EL PROYECTO DE INFORMATIZACIÓN DEL *DCECH*: APLICACIONES Y RESULTADOS*

J. M. BLECUA, G. CLAVERÍA, M. PRAT, C. SÁNCHEZ, J. TORRUELLA
Universidad Autónoma de Barcelona

J. R. MORALA, C. EGIDO, J. LEMEN
Universidad de León

M. BATLLORI, I. PUJOL
Universidad de Girona

El proyecto de informatización del *DCECH*[1] nació, en el año 1989, con el objetivo de relacionar fácilmente los datos que ofrecen los seis volúmenes de este diccionario. Esta investigación surgió por el interés general de los miembros del *Seminario de Filología e Informática* de la Universidad Autónoma de Barcelona por explorar las posibilidades que las herramientas informáticas estaban abriendo en el campo de la filología. El proyecto se está desarrollando actualmente gracias a cuatro ayudas institucionales del MEC y la CIRIT, e implica la colaboración de las universidades de León, de Gerona y de la Autónoma de Barcelona. A lo largo de estos años, conforme se profundizaba en los complejos artículos que componen la obra, el proyecto ha ido evolucionando, tanto desde el punto de vista filológico como desde el meramente computacional.

Este proyecto, cerrado en su primera parte porque informatiza los datos de una obra ya concluida, abre a los filólogos infinidad de posibilidades tanto de estudio de las informaciones de la obra como de homogeneización, ampliación y complementación de los datos que reflejan y hacen evidentes dichas informaciones. Ello, por tanto, convierte esta investigación en un trabajo abierto a largo plazo.

El objetivo primero y fundamental es la digitalización, es decir, la introducción de parte de las informaciones del *DCECH* en bases de datos rela-

* Investigación parcialmente financiada con una ayuda de la DGICYT (PB95-0656) y con el apoyo del Comissionat per Universitats i Recerca de la Generalitat de Catalunya (GRQ95-0544).

[1] J. Corominas y J. A. Pascual, *Diccionario crítico etimológico castellano e hispánico*, 6 vols., Madrid, Gredos, 1980-91.

cionales. No se trata, como se podría pensar a primera vista, de un trabajo mecánico sino que se requiere una labor importante de sistematización y organización de los datos contenidos en el diccionario, tarea nada fácil al considerar que, al ser una obra muy personal, junto con los datos filológicos aparecen impresiones, opiniones y juicios sobre cualquier aspecto de los vocablos estudiados.

El hecho de utilizar como soporte informático bases de datos se debe a que las consideramos el medio más adecuado para contener una obra de este tipo. Las obras lexicográficas, y en especial los diccionarios, casi siempre están construidas sobre dos ejes (*macroestructura* y *microestructura*). Las restricciones físicas del papel hacen que sólo sea posible el recorrido por los datos a través del eje de la macroestructura, perdiéndose así muchísimas informaciones accesibles solamente mediante las relaciones y los saltos entre los datos o referentes contenidos tanto en la macroestructura como, sobre todo, en la microestructura. Las bases de datos, al trabajar también sobre dos ejes (*registros* y *campos*), son el soporte idóneo para este tipo de obras, ya que no solamente permiten llegar a las informaciones por vías distintas al lema, sino que también facilitan la relación de todos los referentes distribuidos en ambos ejes. Con las bases de datos rompemos con la linealidad de consulta y el orden alfabético como único acceso posible en los diccionarios impresos, para pasar a una circularidad de múltiples accesos. Se evoluciona del salto manual entre referentes al salto hipertextual, abriéndose así un sinfín de posibilidades cada vez que se utiliza la obra. El usuario, al realizar una consulta, se adentra en una aventura personal como puede ser el camino a través de la infinidad de puertas que se encuentran en el recorrido hipertextual. Los datos siempre han estado en el diccionario, pero sólo ahora con la informática se puede disponer de los accesos más productivos. De este modo, se pueden obtener informaciones desconocidas incluso para el propio autor del diccionario, el cual, a pesar de haber introducido los datos no ha podido relacionarlos en todas sus posibilidades. Esta limitación humana es la que genera heterogeneidad y otras deficiencias frecuentes en los diccionarios no informatizados.

El uso de un soporte informático, si bien facilita las tareas de revisión y homogeneización de las referencias, exige diseñar una estructura de registros y campos que pueda contener todos los datos del diccionario que se consideran relevantes. Ello implica un estudio profundo de la estructura interna de los artículos y una considerable capacidad de relación para homogeneizar y colocar en su justo lugar aquellos datos que aparecen de formas distintas y en sitios dispersos en la obra.

En la informatización del *DCECH*, los datos seleccionados de cada entrada del diccionario se distribuyen entre los campos de las cuatro bases de datos que se utilizan actualmente: a) una base de datos central, de 23 campos, que recoge las informaciones de carácter general sobre el término estudiado; b) una base de datos etimológica, de 9 campos, que incluye las informaciones genealógicas del lema en cuestión (historia de la procedencia de la palabra); c) una base de datos documental, de 8 campos, que

recopila todas las informaciones referentes a los textos y documentos que atestiguan el lema (primera documentación y otras dataciones posteriores); y d) una base de datos sobre la pervivencia, de 7 campos, que da cuenta de aquellas informaciones relacionadas con la historia de la palabra (distribución geográfica, campo semántico, nivel diastrático, época de uso, etc.). Las informaciones de estas cuatro bases de datos se enlazan mediante un campo común a todas ellas, que es el término estudiado.

El número de bases de datos relacionales puede ampliarse según vaya creciendo el proyecto y surjan nuevos intereses. De este modo es previsible, por ejemplo, que se estructuren nuevas bases de datos exclusivamente para las informaciones de carácter fonético, morfológico, sintáctico, semántico, que siempre podrán entrar en relación con las demás.

Además, está previsto complementar los resultados del proyecto con una bibliografía de artículos aparecidos sobre vocablos u otras informaciones contenidas en el diccionario, que podrá fácilmente cruzarse con las unidades léxicas de corpus diacrónicos no solamente de lenguas hispánicas sino incluso románicas.

Las cuatro bases de datos también se conectan con un archivo que contiene el texto de la obra. Así pueden aparecer en la pantalla del ordenador los resultados de una consulta hecha a las bases de datos (por ejemplo, voces militares provenientes del italiano y que se documentan por primera vez en los siglos XVI-XVII = *esguazar*) y, al mismo tiempo, la reproducción del texto de Corominas y Pascual que hace referencia a dicha consulta. La relación entre los datos seleccionados y tipologizados en las bases de datos y el texto del *DCECH* se consigue a través de marcas o etiquetas añadidas al texto que vinculan las informaciones de las bases de datos con la parte del texto de donde han sido extraídas. De este modo se convierte el *texto* en un *hipertexto*, en el cual, a través de los referentes contenidos en las bases de datos, se pueden ir efectuando saltos entre informaciones del diccionario que se van relacionando indefinidamente.

Como cada artículo del *DCECH* reúne datos e informaciones de carácter muy heterogéneo, nos pareció que la mejor manera de poner orden a esta heterogeneidad con vistas a la explotación informática era el establecimiento de una clasificación de los tipos de información que se pueden encontrar en el diccionario. El etiquetado del complicado mundo de informaciones de cada uno de los artículos-monografía del *DCECH* homogeneíza y «supraordena» lo que inicialmente puede parecer un cúmulo de particularidades imposibles de relacionar.

Formalmente, el etiquetado sigue las normas *Standard General Markup Language* (SGML) en su versión *Text Encoding and Interchange* (TEI), ya que, avaladas por las principales asociaciones lingüísticas (*Association for Computers and the Humanities, Association for Computational Linguistics* y *Association for Literary and Linguistic Computing*) y organismos políticos (Comunidad Europea y Ministerio de Defensa norteamericano), se considera que son las más estandarizadas en el campo de la filología moderna. Con ellas aumentan las posibilidades de reusabilidad de nuestro trabajo, lo abren a

futuros objetivos tanto nuestros como de otros grupos de investigación y le aseguran la no caducidad, ya que, al ser etiquetas declarativas (describen estructuras y contenidos), no dependen de sistemas informáticos específicos, con lo que facilitamos el intercambio de documentos y aseguramos su buena transmisión a través de las redes informáticas.

Estas etiquetas marcan en cada momento el tipo de información que aparece en el diccionario (gráfica, fonética, morfológica, sintáctica, semántica, etimológica) o bien permiten identificar rápidamente, por ejemplo, nombres de lenguas y dialectos aludidos en los artículos (catalán, leonés, árabe, español de América, etc.).

De la explotación del banco de datos hipertextual resultante de la conexión de la base de datos con el texto se pueden derivar mejoras sustanciales para la lexicografía histórica y la investigación etimológica. Con unas pequeñas calas en nuestro banco de datos obtenemos ya algunos ejemplos de posibles resultados tanto (a) referentes al grado de homogeneización de los datos histórico-etimológicos como (b) a las posibilidades de obtener informaciones a partir de la relación de estos datos.

(a) Homogeneización

La homogeneización puede aplicarse de dos maneras diferentes. En el caso de la etimología de la palabra se trabaja con tipologías y estructuras de bases de datos que permiten recoger con un alto grado de sistematización su evolución genealógica. Distinguimos entre grados de fiabilidad de la etimología (cierta, incierta y desconocida), estadios genealógicos diferentes (en el caso de transmisión compleja) y tipo de transmisión (palabra patrimonial, palabra culta, palabra semiculta; derivado de una forma romance, derivado culto; compuesto de formas romances, compuesto culto; cruce, alteración, onomatopeya, etc.) en cada uno de los estadios. Esta triple distinción sistematiza la gran diversidad de redacciones en las que se presenta la información etimológica del primer párrafo de cada entrada del Diccionario: tanto, por ejemplo, la palabra que procede del latín vulgar que a su vez es derivada de una forma del latín (*abarcar*), como la palabra que procede del griego y que en esa lengua es un compuesto (*arqueología*) o un derivado con una forma prefijada (*dímero*), como la palabra que es un cruce entre dos formas (*guedeja*), o una alteración de una forma etimológica (*cerrojo*) o bien es un cultismo que procede de una forma del latín tardío que a su vez procede del griego y en esta lengua es un compuesto (*cenobio, cirugía*), etc. La sistematización en este aspecto es prácticamente completa y, por tanto, las investigaciones y relaciones en el nivel etimológico podrán llegar a ser muy diversas.

En otros casos, es el mismo contenido de la base de datos el que permite una sistematización posterior al proceso de la informatización. Por ejemplo, reflejando el principio de que la etimología es la historia misma de la palabra, el *DCECH* proporciona en ocasiones comentarios referentes a la va-

riación diastrática. Es interesante apreciar cómo este tipo de información se muestra muy heterogénea ya que no aparece en todas las entradas y, cuando se da, no se expresa de una forma sistemática.

Si se profundiza en el estudio de la variación diastrática, se puede observar que los comentarios del *DCECH* son sistematizables en cierta medida:

En general, la mayoría de los lemas que presentan información diastrática pueden analizarse desde dos puntos de vista distintos dependiendo de si se trata de vocablos pertenecientes a la lengua escrita o a la lengua oral. Así, entre los primeros, son frecuentes los términos pertenecientes al lenguaje literario o al lenguaje poético y, entre los segundos, los considerados como característicos del lenguaje rústico, jergal o de germanía. En los relativos al lenguaje literario se destaca, en ocasiones, la vigencia del vocablo [«vivo en la lengua literaria» (*erguir*)]; en otras, en cambio, simplemente se expresa el hecho de que se trata de un vocablo literario [«voz puramente literaria» (*ensalzar*)]. En cuanto a aquellas voces propias del lenguaje poético, el *DCECH* suele presentar comentarios muy concisos que, a veces, pueden complementarse con referencias sobre la difusión de la palabra. Contrástense «palabra del estilo poético» (*ebúrneo*) y «voz poética y rara» (*erebo*).

Cuando se trata de términos marcados como rústicos, o bien se comenta únicamente su adscripción a este lenguaje [«vocabulario rústico» (*gamella*)], o bien se justifica mediante autoridades [«palabra de rústicos, en la *Farsa* de Alonso de Salaya (3r cuarto del siglo XVI), en Mariana y en Cervantes, para el vestido de una pastora en el navarro Arbolanche (1566), 181r6» (*gabán*)]. Lo mismo ocurre con las voces de germanía: en muchas ocasiones, esta información aparece simplemente a partir de la abreviatura *gnía.* (*destebrechar* y *dupa*), mientras que en otras viene complementada por información de carácter histórico [«Por otra parte, *godo* toma ocasionalmente el valor de 'hombre orgulloso' o análogo (...). De ahí que en germanía *godo, godizo* y *godeño* significaran «rico o principal», según el vocabulario de Juan Hidalgo (1609)...» (*godo*)]. Es preciso destacar que el *DCECH* también marca cuándo una acepción concreta de una palabra pertenece a la germanía (*durindaina*, en su 2ª ac.).

Para referirse a la información relativa al habla jergal, el *DCECH* establece una gradación entre términos jergales (*entruchar*), semi-jergales (*gamba*) y casi jergales [«Hoy es voz bastante usada, pero marcadamente familiar, casi jergal, en el sentido de 'mucha gana de comer', 'vivo apetito', 'hambre'» (*gazuza*)]. La riqueza del diccionario se pone de manifiesto cuando se observa que las referencias a voces jergales se extienden también a voces no castellanas. Así se expresa en el artículo de *entruchar* donde se lee «palabra jergal, común al castellano y al port. *entrujar* (...); el caso es que hoy *entruchar* o *entruchilar* es simplemente 'engañar' en Salamanca, y el port. *intrujar, intrugir*, término jergal o meramente popular, es "burlar; lograr; desfrutar com astúcia"».

El *DCECH* distingue, además, entre los registros lingüísticos culto, coloquial, popular [*garrotillo* (*difteria*)] y familiar (*desmirriado*). Esta marca pue-

de completarse con otras indicaciones adicionales: en ocasiones, por ejemplo, se restringe el alcance diastrático del cultismo a la lengua escrita o a la lengua oral como en *dolo* que es «voz culta, curialesca» (*dolo*) y «hoy pertenece aún al vocabulario oral de la gente educada» (*díscolo*)], o en *esculpir* «frecuente desde los clásicos (*Aut.*)», donde se observa la vigencia de este uso desde la época clásica.

Dentro del registro coloquial, caben matizaciones variopintas que van desde una descripción de la situación actual hasta comentarios personales de los autores. Así se manifiesta en *grima*: «En la actualidad *grima*, y especialmente *dar grima*, pertenece al lenguaje coloquial más que al idioma literario, pero es bastante vivo y lo he oído a gente ciudadana de varios puntos de España».

En la especificación y caracterización de los usos lingüísticos primordiales de cada uno de los lemas, se pone de relieve una gran diversidad de contenidos que se desprenden de las diferentes redacciones en las que interviene el término *uso* en el diccionario [«uso muy popular y general» (*garrido*), frente a «debido a su uso cotidiano y humilde *ganso* sabía a labriego» (*ganso*)].

Entre las múltiples informaciones que aparecen en cada uno de los artículos del *DCECH*, cabe destacar además las distintas apreciaciones que se pueden hallar acerca del sector léxico al que pertenece un lema concreto. Este tipo de comentarios se introducen en el diccionario de forma diversa, generalmente a partir de la mención específica de una ciencia o rama del saber. Así ocurre en «voz de botánicos» (*dorso*); «medicina» (*embroca*); «término náutico» (*garete*); y «término culinario latino» (*gajorro*). En otras ocasiones el *DCECH* pone de relieve cómo distintas voces han sido introducidas en la lengua como términos propios de un campo léxico determinado. Por ejemplo leemos en *galón I* «entró en español como término militar y de modas»; en *cripta*, «tomado por vía eclesiástica» (*gruta*); en *declive*, «el vocablo entró como parte de la terminología militar y de fortificaciones»; y en *esguazar* «era italianismo propio de los militares».

La informatización de los datos que recogen la variación diastrática se podría constituir en el inicio de una investigación sobre este aspecto y en el punto de partida de multitud de léxicos sectoriales y de especialidad.

(b) Obtención de informaciones a partir de la relación de los datos

Las posibilidades de obtención de informaciones diversas a partir de la relación de los datos informatizados son también muy importantes.

La gramática histórica y el diccionario etimológico son complementarios a la hora de sistematizar los cambios fonéticos sufridos por una lengua. Mientras que la primera se ocupa sobre todo de los cambios regulares, el segundo se detiene en las particularidades evolutivas que se desvían de lo regular. La gramática histórica no se ocupa del cambio esporádico de *l* a *d* sufrido por *dejar*. Nuestro banco de datos permite identificar este cambio

en una palabra como *dejar* (< *lexar*, ant., forma general hasta h. 1200 < *laxare*, lat.) y también en *dintel* (< *lintel*, fr. medio < *limitalis*, lat. vg. < *liminaris* lat.), que serían los dos únicos lemas de la letra D del diccionario en que el español habría hecho un cambio esporádico de l- > d-, recibiendo ambos explicaciones un tanto dispares y diversas (se apela a la fonosintaxis) que, de buen seguro, merecen un estudio mucho más pormenorizado.[2] Si se establece una relación con otras palabras del diccionario a través de este tipo de evolución, es posible incluir dentro de este mismo grupo formas como *devantar*, procedente de *levantar*, por lo que respecta a la posición inicial de palabra; o étimos como *cadavera* (s.v. *calavera*)[3] o *melezin*[4] (s.v. *médico*), por lo que respecta a la posición interior de palabra.[5]

La conexión de estos cinco vocablos, al presentar un mismo tipo de problema fonético en su evolución, permite poner en duda la caracterización particular y esporádica del cambio (no se trata de casos tan aislados, como se ha intentado explicar), y se constata la existencia de una relación evolutiva biunívoca entre la oclusiva dental sonora y la lateral alveolar,[6] tanto en posición inicial ([l- > d-] *l*intel > *d*intel; *l*exar > *d*ejar; *l*evantar > *d*evantar) como en posición interior ([-l- > -d-] ca*l*avera > ca*d*avera; [-d- > -l-] me*d*icina > me*l*ezina).

Otra clase de problema etimológico que el banco de datos permite estudiar con mayor detenimiento es el referido a las etimologías inciertas, de las cuales encontramos una buena muestra en la letra D, por ejemplo. Los étimos *dado* (del árabe o persa), *daga I* (de origen desconocido), *dardabasí* (de origen desconocido), *daza* (del árabe), *debó* (de origen incierto), *dengue* (voz de creación expresiva), *desmirriado* (del portugués), *despotricar*, *destartalado* (del árabe), *destebrechar* (del oc. septentrional), *dibujar* (del francés antiguo), *disfrazar* (del latín vulgar), *divieso* (del latín), *dolama* (del árabe), *droga* (del celta), *duerna* (del céltico), *dujo*, tendrían en común la imposibilidad (en mayor o menor grado) de explicar con total certeza su origen,

---

[2] El *DCECH* justifica la evolución de *lexar* > *dexar* a partir de la anticipación de la *d* de la preposición *de* en la construcción *dejar de hacer algo* (s.v. *dejar*); y la de *lintel* > *dintel* como un proceso de disimilación de las dos *l* en la construcción con artículo *el lintel*, al igual que sucede en *bulla* > *bulda* (s.v. *dintel*).

[3] Este cambio fonético se explica como el resultado de la confusión en el habla vulgar entre *calvaria* y la forma culta *cadāuer*, palabras de significado próximo (vid. *DCECH*, s.v. *calavera*).

[4] Vid. Y. Malkiel, «Etimología y cambio fonético débil: Trayectoria iberorrománica de MEDICUS, MEDICĀMEN, MEDICĪNA», *Ibérida*, 6 (1961), pp. 127-171 (especialmente las págs. 151-171). Y. Malkiel justifica la variante *melezina*, forma patrimonial predominante entre 1250 y 1500, por una amalgama léxica con *miel*, cruce léxico que introduce una *-l-* antihiática tras la síncopa de la *-d-* latina.

[5] Se podría incluir también el caso de *cauda* > *coda*, esp. ant. > *cola*, explicado por influjo del vocablo *culo* (vid. Lloyd, P. M., *Del latín al español. I: Fonología y morfología históricas de la lengua española*, Madrid, Gredos, 1993, pp. 374-379).

[6] Vid. R. Menéndez Pidal, *Manual de gramática histórica española*, 6.ª ed., Madrid, Espasa-Calpe, 1940, §725b. Este autor considera la evolución *l* > *d* como un caso de error lingüístico (cambio fonético esporádico) que se produce por la equivalencia acústica entre la dental [d] y las líquidas [l] y [r].

considerado por los autores como incierto o desconocido. Además, otros vocablos, como *derribar, derrochar, desbarajustar, desbrevarse, desfalcar, desga, desgaire, desmazalado, desvaído, dije, dislate,* tampoco presentan una total seguridad sobre su origen, aunque se les añade un carácter de mayor probabilidad.

El hecho de poder agrupar todas estas palabras que presentan problemas etimológicos representa ya una gran ayuda para el investigador, pues permite establecer posibles nexos comunes entre ellas, y confiere una visión mucho más amplia y general del fenómeno. De este modo, una clasificación y estudio de estos étimos a tenor de las lenguas de las que proceden puede ayudar a clarificar, si no su origen, sí la veracidad o acierto de los argumentos etimológicos aportados por los autores del diccionario (caso de *dibujar, droga,* etc., por poner algunos ejemplos en cuanto a argumentación etimológica se refiere).

Una de las circunstancias que permite al lingüista una mayor precisión a la hora de inclinarse por una u otra hipótesis etimológica viene dada por lo que, en términos generales, podemos definir como valores diatópicos: la comparación entre las lenguas o entre los dialectos de una misma lengua histórica, los distintos resultados, los diferentes ámbitos de uso de una variante, etc., contribuyen todos ellos, en ocasiones de forma decisiva, a relegar determinadas hipótesis en beneficio de otra más sólida.

En este sentido, como no podía ser menos, el *DCECH* hace uso de una riquísima información geográfica que le lleva a ser el diccionario más completo en información dialectal en el ámbito *castellano* e *hispánico,* incluido el propio *DRAE*: aunque siempre supeditada a la explicación etimológica, la información diatópica es mucho más frecuente y precisa en el *DCECH* que en el *DRAE*[7].

El *DCECH* se convierte así en el mayor banco de datos de información dialectal pormenorizada para miles de entradas de los romances hispánicos. Pero desgraciadamente se trata, como para el resto de sus contenidos, de un banco de datos accesible exclusivamente por la entrada léxica correspondiente y no, por ejemplo, agrupados y ordenados por el lugar en el que se utiliza una variante, por la zona por la que se extienden los resultados de un étimo o, en fin, por el lugar en el que se registra la primera documentación de una voz.

El objetivo consiste, por tanto, en sistematizar todo ese cúmulo de información dialectal que se maneja en el *DCECH* de tal forma que pueda ser reducida a los parámetros propios de una base de datos, para poder luego reutilizarla según los criterios que, en cada momento, se considere más oportuno. El camino no es, sin embargo, fácil y los mayores inconvenientes vienen dados por el complejo entramado bajo el que esa información se presenta en el propio *DCECH.* Se trata de una información que carece de homogeneidad, que no es sistemática —se utiliza sólo cuando es

[7] Compárense, por ejemplo, las entradas *masera* (s.v. *masa*), *magüeto, marón,...* en uno y otro diccionario.

útil—, que, en fin, se presenta de modo muy distinto en cada una de las entradas, todo lo cual la hace difícilmente sistematizable.

En efecto, al lado de multitud de entradas en las que no aparece ninguna indicación diatópica, hay otras muchas en las que se nos ofrecen variadas precisiones geográficas: así, por ejemplo, son usuales en el *DCECH* indicaciones como las siguientes: voz usada en Santander (*mayueta*), en la zona leonesa (*masera*), con resultados en castellano y portugués (*marrano*), en iberorromance o en los tres romances hispánicos (*matizar, matar*) o en los romances de Occidente (*meaja, mazapán*). En estos casos, pese a todo, se manejan unos conceptos que no es difícil pasar a una base de datos. En otros, es preciso, no obstante, soslayar una serie de inconvenientes terminológicos que, en caso de no hacerlo, restarían efectividad a la base de datos. Es el caso de las denominaciones ambiguas: *gallego* frente a *portugués* en la mayoría de las ocasiones pero *gallego-portugués* en otras. Algo similar ocurre con la atomización de los datos procedentes de áreas dialectales: los abundantes Maragatería, Bierzo, Zamora, Asturias, etc. han de tener también —sin abandonar el dato pormenorizado— una referencia común del tipo de *leonés* en aras de esa operatividad a la que se aludía arriba.

Pero no se agotan aquí las referencias diatópicas en el *DCECH*. Fuera de esas informaciones relativas al ámbito de uso de una voz, son muy frecuentes también las referencias geográficas al tratar de la primera forma documentada de una palabra o de la documentación antigua, en general, de un vocablo. Así ocurre, por ejemplo, con las abundantes alusiones a la documentación medieval leonesa (*mesta, merino, masera, melena, miera...*) que, con frecuencia, sirve de referencia inicial en romance.

El inevitable cruce entre diacronía y diatopía lleva incluso a situaciones más complejas: de *mayo*, por ejemplo, se nos da una variante dialectal, el leonés *mao*, que sólo por el contexto ha de entenderse como una variante antigua y no actual; de modo similar, lo que hoy es una palabra de ámbito muy restringido (*mayueta*), se nos presenta en el pasado como voz usual en todo el castellano. Las indicaciones dialectales afectan a variantes formales, a significados específicos, pero también lo hacen a aspectos morfológicos: el género masculino (*el miel*) caracteriza al gallego, al portugués y, con ellos, al leonés occidental, frente al femenino *la miel* propio del castellano.

Todos ellos son detalles anotados por el *DCECH*, pero a los que difícilmente se puede recurrir más allá de los límites de la entrada para la que se citan. Una sistematización de los mismos que nos permitiera contar con amplias series de datos ordenados con criterios diversos haría posible una utilización más completa de todo ese material.

Las posibles aplicaciones de una base de datos en la que se hayan recogido todos estos extremos del *DCECH*, son enormes. Sin abandonar la propia discusión etimológica, podremos comprobar la correspondencia, o no, del área antigua de una serie de voces frente a su expansión actual, o bien, identificar las voces que están en retroceso y comprobar si tienen algo en común; será posible establecer una relación entre las explicaciones sustratísticas y el ámbito geográfico de un vocablo o grupo de vocablos; dibujar

las áreas léxicas más significativas de la Península; en fin, podremos fijar bases para iniciar nuevas investigaciones: imaginemos una consulta en la que aislemos los vocablos de etimología dudosa, con correspondencia sólo en castellano y portugués y, además, escasamente representados —según los datos del *DCECH*— en el área leonesa. Esta relación nos permitiría concentrar nuestros esfuerzos en identificar las posibles variantes leonesas de dichas voces —variantes no manejadas por J. Corominas— y que, sin embargo, es posible que en un buen número de casos guarden la clave para una mejor explicación de la evolución histórica de las voces normativas castellana y portuguesa.