LINGÜÍSTICA
Vol. 24, diciembre 2010: 11-50
ISSN 1132-0214 impresa
ISSN 2079-312X en línea

SOBRE CODIFICACIÓN Y EXPLOTACIÓN DE CORPUS TEXTUALES: OTRA COMPARACIÓN DEL *CORPUS DEL ESPAÑOL* CON EL *CORDE* Y EL *CREA*

ON CORPUS ENCODING AND EXPLOITATION: ANOTHER COMPARISON OF THE *CORPUS DEL ESPAÑOL*, *CORDE* AND *CREA*

GUILLERMO ROJO
Universidade de Santiago de Compostela
España
guillermo.rojo@usc.es

En varios trabajos recientes, Mark Davies ha venido comparando el Corpus del español (CE), construido por él, con el CORDE y el CREA, diseñados y desarrollados en la Real Academia Española. En opinión de Davies, el CE tiene una 'arquitectura' muy superior a las que poseen CREA y CORDE, por lo que su utilización presenta ventajas considerables en diversos aspectos. El propósito de este trabajo es mostrar que la comparación que hace Davies se basa en la consideración de un conjunto muy reducido de factores y que el análisis de todas las características relevantes en este tipo de recursos lleva a la conclusión contraria: el CORDE y el CREA resultan mucho más útiles para la investigación lingüística que el CE.

Palabras clave: *Corpus del español*, CREA, CORDE, diseño, construcción y explotación de corpus

In recent works, Mark Davies compares the Corpus del Español (CE), which he himself designed, with the CORDE and the CREA, which were designed and developed in the Real Academia Española. According to Davies, CE shows a much better 'architecture' than CREA and CORDE, a fact that entails outstanding advantages when it comes to its use. This paper argues that Davies's statement arises from his comparison of a reduced number of features in the three corpora. Moreover, the paper intends to prove that a deep analysis of all the relevant characteristics in the three corpora leads to the opposite conclusion, namely that CORDE and CREA are actually much more useful for linguistic research than CE.

Recibido
09/09/10
Aceptado
19/10/10

Key words: Corpus del español, CREA, CORDE, corpus design, corpus compilation and exploitation

1. Introducción

Con independencia de que la consideremos una disciplina, una metodología o incluso una aproximación filosófica distinta a los hechos lingüísticos (Leech 1992: 106), a estas alturas parece ya innegable que la lingüística de corpus (LC) es la versión actual de la lingüística descriptiva, entendida en el sentido más amplio y más técnico posible. La investigación en nuestro campo ha dado un giro radical y nuestro conocimiento de lo que sucede realmente en las lenguas y de cuál ha sido su evolución ha experimentado un incremento muy notable. En términos generales, creo que la razón fundamental de ello radica en que la LC no solo se opone, como se dice habitualmente, a la lingüística de orientación racionalista, sino también a la lingüística descriptiva tradicional (*cf.* p.e., Rojo 2010), de modo que nuestras hipótesis acerca de los fenómenos lingüísticos pueden partir ahora de un conjunto de datos mucho mayor que el manejado tradicionalmente y, claro está, también disponemos de muchos más datos objetivos, basados en el uso real, para el contraste de nuestras conclusiones.

He mantenido en varias ocasiones que la lingüística española llegó a la LC con un notable retraso con respecto a otras, la inglesa sobre todo, pero consiguió en muy poco tiempo ponerse al nivel de lo que se estaba haciendo en otras tradiciones. En el caso de los corpus diacrónicos, incluso lo superó ampliamente, ya que el volumen y la información contenida en el *Corpus diacrónico del español* (CORDE), construido por la Real Academia Española, están muy por encima de los que se encuentran en sus correlatos ingleses como ARCHER o el *Helskinki Corpus* (*cf.* Rojo 2009). De la variedad y riqueza de los corpus de español existentes en la actualidad da buena idea el reciente trabajo de Briz y Albelda (2009).

Los llamados 'corpus de referencia', como el *British National Corpus* (BNC), el *Corpus de Referencia del Español Actual* (CREA) o, con la adición del componente diacrónico, el *Corpus Diacrónico del Español* (CORDE), son por su propia naturaleza proyectos complejos, que requieren una planificación cuidadosa y suponen fuertes inversiones en recursos humanos y materiales. Estos corpus se encuentran en la actualidad sometidos a las tensiones que producen la explotación directa de lo que se puede encontrar en la red (la línea

conocida como 'web as corpus') por un lado y los conjuntos pequeños, centrados en un tipo determinado de textos, incluso con varias versiones de la misma obra en muchos casos por otro. No puedo entrar aquí en esta cuestión (*cf.* Rojo 2010), pero me interesa aludir a ella para marcar la existencia de una clase de corpus, diferente de otras, que presenta unas determinadas líneas constructivas y tiene la finalidad de proporcionar los datos habitualmente requeridos en una gama amplia de investigaciones lingüísticas.

Aunque no ha sido siempre así, hoy es de aceptación general que lo que interesa de un corpus no es lo que puede mostrar globalmente de un fenómeno lingüístico determinado, sino las diferencias existentes entre lo que contiene ese corpus y lo que se puede obtener de otros o bien las que se dan entre distintos subcorpus (temporales, geográficos, tipológicos, etc.) del mismo corpus. Esta última es una característica clara de los corpus de referencia, que son construidos habitualmente de modo tal que los investigadores puedan hacer diferentes extracciones de datos, más o menos generales, más o menos específicas, del mismo fenómeno, comparar los resultados obtenidos y extraer las consecuencias correspondientes a las diferencias observadas.

Al tiempo, dado que los corpus son proyectos individuales, proyectados y desarrollados en unas ciertas condiciones y con unos objetivos determinados, es necesario que los investigadores conozcan sus características y también sus posibilidades de explotación, para decidir cuál(es) se ajusta(n) mejor a lo que necesitan o qué estrategias deben utilizar para obtener los datos que precisan. No es muy habitual que la comparación de las características de dos o más corpus (no de los datos que contienen, por supuesto) se convierta en el tema básico de un artículo, pero es perfectamente lógico que se hagan trabajos de este tipo, que sin duda constituyen una ayuda complementaria a la labor que, para su propio uso, tiene que realizar cualquier lingüista que pretenda utilizarlos. Es una tarea siempre difícil, pero que se hace especialmente delicada cuando la comparación entre el corpus A y el corpus B se asocia al hecho de que quien la hace tiene implicaciones personales en el diseño, construcción o desarrollo de alguno de ellos. Resulta casi imposible en ese caso evitar el sesgo que los intereses personales pueden introducir en los juicios, con el resultado de la pérdida de la objetividad deseable y exigible en un trabajo científico.

Mark Davies ha desarrollado en los últimos años un asombroso trabajo en la construcción de diversos corpus textuales para español, portugués e inglés, en los que aplica una línea de acceso a los datos muy brillante y efectiva, lo cual merece los parabienes y el agradecimiento de cuantos podemos estudiar los datos que obtenemos gracias a su trabajo. En algunos artículos recientes, Davies (2005, 2008, 2009) ha dedicado mucha atención a comparar las características generales y las posibilidades de explotación del *Corpus del español* (en adelante, CE) y también del *Corpus del portugués* (en adelante, CP) con las del CORDE y el CREA, construidos por la Real Academia Española para su propio uso y puestos a disposición de los investigadores desde 1998. Resumiendo mucho sus planteamientos, Davies considera que en la elaboración y análisis de un corpus hay que atender al conjunto, mayor o menor, de textos que lo componen, pero importa mucho –probablemente mucho más– "the corpus architecture and interface" (Davies 2009: 137). El CORDE y el CREA son, según Davies, individualmente y en conjunto, superiores al CE en el primer aspecto, pero muestran serias deficiencias en los otros dos, en los que CE y CP se muestran mucho más adecuados y útiles para la investigación lingüística.

A mi modo de ver, la línea de pensamiento y argumentación de Davies está excesivamente condicionada por la intención de mostrar la superioridad del CE y el CP, lo cual lleva a una presentación en la que con demasiada frecuencia quedan ocultas las características que no son favorables a su propósito o se desdibuja la realidad de las aplicaciones. No puedo ocultar que yo también estoy personalmente implicado en este asunto y que, en tanto que responsable del desarrollo de CREA y CORDE desde sus inicios respectivos hasta marzo de 2009, estoy interesado en mostrar precisamente lo contrario de lo que cree Davies, de modo que corro riesgos similares. No obstante, sin ocultar esa evidente motivación personal, me esforzaré en adoptar sistemáticamente la óptica de un lingüista que utiliza los datos contenidos en los corpus para hacer investigación lingüística profesional y, por tanto, necesita que sean fiables y puedan ser obtenidos de modo razonablemente cómodo. Trataré de hacer un análisis objetivo de lo que hay en cada caso y aportar siempre los elementos necesarios para que los lectores puedan extraer sus propias conclusiones.

2. LA VISIÓN DE DAVIES

En la más reciente comparación de los corpus que conozco, Davies (2009) establece la diferencia señalada en el apartado anterior entre el conjunto de textos y la arquitectura del corpus y presenta algunos ejemplos del comportamiento del CORDE, el CE y el CP en distintos componentes. En su estructuración general utiliza aspectos léxicos, morfológicos, sintácticos y semánticos y, en algunos de esos casos, emplea niveles de complejidad creciente. Hago un rápido repaso de lo que menciona en cada uno de esos bloques y luego las conclusiones a las que llega.

En el nivel léxico más bajo está la recuperación de los casos de una expresión (*braueza* en su ejemplo). Dejando a un lado los detalles, su conclusión es que, en este aspecto, CORDE y CE (o CP) se comportan de modo similar, aunque el CORDE tiene el problema de que solo devuelve datos "when the word occurs less than 1000 times in the corpus" (Davies 2009: 141) y el CE presenta las ventajas "of showing the frequency in each century and in showing the keyword in context for all words, regardless of frequency" (Davies 2009: 142).

En un nivel de mayor complejidad, "users often want to know how frequent a word was in different centuries or other historical periods. It is at this point that CORDE begins to exhibit some serious weaknesses" (Davies 2009: 142). La razón de ello está en que, siempre según Davies, la presentación de la frecuencia de los casos obtenidos se limita a dar la general (es decir, no normalizada) correspondiente a algunos de los años con mayor acumulación, lo cual es escasamente útil, ya que "a word or phrase may be more common in that year simply because there are more words for that year in the corpus" (Davies 2009:143).

El tercer nivel de las búsquedas léxicas consiste en la posibilidad de obtener "a list of words whose frequency matches certain criteria. For example, it might find all nouns that entered the language in the 1600s, or all words that are used at least five times as much in the 1200s that in the 1300s" (Davies 2009: 144). El CORDE no tiene esta posibilidad, que, en cambio, resulta muy cómoda y rápida en CE y CP.

Para las búsquedas morfológicas, ejemplifica con la expresión *des*m?ento* entre 1200 y 1400[1]. La respuesta del CORDE implica

[1] En realidad, por los datos de la respuesta, su consulta al CORDE corresponde al período situado entre 1200 y 1300 (797 casos en 81 documentos). El CE utiliza unas veces ex-

que, para obtener una idea general de lo que sucede, sería necesario revisar, una a una, las 797 formas, lo cual resulta muy pesado y, además, solo se consigue respuesta cuando los ejemplos no llegan a 1000. En el CE, en cambio, una consulta como la anterior devuelve instantáneamente un cuadro con la distribución de las frecuencias normalizadas de cada una de las formas que encajan en la expresión solicitada para cada uno de los períodos en que está estructurado el corpus (los siglos).

En un nivel de mayor complejidad, el CORDE no está lematizado, de modo que no hay forma de obtener respuesta a consultas sobre casos y frecuencias de *hacer* a lo largo del período que comprende. En un lema como este, con tan enorme variedad de formas y grafías distintas, la posibilidad de recuperar los datos forma a forma es inviable. El CE (y el CP), en cambio, devuelven inmediatamente las formas integradas en el lema en cuestión y su distribución por los tramos considerados, además de las listas de ejemplos en cuanto se solicitan. Dedica algunos párrafos a mostrar las dificultades que encierra el trabajo de lematización automática de un corpus, especialmente en el caso de que contenga textos de diferentes períodos históricos.

Para ilustrar las búsquedas sintácticas, utiliza expresiones del tipo *hacer* + infinitivo y *ser* + participio. Dado que el CORDE no está lematizado, son imposibles. El CE, en cambio, devuelve con rapidez todas las expresiones que responden a esa estructura y sus frecuencias en los diferentes períodos.

En cuanto a búsquedas relacionadas con la semántica, Davies examina en primer lugar la obtención de las 'colocaciones' de un término. Frente a la torpeza del CORDE / CREA en esta línea, que se limita, en la opción correspondiente, a dar la frecuencia general de las 'agrupaciones' más destacadas, de modo que habría que invertir enormes cantidades de tiempo en examinarlas y estudiarlas, el CE y el CP proporcionan instantáneamente la relación de términos candidatos a ser considerados 'colocaciones' de una palabra, con indicación de su frecuencia en las centurias seleccionadas.

---

presiones como 's13' para los años comprendidos entre 1200 y 1299 y otras '12' (en las referencias de los ejemplos) o '1200'. Además de que la correspondencia no es exacta según los criterios habituales de comienzo y fin de siglo, esa duplicidad crea cierta confusión tanto en las consultas como en las respuestas.

En un segundo nivel, el estudio de esos términos puede contribuir al conocimiento de la evolución semántica de una palabra, puesto que "if the words 'nearby' a given word change over time, it may be because the word itself has changed meaning (or is at least being used in a different way" (Davies 2009: 160). Como ejemplo, se refiere Davies a las diferencias en los adjetivos que siguen inmediatamente a *mujer* en textos del siglo XIX y los que hacen lo mismo en textos del XX, de modo que "in this case, the corpus data provides interesting insight into the changing view of women in these two centuries" (Davies 2009: 160).

Por fin, en lo que considera el nivel "most advanced" de las consultas sobre aspectos semánticos, se refiere Davies a la posibilidad de que haya consultas "by semantic fields, rather than just searching for words and phrases" (Davies 2009: 160). Para responder a esa necesidad, el CE y el CP contienen "powerful thesauruses" que permiten, por ejemplo, introducir '[=oscuro]' en la consulta y, a continuación, quien ha hecho la consulta puede "see the frequency of all synonyms over time (and in different genres from the 1900s)" (Davies 2009: 164). Además, en la última versión de la aplicación, es posible que los usuarios creen y conserven en el servidor sus propias listas de términos relacionados. Nada de esto es posible en el CORDE ni en el CREA.

En opinión de Davies, el CORDE "uses an older, now-outdated corpus architecture", que "was never designed for –and is completely inadequate for– most types of linguistic research" (Davies 2009: 165). De ahí derivan, directa o indirectamente, todas las deficiencias de estos corpus. Por ejemplo, el CORDE no puede "show whether the word of phrase was increasing or decreasing from one time period to another, or where it was the most frequent (which is quite useful information for philologists)" (Davies 2009: 165). El CE y el CP (y los demás corpus en los que ha trabajado Davies hasta la actualidad) tienen una arquitectura distinta, diseñada precisamente para responder a las necesidades de la investigación lingüística, de modo que, aunque el CE "was created by just one person in less than a year and a half, and with very limited funds", con el CE y el CP "researchers can examine an extremely wide range of linguistic shifts in ways that are not possible with any other historical corpus" (Davies 2009: 165).

Mutatis mutandis, lo mismo debería aplicarse a la comparación de estos dos corpus con el CREA.

3. La 'arquitectura' de los corpus

Para empezar por el final, resulta realmente impresionante lo que Davies ha conseguido, trabajando solo y casi siempre sin financiación especial, en tan poco tiempo. El CE y el CP son recursos de gran utilidad para todos, por lo que cuantos nos dedicamos a la investigación lingüística debemos reconocer y agradecer la inmensa labor que ha llevado a cabo. Sin embargo, ese obligado reconocimiento no puede basarse en una presentación como la que él hace de las virtudes de los corpus que ha construido y los defectos del CORDE (y, por extensión, del CREA). Es la suya una presentación que, para destacar las que considera ventajas del CE, olvida las características positivas del CORDE (y del CREA) y los rasgos negativos o carencias del CE. Tenerlas en cuenta de un modo objetivo proporcionará, espero, una consideración bastante diferente de la que ha venido difundiendo Davies, pero mucho más real.

La distinción entre los textos contenidos en un corpus y su arquitectura, que es el principio organizador de la presentación de Davies que he analizado en los párrafos anteriores, es superficial e insuficiente. Para entender bien lo que hay en un corpus y lo que se puede obtener de él hay que diferenciar entre los textos, la anotación no lingüística (la codificación), la anotación lingüística (morfológica, sintáctica, semántica o pragmática) y la aplicación de recuperación de datos. Tres de estos cuatro elementos diferentes (es decir, todo lo que no son los textos) están ocultos bajo la denominación de 'arquitectura' en la visión de Davies, de modo que no usaré ese término porque resulta demasiado confuso en este marco. Por 'anotación no lingüística' (codificación) entiendo aquí el conjunto de informaciones extratextuales (datos bibliográficos de cada texto, valores de los parámetros utilizados en la configuración del corpus, como el país, el tipo de texto o el área temática) y también de informaciones textuales (marcas de división interna del texto, citas, fragmentos en otras lenguas, desarrollo de abreviaturas, indicación de hablantes en textos orales y de personajes en obras de teatro, etc.).

Aunque luego volveré sobre algunos de estos elementos, para lo que sigue me interesa especialmente destacar la conveniencia de

que los textos estén codificados en muy diferentes aspectos y, como mínimo, en los que han intervenido en la configuración del corpus, que lógicamente han de ser aquellos que resultan relevantes para la comprensión de los fenómenos lingüísticos (y su evolución en el caso de los corpus históricos). Para una lengua como el español, parecen imprescindibles año (ojo: no siglo), país, área y subárea temática y tipo de texto). Esa codificación, sin duda costosa, es la única forma de lograr que luego, en la fase siguiente, la aplicación de consultas pueda localizar los textos (es decir, los casos contenidos en los documentos) de un cierto período, un determinado país o conjunto de países, etc. Naturalmente, la aplicación de consultas debe ser congruente con el esquema de codificación utilizado. De nada sirve un sistema de codificación muy detallado si la información que contiene no puede ser recuperada y mostrada por la aplicación de consultas. En sentido contrario, una aplicación de consultas rápida y brillante no sirve de mucho para el trabajo profesional si los textos no han sido codificados con los rasgos que utilizan habitualmente quienes hacen investigación lingüística.

En términos todavía muy generales, el CREA y el CORDE tienen un magnífico y detalladísimo sistema de codificación que va acompañado de una aplicación de consultas un tanto envejecida (data de 1998), que no es capaz de mostrar directamente todo lo que hay dentro de los corpus. El CE y el CP, en cambio, tienen una aplicación de consultas muy rápida y cómoda, pero la codificación de los textos es muy deficiente. La pregunta, a la que yo no sé responder, es si esa misma aplicación podría ser utilizada para un sistema de codificación que tuviera también las indicaciones de, por ejemplo, país, área temática, tipo de texto, autor, etc. o, simplemente, abriera la posibilidad de recuperar las frecuencias de una expresión en tramos temporales no prefijados y, por tanto, no precalculados.

Toda la argumentación de Davies se basa en una maniobra fundamental: olvidar la codificación y centrarse en las capacidades de una aplicación de consultas que trabaja con una codificación reducida a la indicación del siglo (y, en el XX, de tipo de texto). Tener en cuenta los demás factores no restará al CE la consideración positiva que merece, pero proporcionará una visión mucho más justa y equilibrada de las posibilidades de ese corpus y también de las que tienen los que Davies compara con él.

La parcialidad con que Davies encara la descripción de las características y posibilidades de los corpus que compara lo lleva a hacer una presentación que estimo poco adecuada de los fines con los que se pueden hacer consultas a los corpus. Por ejemplo, parece bastante peculiar la consideración de que la consulta abierta de un corpus tenga que producir directamente la lista de formas documentadas en un determinado siglo. Esa tarea parece más propia de un trabajo interno que produzca la lista de lemas (si está lematizado) o de formas o ambas en períodos cuya concreción debe ser establecida en función de las líneas generales de la lengua en cuestión, no con los rígidos e inapropiados convencionalismos de los siglos. Lo que sucede, en realidad, es que la configuración del CE asigna cada documento (y, por tanto, cada forma o lema de ese documento) a un siglo determinado, con lo que la recuperación de esa información resulta sencilla. Sin embargo, esa posibilidad, quizá interesante, pero no importante en el sistema de explotación de un corpus, no puede aparecer como un rasgo que dé lugar a una gran diferencia entre las posibilidades del CE y las del CORDE.

La cuestión es en realidad bastante más complicada. Ya he dicho que no creo que sea importante esa posibilidad para un sistema abierto de consultas. Pero aceptemos por un momento que pueda serlo. En ese caso, no parece lógico que las respuestas posibles queden reducidas a una estructuración rígida en centurias (1400-1499, 1800-1899, etc.).[2] Si la distribución temporal es importante (y, sin duda, lo es), lo lógico sería estructurar en períodos que se correspondieran con los habitualmente utilizados en los trabajos sobre historia del español. Todavía mejor: dado que esa estructuración resulta siempre discutible y los elementos evolucionan en épocas distintas y a ritmos diferentes, lo realmente útil y lo único adecuado a las cambiantes necesidades de la investigación es que la determinación de las fechas esté abierta a lo que precise quien hace la consulta y no que sea establecida de modo innegociable simplemente porque esa es la única forma de poder precalcular las frecuencias de cada tramo. Según Davies, la comparación entre lo que sucede en los distintos períodos históricos es imposible

[2] Como se ve, los períodos estructuradores del CE no coinciden con los siglos en la consideración habitual. No tiene mayor importancia, pero muestra un rasgo de interés en la forma de codificar el corpus: se recurre a la simple extracción de las dos primeras cifras del año asignado al texto, pero *cf. infra* para algunos desajustes.

en el CORDE, mientras que [d]ue to the architecture of the *Corpus del español* and the *Corpus do português*, where the corpus 'knows' the frequency of each word and phrase in each historical period, such comparisons are quite simple (Davies 2009: 145).

Está claro que esa afirmación solo es verdadera si forzamos el significado de la expresión 'períodos históricos' a significar única y exclusivamente 'siglos', puesto que eso es lo único que 'conoce' el CE, que, por otra parte, es incapaz de devolver las frecuencias de cualquier tramo temporal que no sea un siglo. 'Época alfonsí', 'Renacimiento', 'Siglo de Oro' y todos los demás 'historical periods' utilizados habitualmente por los investigadores están totalmente fuera del alcance del CE.

Por otro lado, no menos importante que la distribución por períodos (abiertos) parece la geográfica. Pero el CE no permite hacer consultas con respuestas distribuidas por países. Ni por tipos de texto. Ni por áreas temáticas, autores u obras. No todos estos rasgos tienen la misma importancia, por supuesto, pero es difícil evitar la sensación de que un sistema de consultas que no permita seleccionar los ejemplos de una palabra que aparecen en documentos publicados en un cierto período históricamente relevante (no por siglos) o en un país determinado o en textos de un cierto tipo resulta ser un sistema que no responde a las necesidades de la mayor parte de quienes investigan sobre el español. Por supuesto, el problema no está solo en la aplicación de consulta, sino también en el sistema de codificación utilizado o bien en la interacción de ambos.

Parece, pues, que Davies selecciona y pondera objetivos en función de lo que su aplicación de recuperación puede hacer con los escasos rasgos que contienen los textos incluidos en el CE, que han sido obtenidos en fuentes muy distintas y con muy diferentes grados de adecuación filológica. ¿Cuáles pueden ser las demandas reales de los investigadores a un corpus de referencia? Una aproximación no comprometida previamente con lo que aquí se discute (y, por tanto, neutral) puede ser la que Hoffmann (2008) propone. En su presentación de BNCWeb, la aplicación de consultas al *British National Corpus* (BNC),[3] plantea Hoffmann tres problemas que, en su opinión, solo

[3] El BNC es un caso paradigmático de corpus de referencia y, por tanto, sus características generales son similares a las del CREA y, con la adición del factor diacrónico, a las del CORDE.

pueden ser resueltos mediante el estudio de los datos facilitados por un corpus del estilo del BNC: el significado de la palabra *goalless*, la vigencia en el uso del modal *shall* y si son las mujeres o los hombres quienes hablan más de *cars*. El análisis de los datos del BNC muestra en todos los casos que la situación real es distinta de la que los hablantes ingleses creen saber acerca de las características actuales de su lengua. Resolver esas cuestiones, por tanto, solo es posible si tenemos la opción de consultar un corpus, obtener los resultados que corresponden a los subcorpus pertinentes en cada caso y, claro está, de analizar todos los casos relevantes. La de Hoffmann (2008) es, me parece, una aproximación realista, formulada con carácter independiente y que puede servir para establecer las características y posibilidades del CREA y el CORDE por un lado y el CE por otro. Trataré de llevarla a cabo de modo que aparezcan problemas del mismo tipo que los utilizados por Davies, aunque no exista un paralelismo estricto en la ordenación de las cuestiones planteadas.

4. Búsquedas de carácter léxico

Comencemos por una cuestión de carácter léxico, comparable, en el mundo hispánico, a la del significado de *goalless*, pero con algunos elementos adicionales. Puede ser, por ejemplo, acerca del significado actual de *enervar* y derivados. Como es bien sabido, el significado tradicional de este término lo aproxima a 'debilitar', pero hace ya bastante tiempo que ha perdido ese significado (o, cuando menos, lo presenta de forma muy minoritaria) y ha pasado a tener el de 'irritar, poner nervioso'.[4]

[4] No solo en español (*cf.* DRAE 2001, *s.v.*). El mismo cambio se ha dado en gallego (*cf.* DRAG, *s.v.*) y en catalán (*cf.* DDLC, *s.v.*). Resulta muy ilustrativa de las actitudes generales hacia procesos de este tipo (y, por tanto, de la pertinencia de estudiar los usos que muestran los textos) la nota que los traductores al español de la obra de Josep Pla *El Quadern Gris* (Gloria de Ros y Dionisio Ridruejo) añaden a la primera aparición de este verbo en el texto original:

> Pla usa aquí el verbo enervar con un significado opuesto al propio. Enervar quiere decir debilitar, quitar las fuerzas, deprimir. Pla quiere decir excitar o poner nervioso. No le corrijo, pero el lector queda advertido. Y sirva la advertencia para lo sucesivo, pues el empleo erróneo de la palabra es sistemático en sus escritos. (Nota de los traductores a la edición española de *El Quadern Gris*, de Josep Pla: *El cuaderno gris*, Barcelona, Destino, 1966: 48).

Los dos ejemplos de *enervar* en el sentido de 'irritar' que figuran actualmente (agosto de 2010) en el DDLC pertenecen a Pla (*cf.* http://dcc.iecat.net/ddlc/index.asp.).

Si pedimos *enerv** al CE reduciendo el ámbito de nuestro interés al s. XX, obtenemos inmediatamente una relación de las formas que responden a la expresión utilizada, con sus frecuencias totales correspondientes. Un total de 62, la más frecuente de las cuales es *enervante*, que aparece 22 veces. La opción GRÁFICO produce los gráficos de barras, frecuencias totales y frecuencias relativas (casos por millón de palabras) en todos los siglos (aunque solo se haya pedido el XX). Y para este siglo, además, se puede observar lo mismo según los diferentes tipos de texto establecidos: en textos académicos, 0,20 casos por millón, etc.

Pulsando en cualquiera de los elementos de la lista de formas, aparecen las líneas de concordancias acompañados de un número de orden, una indicación del siglo y el nombre (abreviado) del texto al cual pertenece cada una. Pulsando luego en cada una de esas líneas, se puede ver la fecha, el título del texto, su origen y un contexto más amplio. La fecha no siempre aparece (por ejemplo, no está en el primer caso de *enervante* que el sistema devuelve cuando la consulta se centra en el siglo XX) y, en muchos casos, pero no en todos, el origen del texto es un hiperenlace que lleva a los datos que existan en la página de la cual procede (siempre que funcione todavía, claro está). En este primer caso de *enervante*, el texto procede del habla culta de Bogotá, pero no indica fecha, así que hay que recurrir a documentación exterior para conocerla.[5] En términos generales, los datos del texto que puede ver quien hace la consulta son los que aparecen en la fuente original de la que Davies ha obtenido el texto, que puede ser, por ejemplo, una colección de relatos que, además de la historia, contiene únicamente el título del texto y el nombre de autor.

Por tanto, en resumen, lo que se obtiene del CE en una búsqueda de este tipo es la relación de formas que responden a la expresión introducida con sus frecuencias generales y normalizadas, diferenciadas por siglos, y, en los ejemplos del XX, por tipos de texto. Es posible obtener las líneas de concordancias con los datos –no siempre completos– del texto de cada una de esas formas en contextos reducidos y también con un contexto un poco más amplio. Y eso es todo. Los ejemplos correspondientes a cada una de las formas vinculadas a la expresión

[5] La hoja de cálculo en la que Davies ha volcado los datos de los textos contiene una línea para cada una de las entrevistas del habla culta de Bogotá incorporada al CE, pero ninguna de ellas lleva fecha.

usada en la consulta están ordenados de un modo preestablecido y no hay posibilidad de modificarlo.[6] La aplicación permite seleccionar las líneas de las concordancias por bloques y exportarlas a un fichero externo, pero solo contienen el número de referencia, el siglo, el título del texto y las líneas. Hay más texto y más detalles en otra pantalla, pero en ese caso la exportación ha de hacerse caso a caso.

Veamos ahora lo que sucede en el CREA,[7] que, como es bien sabido, contiene unos 160 millones de formas gráficas procedentes de textos de todos los países hispánicos y de los más diversos tipos publicados entre los años 1975 y 2004 (ambos inclusive). La respuesta, habitualmente rápida en consultas de este tipo, dice que hay 288 casos procedentes de 216 documentos y ofrece la posibilidad de consultar una página de estadísticas en la que se encuentra una presentación rápida de esos casos (con frecuencias generales y porcentajes) por años, países y las grandes áreas temáticas que estructuran el CREA.

En opinión de Davies, la tabla que refleja las frecuencias de los años más destacados es inútil. Refiriéndose al CORDE (pero la crítica general sería igualmente aplicable al CREA)

> [t]his table tells us the specific *years* in which the word or phrase is most common, but it is impossible to see the frequency by decade or by century. It does little good to show that the word was the most frequent in 1627, if in fact the word (*braueza* en este caso, G.R.) is much less common in the 1600s than in the 1200s or 1300s
>
> (Davies 2009: 143)

Sin duda tiene razón en ese punto, más acusado en el CORDE, puesto que, al abarcar un ámbito temporal mucho mayor que los veinticinco años previstos inicialmente para el CREA, los casos que puedan aparecer en un año determinado carecen habitualmente de interés. Pero Davies hace una presentación sesgada de lo que sucede en realidad y, además, no menciona aspectos relevantes para la evaluación de los dos corpus y, por tanto, para la información de sus lectores. El dato del año puede ser intrascendente, pero la distribución general de las formas por países y áreas temáticas (también las más destacadas) no

---

[6] El orden de aparición de los ejemplos está determinado en primer lugar por el siglo (comenzando por el más reciente) y luego por algún factor de ordenación que, al parecer, no tiene relación con las características propias del texto.

[7] Para los aspectos que vamos a tratar aquí, es irrelevante que la consulta se haga sobre el CREA o el CORDE.

lo es, puesto que en la mayor parte de las ocasiones (en todas aquellas en las que la distribución es sensible a estos factores) proporciona un utilísimo panorama general de cómo se distribuyen los casos en relación con esos parámetros. ¿Qué ofrece en estos aspectos el CE? Absolutamente nada. Y, por supuesto, no se trata solo de la distribución general de las frecuencias: el CE no permite hacer recuperación selectiva de la información ni por países ni por áreas temáticas ni por tipo de texto ni, claro está, por ninguna combinación de dos o más de estos factores. En realidad, incluso es discutible que se pueda decir que el CE permite la recuperación selectiva por tramos temporales. Los datos están agrupados por siglos y el sistema no permite salirse de esa estructura. La rigidez nunca es buena, pero la estructuración por siglos parece escasamente adecuada para entender la evolución del español en la mayor parte de los fenómenos.

A mi modo de ver, sin duda parcial, pero con auténtico deseo de examinar los corpus y las aplicaciones que los explotan, aquí se encuentra el núcleo de las diferencias entre ambos enfoques. El CE tiene una estructura codificadora que solo atiende a los siglos y, en el XX, a una cierta tipología de los textos. En general, pues, solo un parámetro y, además, sometido a una estructuración fija. Atender a un único factor[8] y hacerlo un modo predeterminado permite que los resultados puedan estar precalculados y, por tanto, facilitar las frecuencias generales y normalizadas de modo instantáneo. El CORDE y el CREA, en cambio, se basan en una riquísima codificación de los textos, incluso jerarquizada en algunos campos, como el correspondiente a las áreas temáticas (*cf. infra*). Es posible, por ejemplo, hacer la consulta *enerv** para los años comprendidos entre 1981 y 1987 (o cualesquiera otros) en el CREA y entre 1782 y 1859 (o cualesquiera otros) en el CORDE. Por esa razón, porque son corpus de estructuración muy rica y detallada, que permiten consultas tan abiertas como dicten las necesidades de los usuarios, no puede tener precalculados los resultados y por ello solo los puede dar cuando se hace la consulta correspondiente. ¿Podrían precalcularse los resultados de las formas ortográficas o de bigramas y trigramas por siglos? Naturalmente que

[8] Me refiero a la organización general del corpus, sin olvidar que una parte (la del siglo XX) tiene también la indicación del tipo de texto. De todas formas, la utilidad inicial de este segundo parámetro está muy condicionada por su aplicación global a un período tan amplio y diversificado internamente como el siglo XX. *Cf. infra*, apdo. 5 para una muestra de estos inconvenientes.

sí, pero no serviría de mucho. Las necesidades de quienes consultan los corpus no suelen estar referidas a esas fotografías generales que distribuyen los resultados por siglos. El precio que hay que pagar en las consultas al no poder disfrutar de un panorama general como el que devuelve el CE queda más que sobradamente compensado con la posibilidad de obtener resultados diferenciados para cualquier tramo temporal. La escasa utilidad de la indicación de los años con las frecuencias más altas en el CORDE se justifica por el hecho de que es un rasgo adoptado del CREA (que tiene un abanico temporal mucho menor) y, sobre todo, por la evidencia de que cualquier persona interesada en obtener la frecuencia de una expresión por tramos temporales, países, tipos de texto, área temática, autor u obra puede conseguirla con suma facilidad. En realidad, para seguir la argumentación de Davies, reproducida más arriba, según la cual los usuarios quieren conocer la frecuencia de una palabra en diferentes siglos u otros períodos históricos y su afirmación de que es aquí donde el CORDE "begins to exhibit some serious weaknesses", las cosas suceden casi exactamente al revés. La ventaja del CE es que da todas las frecuencias que 'conoce' de una vez, diferenciándolas por siglos. El CORDE, en cambio, puede dar las frecuencias correspondientes a cualquier tramo temporal, aunque, precisamente por esa amplitud de posibilidades, necesita que se le diga en cada caso qué es lo que se pide y, en consecuencia, requiere tantas consultas distintas como períodos haya que diferenciar. Parece bastante lógico. ¿Cuál es el corpus que en este punto muestra debilidades? ¿Cuál el más útil para la investigación lingüística profesional?

Lo mismo sucede con los demás factores, pero aquí ya no se puede hacer la comparación simplemente porque el CE no admite esas posibilidades. ¿Se puede considerar que es tan excelente y tan superior a CORDE y CREA como cree Davies un corpus del español que no permite conocer la distribución de una expresión por países? El CREA y el CORDE están organizados para, entre otras cosas, hacer posibles esas búsquedas, lo cual ha supuesto un enorme esfuerzo en la codificación y también en la orientación en el sistema de recuperación que, hay que reconocerlo, es antiguo (la primera versión, que se ha mantenido igual en la mayor parte de estos aspectos, fue abierta al exterior en 1998) y poco brillante en algunos componentes. Pero, a pesar de ello, permite obtener el número de casos (frecuencia general

y normalizada) de una expresión en textos argentinos, de prensa, con el comercio como tema específico y editados entre 1918 y 1939, posibilidad absolutamente inalcanzable para el CE. También es posible, aunque no sea habitualmente utilizada esta opción, la consulta específica de un determinado texto o de un autor concreto.

Creo que lo anterior es suficiente para llegar ya a una conclusión clara. Ante una consulta del tipo *enerv**, el CE solo puede devolver las frecuencias generales y normalizadas agrupadas por siglos, también por tipos de textos en el XX, la relación de formas implicadas en esa expresión y los casos correspondientes a cada una de ellas, limitados a cierto tramo temporal (y, en el XX, a los tipos de texto considerados) si la consulta selecciona esa opción. La aplicación de consultas del CORDE y el CREA no agrupa las formas integradas en la expresión regular, pero permite acotar las consultas utilizando cualquier valor en todos los parámetros que han servido para la configuración del corpus (año, país, tipo de texto, área temática e incluso autor y obra). Absolutamente imposible para el CE es, por citar un ejemplo que he manejado (con fines muy diferentes) en otra ocasión, permitir la investigación de en qué medida sigue García Márquez su conocida recomendación de no utilizar adverbios en *-mente*: **mente* en la casilla de 'Expresión' y 'García Márquez' en la de autor devuelve los resultados correspondientes, tanto en el CREA como en el CORDE.[9]

Hay todavía más diferencias. Davies señala, con toda razón, que las frecuencias generales sirven de poco y hay que utilizar las normalizadas (habitualmente, casos por millón de formas). El CE da, de entrada, las dos cifras para cada uno de los períodos en los que estructura la respuesta. El CORDE y el CREA, en cambio, solo facilitan al principio la frecuencia general, pero proporcionan también el modo de obtener la normalizada: usando el enlace que aparece en la pantalla de respuesta y haciendo la consulta a la nómina se obtiene el número de palabras, el número de documentos y la lista de los documentos con todos los detalles relevantes que corresponden a la zona seleccionada. Con una simple división se puede obtener

[9] Que no son solo adverbios, claro, sino todas las palabras terminadas en *-mente*. La localización de los casos de interés se obtiene con la simple ordenación de los resultados por la forma pivote. *Cf.* Rojo (2009) para más detalles. Es fácil ver que se trata de una búsqueda computacionalmente muy costosa (el 'comodín' está a la izquierda), de modo que requiere bastante tiempo para dar el resultado. La espera se reduce si se ayuda al sistema añadiendo la restricción a 'libros' y a 'Colombia', por ejemplo.

la frecuencia normalizada que corresponde a la general que se ha recibido un momento antes. Sin duda es más cómodo verlo todo de una vez y no tener que hacer el cálculo, pero, de nuevo, las posibilidades del CE se quedan reducidas a los períodos en que está rígidamente estructurado; no hay forma de obtener las frecuencias ni generales ni normalizadas para los ejemplos de un período que no sea un siglo ni de un cierto país, área temática, tipo de texto... Las diferencias tienen el mismo fondo: la aparente brillantez del CE se consigue mediante la imposición de una estructura fija, reducida a la diferenciación por siglos (y tipos de texto en el XX); el CORDE y el CREA tienen que buscar y calcular en cada caso los datos solicitados, pero gracias a ello permiten obtener no solo las frecuencias, sino también los casos correspondientes a cualquier combinación de valores de los parámetros deseada por los usuarios. Esta es la situación real y tenerla en cuenta y valorarla técnicamente permite también juzgar la afirmación de Davies según la cual "[i]f it [el CORDE, G.R.] does 'know' the frequency of all words and phrases in all historical periods, it certainly does not allow researchers to use that information as part of the query" (Davies 2009, 144). El CORDE no 'conoce' con antelación las frecuencias correspondientes a las consultas posibles. No podría, puesto que está basado en la posibilidad de que la búsqueda se refiera a cualquier combinación de valores en los cuatro grandes parámetros utilizados en su construcción. No las 'conoce', pero puede calcularlas. El CE ha precalculado las frecuencias distribuidas por siglo, pero no puede proporcionar ni las correspondientes a otros tramos temporales ni a ningún otro factor relevante en la organización del corpus. ¿Cuál es el corpus más útil para la investigación lingüística profesional?

El CE devuelve los casos correspondientes a la consulta realizada en un orden preestablecido que se basa en primer lugar en el siglo (el más reciente primero) y luego en algún otro factor de organización interna, totalmente desvinculado de las características internas de los textos. En cambio, la aplicación de consultas de CORDE y CREA permite también ordenar la salida de diversas formas, con lo que se recibe una ayuda considerable. Por ejemplo, en un caso como el que estamos analizando, cabe la opción de ordenar los resultados obtenidos por la forma pivote, de modo que las pantallas sucesivas van mostrando, por orden alfabético, todos los casos de *enerva* (95),

luego los de *enervaba* (31), *enervaban*, etc.[10] Es cierto que la aplicación no facilita un resumen de los resultados obtenidos clasificados por períodos (ni por ningún otro factor), pero esa carencia se explica, como ya he señalado, por el carácter abierto con que se han construido estos dos corpus. Téngase en cuenta que esa misma búsqueda podría haber sido adicionalmente restringida a textos colombianos o textos de prensa o publicados entre 1543 y 1861, por ejemplo. Los resultados parecen muy brillantes en la primera aproximación al CE, pero están limitados a un único parámetro e imponen una clasificación rígida que los investigadores solo pueden superar a base de analizar los casos de forma individual. Esta diferencia, evidente para cualquiera que haya utilizado el CREA o el CORDE, permite rebatir las afirmaciones de Davies acerca de las posibilidades del CORDE en la agrupación de los casos de las formas correspondientes a una expresión regular. Según Davies, en una consulta relativa a la expresión *des*?ento* entre 1200 y 1300[11],

> [t]he corpus (el CORDE, G.R.) indicates that there are '797 casos en 81 documentos'. One can then page throught all of the 797 tokens –one by one– and manually count up the total for ecah different form (*desfazimiento*, *desaffiamiento*, etc.) to see how frequently each one occurs. This would, however, take an hour or two.
>
> (Davies 2009: 147)

Cualquiera puede comprobar que la simple petición de ordenación de los resultados por la forma pivote produce una salida en la que se puede ver inmediatamente que hay un caso de *desabimiento*, cuatro de *desacordamiento*, cuarenta y seis de *desafiamiento*, etc.[12] Parece que Davies no ha visto esta posibilidad, puesto que nunca se refiere a la ordenación de los ejemplos, lo cual significa que no ha documentado en el grado esperable las pruebas en las que basa su comparación.

Además de por la(s) forma(s) pivote(s), la aplicación del CORDE (y del CREA) permite ordenar los resultados obtenidos por la forma que ocupa una determinada posición a izquierda y derecha (en una ventana

---

[10] No da directamente el número de casos, pero, como cada ejemplo lleva un número de orden, ese dato se obtiene con suma facilidad.

[11] Davies dice 1200 – 1400, pero ya he indicado que es un error. Las cifras que él da como resultado corresponden a la horquilla temporal que menciono en el texto.

[12] La aplicación, insisto, no da las cifras directamente, pero los ejemplos llevan un número de orden, de modo que una simple resta da el número de casos correspondientes a cada forma.

de un máximo de cinco formas a cada lado). Así, haciendo la búsqueda conjuntamente sobre *enervante* y *enervantes* y ordenando la salida por la forma que ocupa la posición inmediatamente anterior, se ve rápidamente que hay cuatro casos de *acción enervante*, pero solo dos de *clima enervante* y otros dos de *olor enervante*. Ciertamente, esas posibilidades no equivalen a la obtención de todas las formas que coaparecen con *enervante* o *enervantes* en un margen contextual determinado, pero resulta muy útil y cubre una buena parte de las necesidades que puede tener quien requiera datos de este tipo. El resto, lo que la aplicación de consultas no proporciona porque no ha sido diseñada para ese tipo de explotación, puede ser conseguido fácilmente con medios complementarios (*cf. infra*) gracias a las facilidades de exportación de resultados.

Las posibilidades de ordenación de los resultados no se reducen a las formas consultadas o las que están en su entorno. Resulta habitualmente mucho más útil la ordenación de los ejemplos por año (no solo por siglos). En el ejemplo que he venido utilizando (las expresiones correspondientes a *enerv**), el uso de esta opción en la configuración de la salida da inmediatamente la indicación de que el primer caso de *enervar* registrado (en el CORDE) procede de la traducción de la *Eneida* hecha por Villena (1427-1428). Es evidente que esta posibilidad es mucho más adecuada y útil que la existente en el CE, que ordena, en bloques, por siglos, pero luego organiza la salida en una secuencia que no se relaciona, hasta donde puedo ver, con factores propios de los textos. Ello significa que la localización de la primera documentación de una forma surge inmediatamente en el CORDE y en el CREA, pero obliga a revisar, uno a uno, todos los casos del siglo más antiguo de cada respuesta y tratar de hallar la fecha de cada texto[13]. Con un ejemplo adicional, la consulta sobre la forma *élite* (con tilde) y la ordenación por años de los 103 casos contenidos en el CORDE da inmediatamente la indicación de que

[13] Ya he indicado que las fechas no están siempre en las fichas de los textos, lo cual supone una incomodidad importante. En otros casos, hay fallos en la organización cuya causa no se me alcanza. El ejemplo más antiguo de *enerva* que registra el CE procede del libro tercero de la *Política indiana* de Solórzano Pereira. La ficha proporcionada por la aplicación data el texto en 1614, pero el enlace a la página correspondiente de la Biblioteca Virtual Cervantes da fallo (no es un simple cambio de nombre: la BVC ya no contiene ese texto) y la relación de concordancias lo atribuye al período 1500-1599. En realidad, el primer tomo de esta obra apareció, en latín, en 1629 y el segundo, en 1639. La traducción al castellano, hecha por el propio Solórzano Pereira, con ampliaciones, fue publicada en 1647. *Cf.* Tomás y Valiente 1996: xxiv-xxv.

las dos primeras documentaciones están en el *Facundo* de Sarmiento (cuya primera edición es de 1845) y las dos siguientes, unos cuarenta años después, a los *Recuerdos de viaje*, del también argentino Lucio Vicente López. Todos los datos pertinentes aparecen en las pantallas de resultados. La misma consulta al CE da como resultado la atribución al período 1800-1899 de veinte de los setenta y siete casos registrados, pero la localización del más antiguo de ellos exige ir abriendo las fichas de los textos, una a una, e identificar la fecha más antigua. ¿Qué es lo que, para un problema tan habitual como este, 'conoce' el CE? La aplicación solo vincula los textos a los siglos y en el interior de cada siglo hay que recurrir a los datos complementarios (las fichas) para obtener lo que necesitamos. En realidad, como hemos visto ya, el problema no está solo en la aplicación de consultas, sino en el bajísimo grado de codificación que tienen los textos incluidos en este corpus.

Además, el CREA y el CORDE admiten la ordenación de los casos obtenidos por país, área temática o tipo de texto. Queda claro, me parece, que la combinación del grado de detalle y la multiplicidad de parámetros que se pueden introducir en la formulación de la consulta con las diferentes ordenaciones admisibles en la devolución de los resultados hacen tanto del CREA como del CORDE, a pesar de la antigüedad de la aplicación de recuperación, unos recursos insustituibles para la investigación del español, con evidentes deficiencias, pero notablemente superiores al CE, como espero haber demostrado con este primer bloque de ejemplos.

5. BÚSQUEDAS DE CARÁCTER GRAMATICAL

Veamos ahora cuál es el comportamiento de los corpus que estamos comparando en la investigación de cuestiones gramaticales del estilo de la planteada por Hoffmann sobre los verbos modales del inglés contemporáneo. La diferencia de partida es muy clara: el CE tiene la enorme ventaja de estar parcialmente lematizado, mientras que el CORDE y el CREA carecen de este nivel de anotación. En consecuencia, búsquedas tan simples como todos los casos de un cierto verbo, de un sustantivo seguido por dos adjetivos, de todos los casos de futuro de subjuntivo o del verbo *ir* + *a* + infinitivo no son posibles de forma directa ni en el CORDE ni en el CREA y sí lo son en el

CE. Es una diferencia muy clara, que reduce las posibilidades de uso del CREA y el CORDE para estudios netamente gramaticales. Sería deseable que la RAE diera cuanto antes los pasos necesarios para poner a disposición de todos los investigadores versiones anotadas lingüísticamente de estos dos corpus.

Pero eso no significa que no sea posible utilizar el CREA y el CORDE en estudios gramaticales. Dado que el problema está en la ausencia de anotación lingüística, no es difícil prever la existencia de un buen número de problemas en los que la rica codificación de estos corpus (aspecto en el que el CE es muy deficiente) permitirá obtener resultados más fiables que los que se pueden alcanzar con el CE. Mi propio trabajo en los últimos años me ha llevado a enfrentarme en varias ocasiones con situaciones en las que es precisamente eso lo que sucede.

Un caso que me parece ilustrativo es el del uso actual de las formas correspondientes al copretérito de subjuntivo (*cantara* y *cantase*) y las compuestas correspondientes, así como las líneas generales del proceso mediante el cual se ha llegado hasta ahí. El CE permite recuperar en solo dos consultas (una para cada forma verbal) los datos con las frecuencias generales. Los reproduzco aquí en el cuadro número 1, al que añado los porcentajes correspondientes.

	-ra	*-se*	Totales	%-*ra*	% -*se*
XVIII	13 778	34 657	48 435	28,45	71,55
XIX	47 335	40 891	88 226	53,65	46,35
XX	37 271	4474	41 745	89,28	10,72
Totales	98 384	80 022	178.406		

Cuadro 1. Frecuencias generales de las dos formas del pretérito de subjuntivo en tres siglos distintos según el *Corpus del español*. Fuente: *Corpus del español* (http://www.corpusdelespanol.org [comprobado el 26/7/2010]). Elaboración propia.

Algo no muy distinto aparecería para las formas compuestas relacionadas, pero no es necesario aquí ir más allá. Los datos muestran con toda claridad que la utilización de estas dos formas verbales ha dado un vuelco total en los últimos trescientos años de historia del español. La forma en *-se* ha pasado de suponer el 71,55% del total de ambas en el siglo XVIII a representar un exiguo 10,72% en la totalidad del siglo XX. La consideración general puede tomar estos datos en bruto, puesto que aquí se trata del conjunto de valores que poseen

estas dos formas, de modo que ver la distribución cuantitativa general tiene sentido y consistencia plena.

En los datos correspondientes al siglo XX es posible dar un paso más y obtener la distribución por tipos de texto, que es la que figura en el cuadro 2.

Tipo de texto	-ra	-se	Totales	% -ra	% -se
Académico	4815	394	5209	92,44	7,56
Periodístico	6223	797	7020	88,65	11,35
Ficción	18637	2530	21167	88,05	11,95
Oral	7596	753	8349	90,98	9,02
Totales	37 271	4474	41 745		

Cuadro 2. Frecuencias generales de las dos formas del pretérito de subjuntivo en diferentes tipos de textos del siglo XX según el *Corpus del español*. Fuente: *Corpus del español* (http://www.corpusdelespanol.org [comprobado el 26/7/2010]). Elaboración propia.

Los porcentajes correspondientes a la forma en *-se* son siempre bajos, pero sorprende, dado lo que creemos saber acerca de la evolución del uso de estas dos formas alternantes, que el porcentaje más reducido de *-se* corresponda a los textos de tipo académico, seguidos luego por los orales. Con los datos del cuadro parece que el proceso de sustitución de *-se* en todos sus valores por *-ra* está más avanzado en los textos académicos que en los orales y en cualquiera de ellos más que en los periodísticos o de ficción. No parece que eso sea exactamente así, de modo que estamos ante un efecto perverso producido probablemente por el excesivo tamaño del tramo temporal considerado (la única posibilidad que permite el CE, como hemos visto repetidamente).[14]

Esto es todo lo que puede obtenerse con el CE: frecuencias por períodos fijos y excesivamente amplios y posibilidad de observar diferencias según tipos de texto en el siglo XX (naturalmente, para todo el siglo XX en conjunto). La evolución es clara, pero parece evidente que no hay lingüista que se pueda quedar conforme con una presentación tan general. Lo malo es que el CE no permite pasar de ahí. Cualquier modificación en los períodos considerados (o el estudio por países, áreas temáticas, etc.) supondría un trabajo ímprobo,

[14] Es probable que el CE haya otorgado la consideración de orales a textos que, en realidad, pertenecen a otros tipos, con la consiguiente deformación de los datos que facilita. *Cf. infra*, nota 19.

puesto que la reducción de la consulta a los casos correspondientes a un siglo devuelve los ejemplos ordenados según los textos de los que proceden, así que no hay modo razonable de refinar la distribución temporal y es absolutamente imposible estudiar la distribución por países, tipos de texto, etc.

El CORDE y el CREA, en cambio, tienen una codificación muy detallada y un sistema de consultas abierto, que permiten, por ejemplo, trabajar con períodos de veinticinco años (o de diez o de treinta y dos si es aconsejable) para poder observar con el detalle necesario en cada caso la evolución experimentada por estas dos formas en los últimos trescientos años (o setenta o ciento cuarenta). El fuerte inconveniente de la falta de anotación gramatical puede paliarse haciendo las búsquedas correspondientes a las dos formas del pretérito de subjuntivo de los verbos más frecuentes. Ese fue el enfoque que adopté en Rojo (2008), reduciendo las búsquedas a las formas más habituales (1ª. y 3ª. de singular por un lado y 3ª. de plural por otro) de los 11 verbos más frecuentes del español, tanto en el CORDE como en el CREA. Ciertamente, hay que hacer un buen número de consultas distintas[15], pero la riqueza de los datos resultantes[16] compensa el esfuerzo necesario para obtenerlos, como muestran los que reproduzco aquí como cuadro 3 (Véase en la página siguiente, ligeramente adaptado del cuadro 4 de Rojo 2008).[17]

No hace falta entrar en detalles sobre el problema en cuestión para concluir que la única estructuración de los resultados permitida por el CE es excesivamente general y un tanto engañosa. Desde el primer tramo de veinticinco años del siglo XX hasta el último se produce una evolución cuantitativa muy superior a la que tuvo lugar entre 1700 y 1900. Considerar el siglo XX como una unidad distorsiona los resultados, de modo que es necesario que la recuperación de datos pueda hacerse de modo mucho más flexible. Eso lo permiten el CORDE y el CREA, pero no el CE.

---

15 Naturalmente, el número de consultas necesarias está en función de la cantidad de verbos analizados, el número de tramos considerados y el número de formas. Consultas del tipo 'tuviera o tuvieran' reducen a la mitad las precisas si se pretende obtener cada número por separado. Algo parecido sucede con los tramos temporales, los tipos de texto, los países, etc.

16 Téngase en cuenta que son datos parciales (tres formas de once verbos muy frecuentes) y con algunas complicaciones derivadas de la existencia de homografías, entre las que *fuera* es el caso más destacado. No obstante, todo indica que son datos fiables y significativos.

17 Nótese que en el cuadro falta el último período del CREA (2000-2004). Cuando se hizo la recuperación de datos, en 2006, este período estaba sin cerrar, de modo que preferí no tomarlo en consideración para estos recuentos.

	-ra	*-se*	Total	% *-ra*	% *-se*
1700-1724	823	1879	2702	30,46	69,54
1725-1749	1546	2144	3690	41,90	58,10
1750-1774	1314	1490	2804	46,86	53,14
1775-1799	1191	1547	2738	43,50	56,50
1800-1824	1271	1805	3076	41,32	58,68
1825-1849	2975	3484	6459	46,06	53,94
1850-1874	3658	2694	6352	57,59	42,41
1875-1899	10 258	7394	17 652	58,11	41,89
1900-1924	5158	4952	10 110	51,02	48,98
1925-1949	7305	3441	10 746	67,98	32,02
1950-1974	8798	3681	12 479	70,50	29,50
1975-1999	42 058	9039	51 097	82,31	17,69
Totales	86 355	43 550	129 905		

Cuadro 3. Frecuencias absolutas y relativas de las formas del tipo *cantara* y *cantase* de once verbos muy frecuentes en los textos españoles de CORDE y CREA. Elaboración propia (=Rojo 2008, tabla 4) a partir de los datos obtenidos en los corpus (www.rae.es). [Datos obtenidos en marzo de 2006].

En términos más generales, las diferencias entre los conjuntos de datos que se pueden obtener de cada corpus quedan plasmadas, para este caso, en los dos gráficos siguientes.[18]

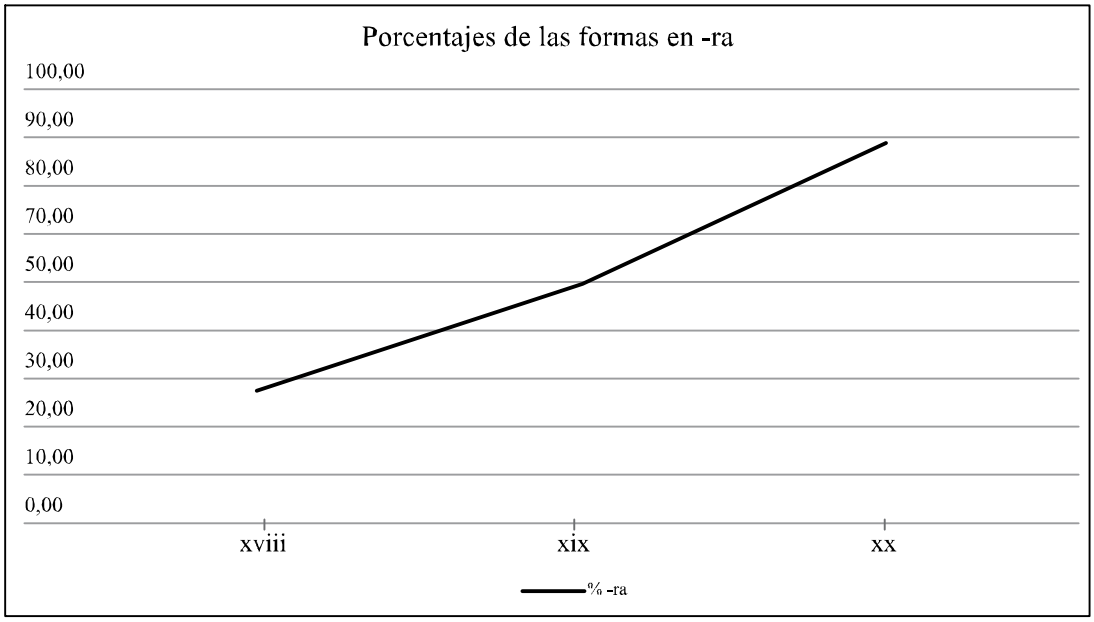

Gráfico 1. Evolución de los porcentajes de uso de las formas en *-ra* según los datos del *Corpus del español*. Elaboración propia a partir de los datos obtenidos en el CE (www.corpusdelespanol.org).

[18] Nótese que el gráfico número 2 toma en consideración únicamente los textos españoles.

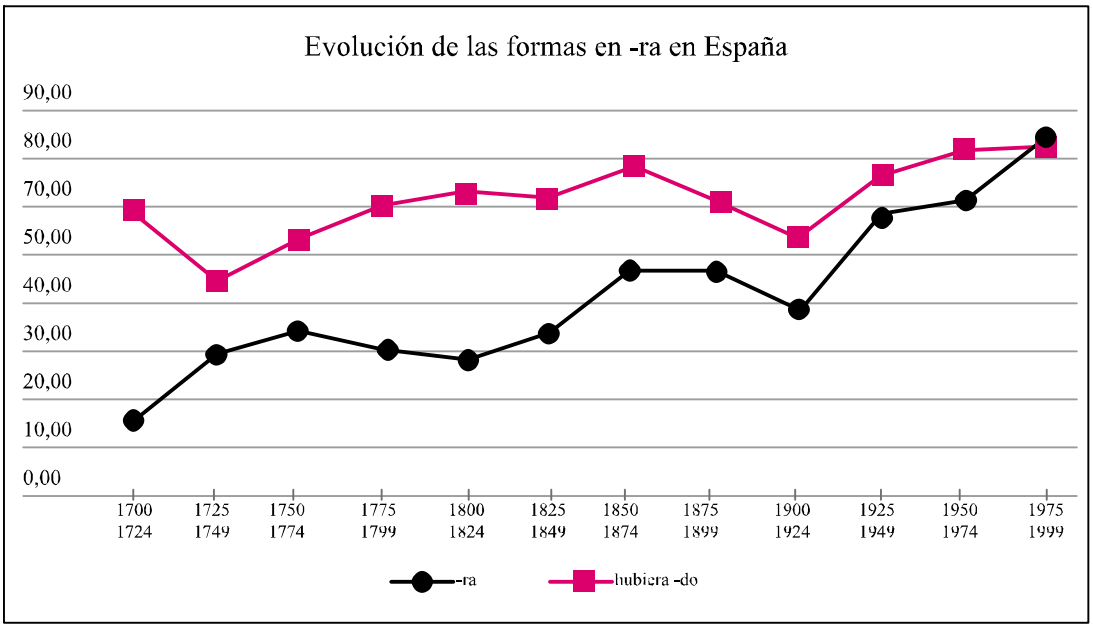

Gráfico 2. Evolución de los porcentajes de uso de las formas simples y compuestas en textos españoles de CORDE y el CREA. Elaboración propia a partir de los datos obtenidos en los corpus (http://www.rae.es). Reproducción del gráfico 2 de Rojo (2008: 171)

Las diferencias en el grado de detalle que se puede alcanzar en cada caso son evidentes y no parece necesario insistir en la muy distinta utilidad que para la investigación lingüística tienen las posibilidades de explotación del CE por un lado y el CORDE y el CREA por otro. El cuadro 3 y el gráfico 2 muestran que el proceso general de expansión del uso de las formas en *-ra* presenta fases de estabilización o incluso de retroceso cuyas causas habrá que investigar, pero que son indetectables si el análisis no llega a este grado de detalle, absolutamente imposible de conseguir con el CE. Pero la divergencia va mucho más allá de los distintos grados de detalle con que se pueden observar los resultados en la línea temporal. Más importante todavía es la distancia que establece el hecho de que el CE no permite obtener las frecuencias por países (o agrupaciones de países) ni por áreas temáticas ni por tipos de texto (salvo, en este último caso, para la totalidad del siglo XX). La bibliografía existente sobre la cuestión (*cf.* Conde Silvestre 2007: 152-153 y Blas Arroyo 2005: 85-90 para los aspectos sociolingüísticos) se refiere habitualmente a diferentes ritmos y porcentajes de sustitución en España y América, aunque parece claro que en todos los países hay diferencias según el carácter de los textos. Dado que no es posible entrar aquí en el fondo de la cuestión, véase el cuadro 4 como muestra del tipo de datos que es posible obtener del CREA

(y del CORDE). Me interesa especialmente resaltar que la adición de la anotación lingüística necesaria reduciría el número de consultas precisas (solo una por cada forma gramatical y tipo de texto en lugar de tantas como verbos tomados en consideración) y daría los datos correspondientes a todos los verbos, no solo a los once más frecuentes, pero no supondría aumento, con la misma 'arquitectura', del grado de detalle (la 'granularidad') que es posible conseguir en la versión actual. En el CE, en cambio, se requeriría una remodelación total del sistema de codificación, lo cual parece sencillamente imposible.

	-ra(n)	***-se(n)***	**N**
Argentina	84,15	15,85	4346
Bolivia	93,60	6,40	375
Chile	90,67	9,33	2465
Colombia	95,10	4,90	1672
Costa Rica	92,67	7,33	505
Cuba	87,16	12,84	1822
Ecuador	90,28	9,72	247
El Salvador	93,48	6,52	184
España	79,46	20,54	28 780
Estados Unidos	91,24	8,76	502
Guatemala	92,22	7,78	424
Honduras	86,43	13,57	199
México	84,39	15,61	5799
Nicaragua	92,59	7,41	378
Panamá	93,44	6,56	259
Paraguay	82,81	17,19	128
Perú	87,70	12,30	1374
Puerto Rico	84,09	15,91	748
Rep. Dominicana	93,30	6,70	522
Uruguay	91,70	8,30	675
Venezuela	89,86	10,14	1983
Totales	83,09	16,91	48 759

Cuadro 4. Porcentajes correspondientes a las formas en *-ra* y en *-se* en diferentes países según los datos del CREA. Elaboración propia a partir de los datos obtenidos en el corpus (http://www.rae.es/creanet.html [recogidos el 11/1/2010]).

De modo semejante, el CREA (y el CORDE) permiten obtener las frecuencias correspondientes a algunos de los diversos tipos de texto en que está organizado. Los resultados, que aparecen en el cuadro 5,

presentan diferencias evidentes con los que proporciona la rígida consideración de todo el siglo XX en un único bloque (*cf. supra*, cuadro 2):

	-RA(N)	*-SE(N)*	N=
Ficción	81,49	18,51	41 296
Libros (no ficción)	83,52	16,48	26 096
Periódicos	85,43	14,57	10 593
Revistas	89,68	10,32	2 499
Orales	90,39	9,61	2 507

Cuadro 5. Porcentajes correspondientes a las formas simples en *-ra* y en *-se* en diferentes tipos de texto reconocidos por el CREA. Elaboración propia a partir de los datos obtenidos en los corpus (http://www.rae.es) el 25/8/2010.

Incrementando el número de búsquedas individuales, tanto el CORDE como el CREA permiten obtener las frecuencias correspondientes al cruce del país con el tipo de texto, el área temática o diferentes tramos temporales, además de combinar tres o incluso más de esos parámetros. Como es obvio, cada rasgo adicional aumenta las búsquedas necesarias, pero el diseño de estos corpus lo permite y, como he indicado, la incorporación de la anotación morfosintáctica simplificará enormemente el número de operaciones e incrementará su validez general. En el CE, nada de esto, tan importante, es posible ni puede serlo porque su diseño no lo admite.

6. Otros aspectos relevantes

Los dos casos que hemos analizado muestran, creo que de forma muy clara, las ventajas y los inconvenientes del CE por un lado y del CORDE/CREA por otro. El panorama resultante es muy distinto del trazado por Davies, que no menciona características y posibilidades importantes del CREA y el CORDE y parece reducir la finalidad de los corpus a proporcionar tablas de frecuencias por siglos de aquello que se puede calcular con facilidad cuando se renuncia a tratar muchos de los rasgos de los que la investigación lingüística no puede prescindir.

Aunque no entran en la comparación que hace Davies, hay algunos aspectos de interés a los que debo hacer referencia, siquiera de pasada. En primer lugar, hay que tener en cuenta que la recogida y codificación de los textos que componen el CORDE y el CREA ha sido muy cuidadosa, aunque, por supuesto, puede contener errores. La adecuada codificación de los textos permite luego que los usuarios puedan ver

directamente todos los datos pertinentes acerca del texto en la misma pantalla de devolución de las concordancias, mediante la barra de desplazamiento horizontal o bien en un 'bocadillo' que aparece cuando el puntero se sitúa sobre la palabra pivote. El CE presenta, en una pantalla distinta, el año (pero no siempre), el autor, el título y la procedencia del texto, que es en muchos casos (pero no en todos) el enlace a los datos del texto en la fuente original. Esa información no existe en muchos casos. Por ejemplo, los enlaces a muchas páginas de la BVC dan fallo ahora (agosto de 2010) bien porque han cambiado de nombre, bien porque esos textos ya no están en ella (*cf.* lo apuntado *supra* para la *Política indiana*). En otros casos, la ficha, probablemente completa para su misión original (distinta de la integración en un corpus), no da los datos que necesitaríamos (el país de origen, por ejemplo). Esta falta de detalle tiene gran importancia y no solo para la recuperación de la información. En la investigación tradicional, cuando los textos eran expurgados 'a mano' y casi siempre por la misma persona que luego hacía la investigación, el valor que podía atribuirse a un ejemplo obtenido derivaba con naturalidad del conocimiento del texto en cuestión. Cuando, ahora, los corpus textuales nos permiten recuperar en segundos cientos o miles de casos de una determinada expresión, esa valoración es imposible para muchos de ellos, puesto que no situamos bien los textos de los que proceden. Si a ello se suma el desconocimiento de las características que nos permitirían aproximarnos a su valoración, el resultado puede desembocar en una valoración homogénea de ejemplos que deberían tener una consideración muy distinta. El CE simplemente no puede dar esos datos porque, como dice Davies en varias ocasiones, no los 'conoce', no están en la codificación del texto.

En un nivel más relacionado con la organización interna del corpus, creo importante señalar que, tanto en el CREA como en el CORDE, los datos que se facilitan de cada texto, así como los que figuran en la nómina asociada, son parte de la codificación del propio texto, no residen en documentos ni aplicaciones exteriores. Por supuesto, eso no evita la existencia de errores, pero sí hace que sean congruentes y, por tanto, más fáciles de detectar y corregir. Los datos de los textos en el CE están distribuidos en lugares diferentes, lo cual, además de los continuos fallos en los enlaces a las páginas de las que proceden los textos, producen discrepancias como la señalada antes para la obra de Solórzano Pereira o la mucho más llamativa aparición de 1582 como fecha de las dos partes del *Quijote*, con lo cual, por cierto, todo lo que pueda haber de peculiar

en la obra se transfiere al siglo XVI. Por supuesto, no son los detalles, más o menos anecdóticos, lo que importa, sino la forma en la que el CE ha sido construido y el riesgo continuo que ello supone.

Tanto el CE como el CREA/CORDE permiten la exportación de las concordancias (o de los fragmentos más amplios que se pueden obtener también). Es una característica de gran importancia, porque, aunque la argumentación de Davies parece olvidarlo con mucha frecuencia, la aplicación de consultas llega hasta un cierto punto, punto a partir del cual en la mayor parte de las ocasiones es necesario recurrir a otros programas con los que se pueda llevar a cabo el trabajo adicional que el investigador necesita. La exportación, por tanto, debe ser cómoda e integrar, además del texto de la concordancia, todos los datos del texto del cual procede el fragmento. El CE permite la selección y exportación de un número amplio de concordancias en cada paso, pero, naturalmente, no facilita más que lo que aparece en esa pantalla, esto es, el siglo, el título (abreviado) y la concordancia. El CREA y el CORDE, en cambio, tienen una limitación a veinticinco ejemplos de cada vez (debida a las circunstancias de la época en que fue diseñada), pero trae consigo el texto y todos los datos que lo sitúan en el conjunto del corpus, esto es, país, año, autor, obra, área temática, etc. Trabajar con los datos temporales (el año) con la exportación del CE supone que hay que partir de la concordancia amplia, lo cual exige ir ejemplo a ejemplo, o bien añadir por algún otro procedimiento (manual) el año que corresponde a cada caso. Cualquiera de las dos vías es costosa e incómoda, pero posible, cosa que no sucede con los demás rasgos de interés en la valoración y situación del ejemplo. Todos esos datos, en cambio, están presentes en las exportaciones del CREA y el CORDE.

A las posibilidades de búsqueda de palabras y expresiones añaden el CORDE y el CREA una aplicación para la consulta de la nómina, que utiliza, como es natural, los mismos parámetros que estructuran los textos. Los interesados pueden conocer, por ejemplo, el número de comedias con fechas comprendidas entre 1580 y 1640 contenidas en el CORDE, sus autores, títulos, fechas, editores críticos del texto incorporado para cada una de ellas y el número de palabras que tiene cada una y suman en conjunto, con lo cual es posible calcular la frecuencia normalizada de una determinada expresión en esta clase de textos. Lo mismo para textos de los diferentes países, áreas temáticas o equivalentes o cualquier combinación de dos o más de estos rasgos. Cabe también la búsqueda de

una determinada obra o de obras de un determinado autor. El correlato de esta base de datos en el CE es una hoja de cálculo que contiene los escasos datos de los textos que maneja, de modo que no es posible realizar búsquedas ni agrupaciones como las mencionadas porque no hay datos sobre el país ni el área temática ni el tipo de texto más que, en este último parámetro, en los pertenecientes al siglo XX.

De gran interés para el análisis lingüístico y filológico es la posibilidad que brindan tanto el CREA como el CORDE de recuperar los textos correspondientes a la consulta realizada con las marcas de codificación introducidas (todas o una parte de ellas). Es posible así estudiar características que se han expresado en el texto mediante una tipografía distinta, comprobar cuándo y cómo se ha desarrollado una abreviatura o la responsabilidad de un cambio sobre el texto original.

Los dos casos que he expuesto hasta aquí, inspirados en la presentación que hace Hoffmann (2008) de las líneas de investigación lingüística relacionadas con la explotación de corpus, cubren también lo más importante de los diversos aspectos (léxico, morfosintáctico y semántico) en diversos niveles de profundidad utilizados por Davies y en todos los aspectos relevantes, tanto de la organización del corpus y la codificación de los textos como de los sistemas de obtención de resultados. Creo que la conclusión es clara: la única ventaja real que tiene el CE sobre CORDE y CREA radica en que está parcialmente lematizado. Todas las demás ventajas aparentes implican el sacrificio de posibilidades de obtención de datos que cualquier lingüista considera imprescindibles: al reducir el parámetro temporal a la estructuración en siglos y la tipología textual a los correspondientes al XX es posible 'congelar' una amplia serie de cálculos que proporcionan esa apariencia brillante a la que me he referido en varias ocasiones. Pero el precio es muy alto porque la rigidez de la estructuración temporal fuerza agrupaciones de datos incongruentes con las necesidades de la investigación y, además, el procedimiento adoptado impide todas las demás vías de acceso a los datos necesarios.

En el nivel de mayor complejidad de los aspectos semánticos sitúa Davies la posibilidad de trabajar con "semantic fields, rather than just searching for words and phrases" (Davies 2009: 160). Ni el CORDE ni el CREA poseen nada parecido a la anotación semántica ni la vinculación de supuestos sinónimos, de modo que no tiene sentido intentar la comparación en este aspecto. Sin embargo, lo menciono aquí porque creo que el análisis objetivo de lo que el CE muestra en

este punto proporcionará elementos importantes para la evaluación independiente de este recurso.

Davies ha dotado al CE de 'powerful thesauruses' gracias a los cuales, al pedir los equivalentes del lema *oscuro* se obtiene "the frequency of all synonyms over time (and in different genres from the 1900s)" (Davies 2009: 164). Su tabla 14, que refleja parcialmente los resultados obtenidos, muestra, por ejemplo, que "*modesto* and *sombrío* have decreased since the 1800 (per million words), whereas *pesimista* and *opaco* increased from the 1800s to the 1900s" (*ibídem*). Habría que añadir, continúa, algún tipo de 'historical thesaurus' para la época medieval, pero lo importante es que el sistema está preparado para ello y puede incorporarlos en el momento en que exista.

Si se pide a la aplicación de consultas del CE lo propuesto por Davies, restringiendo la búsqueda al siglo XX para no introducir ahora complicaciones derivadas de factores históricos (no solo de la época medieval, naturalmente), devuelve la indicación de que hay 58 lemas que pueden ser considerados sinónimos de *oscuro*, que suman un total de 21161 ejemplos en los textos. Reproduzco los veinte primeros 'sinónimos' en el cuadro 6.

1	bajo [s]	7992
2	difícil [s]	3331
3	negro [s]	2446
4	pobre [s]	1772
5	oscuro [s]	749
6	sencillo [s]	580
7	cerrado [s]	577
8	cubierto [s]	451
9	moreno [s]	391
10	humilde [s]	284
11	pardo [s]	275
12	misterioso [s]	234
13	nocturno [s]	216
14	modesto [s]	201
15	bruno [s]	151
16	denso [s]	149
17	confuso [s]	147
18	incierto [s]	124
19	incomprensible [s]	106
20	apagado [s]	106

Cuadro 6. Relación de los 20 'sinónimos' más frecuentes de oscuro (solo para textos del siglo XX) según el CE. Fuente http://www.corpusdelespanol.org (datos obtenidos el 23/8/2010).

Tal cantidad y variedad de lemas sinonímicos de *oscuro* solo parece alcanzable a base de seguir la línea de establecer la asociación de todos los significados posibles de *oscuro* con todos los lemas que puedan presentar un valor similar en alguno(s) de sus usos. Es una aproximación automática débil, que requeriría el análisis detenido de cada una de las supuestas líneas de sinonimia, lo cual es difícil de llevar a cabo por medios manuales y simplemente imposible cuando se aplican procedimientos automáticos no especializados en el análisis semántico. La supuesta sinonimia de *oscuro* y *bajo*, *humilde* y similares se basa sin duda en el reconocimiento de la existencia de acepciones como la que aparece en tercer lugar de la edición actual del DRAE:

> 3. Dicho del linaje de una persona: Humilde, bajo o poco conocido
> (DRAE 2001, s.v. oscuro).

Saltar del marco controlado de esa definición, con el contorno semántico bien especificado, a los textos produce el reconocimiento de miles de equivalencias (falsas) entre *oscuro* y *bajo*, *oscuro* y *humilde* o, cambiando de acepción, *oscuro* y *pesimista*, etc. Abrir los casos vinculados produce la temida comprobación de que este módulo de la aplicación no solo es inútil, sino que puede ser un instrumento de fuerte deformación de los datos. Véanse, como muestra, los diez primeros casos de 'sinonimia' entre *oscuro* y *bajo*:[19]

1 19-OR Entrevista (ABC): ...pasado su momento y se habían olvidado, como los conciertos de laúd, como bajo continuo y también como instrumento solista en conciertos del siglo XVII y XVIII – el...

2 19-OR Entrevista (ABC): ...en su sitio, afinadas de la misma manera. Las otras cuatro cuerdas, bajos supletorios, añaden unos armónicos nuevos que suenan aunque yo no las toque, y...

3 19-OREntrevista (ABC): ...Viena en Madrid, porque tenemos tradición y calidad. Pero no hay que mirarlo bajo el prisma del espectáculo, sino del de la calidad. - ¿ Es usted

[19] Nótese, además, la discutible consideración de una entrevista publicada en un periódico como texto oral. Por otro lado, si necesitamos más datos (el año, por ejemplo) y pulsamos en el enlace de la fuente, llegamos a la edición del periódico ABC (de Madrid) del día en que se hace la consulta.

4 19-OR Entrevista (ABC): ...Fue "La Bohèmen, uno sus títulos emblemáticos, junto a Glanni Raimondi y bajo la dirección de Rafael FrGhbeck. Días más tarde repetía en Oviedo. Esto era...

5 19-OR Entrevista (ABC): Pero recibí muy buena enseñanza. Comencé a actuar en teatro. Hice de cadáver bajo un sofá en una obra mientras otros chicos dejaban caer cacahuetes sobre mi cara para...

6 19-OR Entrevista (ABC): ...mí ha mantenido durante un período de su vida que ha tenido que ser, bajo cualquier punto de vista imaginable, mucho peor que Rugby. No puedo concentrarme...

7 19-OR Entrevista (ABC): ...Castro salió extrañamente ileso. Luego la fecha vino a denominar a grupos terroristas que bajo la bandera del M26 - 7 actuaron en Santiago de Cuba y en La Habana

8 19-OR Entrevista (ABC): ...primera y la segunda jornadas tienen, respectivamente, 651 y 675 versos, número bajo para una comedia en tres actos, y más parejo con el de las piezas...

9 19-OR Entrevista (ABC): ...más prestigioso festival de música, el de Salzburgo,concluye esta semana su primera edición bajo el mandato directo de Gerard Mortier, que ha heredado el puesto pero ninguno de...

10 19-OR Entrevista (ABC): En su despacho del Festpielhaus, el nuevo director del Festival de Salzburgo está sentado bajo un gran retrato del difunto -y controvertido- dramaturgo austríaco Thomas Bernhard. No...

No parece que esto pueda servir para la investigación de los campos semánticos. Sin duda, Davies aplica algunos filtros contextuales sobre los ejemplos de los candidatos a lemas sinonímicos,[20] pero no parece servir de mucho. Es sorprendente, por ejemplo, que *oscuro* aparezca entre los sinónimos de *oscuro* y mucho más que solo 749 casos de los 1880 que el CE documenta para este lema en el siglo XX lo sean. Es decir, *oscuro* solo significa lo mismo que *oscuro* en el 39,84% de sus apariciones. Si se abre la búsqueda de sinónimos a todo el abanico temporal del CE se puede observar que *cubrir* aparece en sexta posición (por número de casos). Con resultados de este tipo, la

[20] El CE contiene 11 943 casos del lema bajo y 2745 del lema *pobre* en textos del siglo XX.

brillantez de los cuadros de frecuencias, que permite a Davies hacer afirmaciones absolutamente gratuitas como la reproducida antes acerca de las fluctuaciones de *optimista* o *modesto* como sinónimos de *oscuro*, es una trampa, puesto que no hay nada en la realidad lingüística que autorice una conclusión de este tipo. No vale el procedimiento y, en consecuencia, no valen los datos. Los 'equivalentes sinonímicos' del CE son una capa externa inservible y su utilización supone un alto riesgo.

7. CONCLUSIÓN

Creo que lo expuesto hasta aquí es representativo y suficientemente revelador de las ventajas e inconvenientes que presentan el CE de un lado y el conjunto CREA/CORDE de otro. En términos muy generales, el CE está parcialmente lematizado y tiene un sistema de consultas muy ágil y brillante, capaz de devolver en décimas de segundo las frecuencias generales y relativas de las formas correspondientes a una expresión regular o los candidatos a colocaciones de una palabra y mostrar cuadros con la distribución de esas frecuencias por siglos y, en el caso del XX, también del tipo de texto. A esas capacidades ha añadido Davies otras utilidades cuyo interés para la investigación me parece, cuando menos, discutible.

El CREA y el CORDE, en cambio, están basados en una codificación muy cuidadosa y detallada que permite recuperar los datos que interesan estableciendo el filtro por cualquiera de los parámetros que han intervenido en la construcción de los corpus (año, país, tipo de texto y área temática), así como autor y obra. Por ejemplo, la búsqueda puede centrarse en un país (Argentina) o varios países (Argentina, Uruguay y Paraguay) o todos los países. Algo parecido sucede con los demás parámetros, pero conviene destacar además que las áreas temáticas están jerarquizadas, de forma que es posible centrar una búsqueda en textos que traten de empleo o bien de industria o bien de economía, de cualquiera de las tres o, más en general, de cualquier tema incluido en el hipercampo correspondiente (política, economía, comercio y finanzas). Por supuesto, todos esos parámetros son combinables en una búsqueda única (casos de una expresión en textos de un cierto período, un país determinado, en textos de prensa y sobre temas económicos). La razón de todo ello es suficientemente clara:

lo que nos interesa en el análisis de un corpus o varios corpus es la comparación de lo que se encuentra en diferentes segmentos (temporales, geográficos, etc.). El CORDE y el CREA han sido diseñados para permitir que los investigadores obtengan los datos que necesitan en una consulta abierta a sus necesidades, siempre cambiantes. El CE solo permite la diferenciación por siglos y, únicamente para el XX, con tipos de textos. Esa fortísima restricción le permite precalcular algunos resultados y, por tanto, facilitar con rapidez esa información, pero creo que esa posibilidad implica pagar un precio excesivamente alto: no poder trabajar con parámetros como país, área temática o tipo de texto, de evidente importancia para el análisis del español, y reducir las posibilidades de análisis diacrónico a la consideración de los siglos en bloque (*cf. supra*). Lo que hemos visto sobre la evolución de las formas en *-ra* muestra lo peligroso de estas agrupaciones tan amplias y convencionales.

A esas evidentes ventajas del conjunto formado por el CORDE y el CREA sobre el CE hay que añadir que una buena parte de las desventajas o deficiencias señaladas por Davies no son reales. Estos dos corpus se basan en la posibilidad de seleccionar la información por los parámetros que han entrado en su configuración y también en la de utilizarlos para ordenar los resultados obtenidos. Por ejemplo, no agrupan en una entrada única las expresiones obtenidas como resultado de una consulta, pero sí permiten organizar la salida por las distintas formas relacionadas con una expresión regular, la que va inmediatamente a su derecha, por años, por países... Como siempre, la aparente comodidad del CE se basa en la restricción de las posibilidades de búsqueda y recuperación. El CREA y el CORDE son tan abiertos como lo requiere la investigación lingüística. Este factor hace también que el recorte de resultados a un máximo de 1000 (derivada de la antigüedad de la aplicación) tenga una importancia mucho menor que la que le atribuye Davies: se puede reducir el ámbito por cualquiera de los parámetros empleados, con lo que los resultados deseados aparecen, completos, tras la fragmentación necesaria para la consulta.

En términos muy generales, es cierto que el CORDE y el CREA no conocen "the frequency of all words and phrases in all historical periods" (Davies 2009: 144). Tampoco las facilita el CE, que solo

'conoce' las correspondientes a los siglos. La diferencia, la gran diferencia, está en que el CORDE y el CREA pueden calcular, cuando se pide, la que corresponde a cualquier tramo temporal, país, área temática, tipo de texto, autor y obra.[21] El CE solo proporciona las frecuencias correspondientes a los siglos y a los tipos de texto en el XX y no conoce ni puede calcular ninguna otra. No parece haber dudas acerca de qué es lo más útil para los investigadores.

La lematización parcial que posee es, sin duda, una ventaja del CE, cuya utilidad se ve reducida por las restricciones ya examinadas en la recuperación de datos. Dado que ni el CORDE ni el CREA están lematizados, no entro en este punto y omito las consideraciones que podría hacer sobre los resultados observables.[22]

La cantidad de textos y la información interna y externa que contienen son muy superiores en CORDE y CREA. Es lógico que sea así, puesto que el CE ha sido construido integrando textos que ya estaban en formato electrónico, mientras que tanto CORDE como CREA han digitalizado miles de textos y los han codificado de un modo que permite la recuperación selectiva de la información en un grado muy alto.

Hay ciertas divergencias en la imagen que se tiene de la construcción y la apertura a la consulta externa de un corpus textual. Sin duda existe una línea, creo que muy minoritaria, según la cual la consulta al corpus debería devolver la solución al problema planteado. La otra, la mayoritaria, considera que los corpus son conjuntos de textos bien articulados y adecuadamente codificados en los que los lingüistas pueden buscar y hallar los datos externos que deben procesar y estudiar para intentar conocer lo que sucede o ha sucedido en una lengua. En esta dirección, la posibilidad de trabajar con corpus distintos o con subcorpus diferentes del mismo corpus es una cuestión crucial. Y un corpus será tanto más útil cuanto más abierto esté a las posibles necesidades de quienes lo consulten.

---

[21] *Cf.* los datos sobre el uso de adverbios en *-mente* en obras de García Márquez mencionado *supra.*

[22] *Cf.* Davies 2008 para la exposición del método que ha seguido para resolver los numerosos y complejos problemas que se plantean en un trabajo de este tipo. Es curioso, por ejemplo, que la forma *braueza*, que, como hemos visto, utiliza para comparar el CE con el CORDE no haya sido lematizada a *braveza*. Más revelador del carácter del trabajo realizado me parece que solo exista un lema *cerca*, que engloba al sustantivo y al adverbio, solo un lema *duro*, etc.

Los usuarios, por su parte, no deben esperar obtener directamente la solución al problema que se han planteado, sino los datos contenidos en el corpus necesarios para ello. Por supuesto, sé que hay otras opciones, pero estoy convencido de que un buen corpus, un corpus útil, es aquel que, sin necesidad de llegar al radicalismo de Sinclair sobre la inutilidad de toda anotación, ha sido conformado en la línea de la prudencia y la neutralidad de la anotación señalada por Leech (1992). Además, dado que no es imaginable que los corpus generales permitan recuperar directamente los datos correspondientes a construcciones de cierta complejidad, el objetivo debería establecerse en facilitar que los usuarios puedan descargar datos no totalmente refinados para filtrarlos luego mediante la utilización de los programas adecuados; esto es, la técnica de la anotación manual a la que se refieren Smith, Hoffmann & Rayson (2008). Por último, la extracción de datos y su procesamiento individual está en la dirección apuntada repetidamente por autores como Gries (p.e. 2006), que han insistido no solo en la necesidad de mejorar la formación estadística de los lingüistas, sino también de una cierta soltura en la preparación de rutinas que permitan procesar la información contenida en los textos (*cf.* Schulte 2009 para algo similar en su enfoque diacrónico). El CREA y el CORDE se inscriben plenamente en esta orientación y, en ella, resultan mucho más útiles que el CE. Por número de textos, por su codificación externa e interna, por las posibilidades de selección de textos según diferentes parámetros que brinda el sistema de búsquedas e incluso por las facilidades de exportación que ofrecen. El panorama presentado por Davies es muy diferente del que resulta de un análisis objetivo, como espero haber demostrado.

Referencias bibliográficas

Blas Arroyo, José Luis. 2005. *Sociolingüística del español*, Madrid, Cátedra.

Briz Gómez, Antonio y Marta Albelda Marco. 2009. Estado actual de los corpus de lengua española hablada y escrita: I+D, en *El español en el mundo. Anuario del Instituto Cervantes 2009*, Madrid, Instituto Cervantes.

Conde Silvestre, Juan Camilo. 2007. *Sociolingüística histórica*, Madrid, Gredos.

CORDE, Banco de datos en línea / *Corpus diacrónico del español*, http://www.rae.es (versión cerrada en abril de 2005).

Corpus del español (construido por Mark Davies). En línea: http://www.corpusdelespanol.org.

CREA, Banco de datos en línea / *Corpus de referencia del español actual,* http://www.rae.es (versión cerrada en junio de 2008).

Davies, Mark. 2005. Advanced research on syntactic and semantic change with the *Corpus del español*, en C. Pusch, J. Kabatek, W. Raible (eds.), *Romance Corpus Linguistics II: Corpora and Diachronic Linguistics*, Tubinga, Gunter Narr: 203-214.

Davies, Mark. 2008. Spanish and Portuguese Corpus Linguistics, en *Studies in Hispanic and Lusophone Linguistics*, 1: 149-186

Davies, Mark. 2009. Creating Useful Historical Corpora: a Comparison of *CORDE*, the *Corpus del español*, and the *Corpus do português*, en A. Enrique-Arias (ed.), *op. cit.*: 137-166.

Enrique-Arias, Andrés (ed.). 2009. *Diacronía de las lenguas iberorrománicas. Nuevas aportaciones desde la lingüística de corpus*, Madrid/Frankfurt, Iberoamericana/Vervuert.

Gries, Stefan Th. 2006. Some Proposals towards a More Rigorous Corpus Linguistics, *Zeitschrift für Anglistik und Amerikanistik*, 54, 2: 191-202.

Hoffmann, Sebastian. 2008. Looking at language in use: some preliminaries, en S. Hoffmann, S. Evert, N. Smith, D. Lee & Y. Berglund Prytz, *Corpus Linguistics with BNCWeb - a Practical Guide*, Frankfurt, Lang: 1-12.

Institut d'Estudis Catalans, *Diccionari descriptiu de la llengua catalana*. Versión en línea. http://dcc.iecat.net/ddlc/index.asp. [Cit: DDLC]

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance, en Svartvik (ed.), *op.cit.*: 105-147.

Real Academia Española, 2001, *Diccionario de la lengua española*, 22ª ed., Madrid, Espasa-Calpe. [Cit.: DRAE]

Real Academia Galega.1997. *Diccionario da Real Academia Galega*, A Coruña, RAG. [Cit.: DRAG]

Rojo, Guillermo. 2008. De nuevo sobre la frecuencia de las formas *llegara* y *llegase*, en J. Albrecht und F. Harslem (eds.), *Heidelberger Spätlese. Ausgewälhlte Tropfen aus verschiedenen Lagen der spanischen Sprach- und Übersetzungswissenfschaft. Fetschrift anlässlich des 70. Geburtstages von Prof. Dr. Nelson Cartagena*, Bonn, Romanisticher Verlag: 161-182.

Rojo, Guillermo. 2009. El papel de los corpus en el estudio de la historia del español, intervención en la mesa redonda sobre Los corpus diacrónicos en la historia de la lengua española celebrada en el *VIII Congreso internacional de historia de la lengua española*, Universidade de Santiago de Compostela (14-18 de septiembre de 2009). En prensa en *Actas* del Congreso. (http://gramatica.usc.es/~grojo/En_prensa/Rojo_Corpus_diacronicos.pdf)

Rojo, Guillermo. 2010. Aguja de navegar corpus. En prensa en *Actas* del *XII Congreso de la Sociedad Argentina de Lingüística* (Mendoza, 6-9 de abril de 2010) (http://gramatica.usc.es/~grojo/En_prensa/Aguja_navegar_corpus.pdf).

Schulte, Kim. 2009. Using non annotated diachronic corpora: benefits, methods and limitations, en A. Enrique-Arias (ed.), *op.cit.*: 167-180.

Smith, Nicolas; Sebastian Hoffmann & Paul Rayson. 2008. Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations, en *Literary and Linguistic Computing*, 23, 2: 163-179.

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

Tomás y Valiente, Francisco. 1996. Introducción a la edición de la *Política indiana* de Juan de Solórzano Pereira a cargo de F. Tomás y Valiente y A. M. Barrero, Madrid, Biblioteca Castro.