenómenos de frecuencia no demasiado alta, lo cual ha llevado a algunos autores a utilizar los 2,4 millones de citas que contiene el CD-ROM de la segunda edición del *OED* como si se tratara de un corpus, con lo que se ha constituido una 'dictionary-based corpus linguistics', en feliz expresión de Merja Kytö (cf. Mair 2004: 123). No puedo detenerme aquí en esta interesante cuestión, a la que aludo únicamente para resaltar el hecho, realmente inusual entre entre nosotros, de que en este

5 La versión que está actualmente en la red (http://www.rae.es) contiene 4406 textos, de los cuales 530 son compuestos e integran un total de 30 078 documentos anidados. Hay, además, 1058 textos pendientes de la revisión final; una parte de ellos corresponden a ediciones distintas, más adecuadas, de obras ya incorporadas.

6 no solo se trata de la escasa atención prestada a los textos no literarios. En palabras de Iglesias Feijoo (2005: 25), [c]uando para nuestro oprobio, hay cientos de obras fundamentales de nuestra literatura faltas aún de ediciones críticas, debe concluirse que estamos trabajando a cada paso sobre textos provisionales, precarios, acaso mendaces, siempre inseguros. Ante un panorama tan inestable, muchas veces se llega a pensar si no sería oportuno hacer un llamamiento universal, una especie de convocatoria de estados generales del hispanismo con una única finalidad: la de proponer que toda una generación se dedique a cubrir ese vacío, a editar los textos mayores y menores, y solo después pasar de nuevo a emplearnos en las elegantes tareas de la lucubración ensayística.

punto la lingüística española goza de una situación considerablemente mejor que la inglesa[7].

8. Al hacer la caracterización general de los corpus, suele decirse que deben ser representativos de la lengua, variedad lingüística o estado de lengua a que se refieran y, al tiempo, que deben estar equilibrados, esto es, mostrar un cierto balance entre los diferentes tipos de textos integrados en él. A mi modo de ver, conviene centrar bien la importancia de la tan manida cuestión de la representatividad de los corpus, que es un concepto que presenta bastantes dificultades. Se entiende que una muestra debe ser representativa y se considera que lo es cuando reproduce en la escala adecuada las características generales de la población de la que ha sido extraída. No hace falta pensar mucho sobre la cuestión para caer en la cuenta de que lo que sucede en nuestro caso es, simplemente, que no conocemos las características de la población, de modo que difícilmente es posible determinar cuáles debería tener la muestra.

El ajuste en la visión es imprescindible y realmente importante, pero no da lugar a cambios conceptuales. En realidad, la exigencia de representatividad en sentido estadístico estricto es fuerte o muy fuerte en los corpus de pequeño tamaño o bien en aquellos que no permiten hacer recuperación selectiva de la información que contienen. En casos de este tipo, la representatividad es crucial porque constituye la única garantía de que los resultados obtenidos reflejan lo que sucede en la variedad estudiada y no producen una visión sesgada. En estos corpus, pues, es forzoso tratar de aproximarse a un ideal que sabemos inalcanzable y garantizar que el conjunto es congruente con la variedad a la que pretendemos extrapolar los resultados obtenidos en la muestra.

Distinta es la situación existente en los corpus de gran tamaño y en los que permiten la recuperación selectiva de información, esto es, la posibilidad de obtener los resultados que se dan en los textos de un cierto autor, cierta época, cierto país, cierto tipo, etc. o cualquier combinación de dos o más de esos factores. Lo que surge a partir de aquí es la opción de trabajar, no ya con la frecuencia general en el corpus, casi nunca interesante como dato aislado, sino con la frecuencia en los diferentes subcorpus virtuales que se puedan establecer, de forma que los resultados sean comparables entre sí. Es evidente que los datos necesarios para estudiar el modo en que las lenguas cambian tiene que utilizar este procedimiento. El círculo se cierra si, como sucede en el *CORDE* y en el *CREA*, existe la posibilidad cómoda de conocer el tamaño del subcorpus que hemos utilizado y, por tanto, obtener una frecuencia relativa comparable con la que resulta del análisis de otro subconjunto de tamaño distinto. Lo realmente

7 Es evidente que el poder trabajar con la totalidad del texto contenido en todas las citas (utilizadas en el *OED* en este caso) da a estos materiales una utilidad diferente de la que tienen originalmente, puesto que las fichas han sido obtenidas como ejemplos (especialmente ilustrativos) del uso de una determinada palabra. Al acceder a todo el texto de la ficha, se pueden estudiar fenómenos diferentes de los que estaban en la mente de quien la hizo y de quien decidió seleccionarla como ejemplo ilustrativo de un empleo determinado. En cierto modo, esta utilización, imposible sin medios informáticos, remedia parcialmente el inconveniente señalado por Quirk (cfr. *supra*). Además de Mair (2004), vid. Hoffmann (2004) para el estudio amplio de las implicaciones de usar una colección de citas como corpus y las características de las empleadas en el *OED*. Mark Davies elaboró una útil aplicación que permitía el acceso a este conjunto de citas (unos 37 millones de formas) al estilo habitual en la consulta de corpus, pero ha tenido que retirarla del acceso libre.

importante, pues, es que el corpus contenga textos de todas las épocas, tipos, estilos y países que estén implicados en la conformación de aquello que se quiere estudiar. Esto es, que esté equilibrado en el sentido más general de la expresión.

Un corpus es un agregado de textos y, como es obvio, contiene lo que está en los textos que lo componen. En los corpus de gran tamaño (cientos de millones de formas), las peculiaridades individuales se diluyen en la masa de datos y, por tanto, no crean las dificultades que supondrían en conjuntos de menor tamaño. No obstante, es innegable que esas peculiaridades permanecen y tienen un efecto que debemos conocer y explicar. Puede ilustrarse este extremo con dos aspectos relativamente bien conocidos. En primer lugar, la bibliografía contiene numerosas alusiones a casos de, por ejemplo, frecuencias inesperadamente altas de ciertas formas como consecuencia de la inclusión en el corpus de determinados textos. Así, en el *CREA*, por ejemplo, la palabra *trujamán* se registra 115 veces (esto es, 0,75 veces por millón). No es difícil ver que la mayor parte de las apariciones (88) se deben a la presencia de una obra de teatro colombiana titulada *Los diez días que estremecieron el mundo*, uno de cuyos personajes se llama precisamente *Trujamán*, de modo que cuentan todas las marcas de introducción de sus parlamentos, debidamente codificadas, y, claro está, todas las ocasiones en que los demás personajes se refieren a él. Es inevitable que cada texto introduzca sus características en el conjunto del corpus, pero lo importante es que, frente a lo que sucede en la aproximación tradicional, los factores individuales no den lugar a distorsiones que podrían deformar nuestra visión de los fenómenos que queremos estudiar.

De otra parte, es bien conocida la repugnancia que, según sus propias manifestaciones, siente Gabriel García Márquez hacia los adverbios en *-mente*[8]. El análisis de las seis obras de este autor contenidas en *CORDE* y *CREA* produce los resultados siguientes:

8 En *Vivir para contarla*, dice García Márquez haciendo referencia a sus comienzos en el periodismo:
La práctica terminó por convencerme de que los adverbios de modo terminados en *mente* son un vicio empobrecedor. Así que empecé a castigarlos donde me salían al paso, y cada vez me convencía más de que aquella obsesión me obligaba a encontrar formas más ricas y expresivas. Hace mucho tiempo que en mis libros no hay ninguno, salvo en alguna cita textual. No sé, por supuesto, si mis traductores han detectado y contraído también, por razones de su oficio, esa paranoia de estilo (p. 316).
Los cuatro casos que he podido registrar en *Vivir para contarla* (*tremendamente*, p. 398; *coincidencialmente*, p. 436; *solamente*, p. 436; *valerosamente*, p. 571) son, en efecto, citas textuales. La primera de ellas, de un texto del propio García Márquez en sus primeros años de trabajo periodístico. González Egido (2004: 68) no tiene en cuenta este hecho.

Cuadro 1: Distribución de los adverbios en *-mente* en diferentes obras de Gabriel García Márquez contenidas en *CREA* y *CORDE*. Fuente: http://www.rae.es.

Obra	Fecha de publicación	**Adverbios en *-mente* distintos**	**Total de casos**	**Total de formas**	**1 adverbio en *-mente* cada *n* palabras**
La hojarasca	1955	57	84	34 157	406,63
El coronel no tiene quien le escriba	1958	24	50	17 520	350,4
Cien años de soledad	1967	87	208	137 888	662,92
Crónica de una muerte anunciada	1981	0	0	27 960	---
El amor en los tiempos del cólera	1985	0	0	144 685	---
Vivir para contarla	2002	4	4	179 715	44 928

Elaboración propia.

No es difícil imaginar el terrible efecto que la utilización de estos materiales tendría para el estudio de la evolución del uso de adverbios en *-mente* en los últimos cincuenta años del español si fueran los únicos empleados o formaran parte de un conjunto reducido, en el que tuvieran un peso porcentual importante. Sin embargo, la exigua cantidad de los cuatro casos que García Márquez aporta al *CREA* apenas tiene fuerza para reducir las 4141 formas distintas terminadas en *-amente* que hay en el corpus y suman un total de 474 258 casos. Si consideramos provisionalmente válida esta cifra en el supuesto, bastante conservador, de que los terminados en *-amente* que no son adverbios se compensan con los que acaban en *-emente*, *-imente* o *-amente* y sí lo son, obtenemos una media de un adverbio en *-mente* cada 321,68 palabras, bastante próxima a la que se da en las dos primeras obras de García Márquez consideradas. Desde diferentes ángulos, queda claro en ambos casos que estas peculiaridades permanecen, pero pueden ser debidamente entendidas y explicadas.

9. Me gustaría cerrar esta intervención insistiendo en que el *CORDE* es un corpus diseñado con unas ciertas características y una determinada composición, no lo que se ha dado en llamar un corpus oportunista, construido sobre materiales digitalizados con anterioridad y con propósitos diferentes. Gracias a ello, es un corpus abierto a muy diferentes tipos de consulta, de modo que, además de servir a los trabajos de la RAE, resulta útil a todos los investigadores del español. En otros corpus, las búsquedas parciales se limitan a la posibilidad de obtener los datos correspondientes a diferentes siglos y poco más. Esa estructuración interna en unos pocos subconjuntos rígidos permite llevar a cabo ciertos procesos parciales y ‚congelar' los resultados para ponerlos a disposición de los investigadores.

Esa técnica no es la única y, probablemente, tampoco la mejor. La organización de los datos por siglos, relativamente fácil de poner en práctica, choca, sin embargo, con los muchos problemas que plantea la periodización de la historia del español, que ha sido objeto de numerosas discusiones y previsiblemente lo

seguirá siendo durante algún tiempo, de modo que un sistema de recuperación de la información que presente la distribución en siglos como única posibilidad de organización de los datos resultará excesivamente rígido y escasamente útil para la mayor parte de las búsquedas.

Pensando en las múltiples y cambiantes necesidades de quienes iban a consultar el corpus, el equipo encargado de construir el *CORDE* optó desde el principio por un diseño abierto, en el que, a base de invertir grandes esfuerzos en la estructuración de la información, la consiguiente codificación de los textos y la confección de los índices, es posible obtener los datos correspondientes a un lapso temporal cualquiera, de uno o más países, de un determinado autor, de un cierto tipo de textos, área o subárea temática, etc. o la combinación de dos o más de estos factores. Sigo creyendo que, para un proyecto de este tipo, ese es el enfoque más adecuado, puesto que es el que permite a los investigadores fijar exactamente las condiciones que deben reunir los textos sobre los que quieren hacer la recuperación.

Naturalmente, esa opción tiene un cierto precio, que puede resumirse en la imposibilidad de calcular previamente los resultados correspondientes a los textos seleccionados. Aquí es cada investigador quien debe hacer esa parte del trabajo, descargando los ejemplos de su interés y procesándolos con alguno de los muchos programas disponibles para ello.

Bibliografía

Dyson, Freeman (1997): *Imagined worlds*. Harvard University Press, 1997. Cito por la traducción española de Joandomènec Ros *Mundos del futuro*. Barcelona: Crítica.

González Egido, Luciano (2004): "De nombres, verbos, adjetivos y adverbios", *Pliegos de Yuste* 2, pp. 67-70.

Hoffmann, Sebastian (2004): "Using the *OED* quotations database as a corpus – a linguistic appraisal", ICAME 28, pp. 17-30.

Iglesias Feijoo, Luis (2005): "En el texto de Calderón. Teatro y crítica textual, a propósito de *La vida es sueño*", en Melchora Romanos, Florencia Calvo y Ximena González (eds.): *Estudios de teatro español y novohispano*. Buenos Aires: Universidad de Buenos Aires, pp. 23-55.

Leech, Geoffrey (1992): "Corpora and Theories of linguistic performance", en Svartvik, pp. 105-122.

Kuhn, Thomas S. (1962): *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press. Hay traducción española de Agustín Contín (1971): *La estructura de las revoluciones científicas*. México D.F.: Fondo de Cultura Económica.

Mair, Christian (2004): "Corpus linguistics and grammaticalisation theory. Statistics, frequencies and beyond", en Hans Lindquist y Christian Mair (eds.): *Corpus Approaches to Grammaticalization in English*, Amsterdam / Philadelphia: John Benjamins, pp. 121-150.

Murray, James Augustus Henry (1879a): "An Appeal to the English-Speaking and English-Reading Public to Read Books and Make Extracts for the

Philological Society's *New English Dictionary*". Utilizo el facsímil electrónico de la segunda edición de la *appeal* (24/6/1879) que se encuentra en la página del *OED*: www.*OED*.com/archive/appeal-1879-06/p1.html [descargado el 29/7/2009].

MURRAY, James Augustus Henry (1879b): "Eighth Annual Address of the President to the Philological Society, Delivered at the Anniversary Meeting", *Transactions of the Philological Society*, 1877-79, pp. 561-586.

QUIRK, Randolf (1992): "On corpus principles and design", en Svartvik (1992), pp. 457-469.

RESTREPO, Félix (1945): "La continuación del *Diccionario de construción y régimen de la lengua castellana*", *Thesaurus* 1/3, pp. 429-432.

SVARTVIK, Jan (ed.) (1992): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (= Trends in Linguistics. Studies and Monographs, 65). Berlín: Mouton - de Gruyter.