Lexicografía e informática. Aplicación a la lengua catalana

Joaquim Rafel i Fontanals

Universitat de Barcelona
Institut d'Estudis Catalans

Si tuviera que resumir muy brevemente las características de lo que llamamos *lexicografía moderna* frente a la lexicografía tradicional, me referiría básicamente a dos aspectos: 1) la asunción progresiva de los avances que se han producido en la lingüística en los últimos decenios, y 2) la incorporación de las innovaciones tecnológicas que han tenido lugar en estos mismos años por lo que respecta al tratamiento masivo de la información, gracias a los progresos extraordinarios de la informática.

El interés de los estudiosos por la nueva tecnología y concretamente por la aplicación de los ordenadores a la lexicografía se remonta a los años 60 del siglo pasado. A partir de este momento, se realizan trabajos pioneros de naturaleza diversa. Sin poder entrar en los detalles de la evolución que se ha producido desde entonces, podemos decir que hoy en día es inconcebible un proyecto lexicográfico desligado del tratamiento informático.

Hay que decir, sin embargo, que la utilización de la informática en lexicografía puede tener un carácter más superficial, o más nuclear. El ordenador puede ser utilizado meramente como un instrumento que facilita la labor material de confección o de edición del texto o puede ser integrado de una manera plena en el trabajo lexicográfico propiamente dicho. No voy a decir nada sobre la primera opción, porque carece de interés desde el punto de vista de la lexicografía; en cambio, la incorporación plena de la informática en el trabajo propiamente lexicográfico, transforma completamente los métodos tradicionales de esta actividad. Por otra parte, la informática puede tener una importancia mayor o menor, según el tipo de proyecto que se quiere llevar a término. Francis E. Knowles (1990) distingue desde este punto de vista tres situaciones bien diferenciadas: 1) La situación banal que supone llevar a cabo una edición ampliada de un diccionario existente. 2) La realización de un diccionario con un nuevo diseño, pero partiendo de material lexicográfico

preexistente, compilado para diccionarios anteriores. 3) La elaboración de un diccionario partiendo de nuevos planteamientos lexicográficos. Ni que decir tiene que esta tercera opción es la más interesante desde el punto de vista metodológico y la que explota en mayor manera las posibilidades que ofrecen las técnicas y los recursos informáticos en cualquiera de las fases que podemos distinguir en el quehacer lexicográfico, es decir, en la compilación y preparación de los datos (fase prelexicográfica), en la elaboración de la obra propiamente dicha (fase lexicográfica), y en los métodos de divulgación o difusión de los resultados (fase postlexicográfica).

Por lo que respecta a la primera fase, la posibilidad de crear grandes repertorios de textos y de utilizarlos como base empírica para la construcción de nuevos diccionarios establece una frontera decisiva entre los procedimientos manuales de recopilación selectiva de la documentación para la elaboración del diccionario y la llamada lexicografía basada en corpus. A partir de estos planteamientos se establece una clara frontera entre esta nueva modalidad, que ha ido ganando terreno en los últimos años, y la lexicografía no basada en corpus, que representa una etapa cada vez más superada en el método lexicográfico. Las técnicas lexicográficas fundadas en criterios científicos rigurosos y objetivos se basan actualmente en planteamientos de esta naturaleza, después que los responsables de la elaboración de diccionarios han ido tomando poco a poco conciencia de la falta de solidez y de rigor de la práctica tradicional, que parte generalmente de lo que dicen otros diccionarios y del conocimiento que el lexicógrafo posee de la propia lengua, y solo en algunos casos utiliza vaciados selectivos de textos hechos manualmente. Los procedimientos informáticos permiten, en cambio, constituir corpus textuales de gran extensión, organizarlos internamente y extraer la información deseada para la descripción lexicográfica.

Con la aplicación de la tecnología informática, la fase de elaboración del diccionario propiamente dicha (la fase diccionarística, si utilizamos la terminología divulgada por Bernard Quemada) experimenta una transformación radical en relación con los procedimientos tradicionales. Esta transformación afecta fundamentalmente a dos aspectos: la redacción o composición del diccionario propiamente dicha y el acceso a la información necesaria para llevar a cabo la labor lexicográfica. En un sistema de trabajo plenamente informatizado, el esfuerzo del lexicógrafo puede concentrarse en el núcleo de la labor plenamente lexicográfica, es decir, en la determinación del significado de las unidades léxicas, en la redacción de las definiciones y en las otras cuestiones que cada proyecto de diccionario prevea (información de

carácter sintáctico, ejemplificación, coocurrencias más frecuentes, etc.); el redactor se despreocupa, por tanto, de cualquier cuestión relacionada con la disposición física de los distintos elementos en las posibles ediciones de la obra, puesto que trabaja directamente sobre los campos de una base de datos, que corresponden a los distintos elementos de la estructura del diccionario. Esta disposición en forma de base de datos permite además llevar a cabo un control exhaustivo de los más diversos aspectos del diccionario, lo cual asegura la máxima coherencia sistemática, cosa muy difícil de conseguir, por no decir imposible, en los diccionarios redactados de manera secuencial.

Los procedimientos informáticos representan también una revolución por lo que respecta al acceso a las distintas fuentes que proporcionan la información necesaria para que el redactor del diccionario pueda llevar a cabo cómodamente su labor. Así el redactor en su estación de trabajo puede disponer no solamente de la interficie de redacción que le permite introducir los datos correspondientes a los distintos elementos estructurales del diccionario que va a crear, sino también del acceso (telemático o local) a todas aquellas fuentes de información necesarias para su labor.

La informática tiene también un papel decisivo en la fase postlexicográfica, es decir, en todo lo que se refiere a la divulgación de la obra. Actualmente, la difusión en forma de libro impreso es solo una posibilidad al lado de otras formas de difusión electrónica.

Con estas palabras he pretendido esbozar lo que podríamos llamar el marco general de la zona de intersección entre la lexicografía y la tecnología informática. La parte específica de mi intervención se refiere, sin embargo, al modo como este marco ha condicionado las realizaciones concretas de la lexicografía catalana, específicamente en el ámbito institucional académico, es decir, en el ámbito del Institut d'Estudis Catalans. Para ello hemos de remontarnos al comienzo de los años 80 del siglo pasado, cuando la Sección Filológica del Institut d'Estudis Catalans se plantea la forma más adecuada de organizar su actividad lexicográfica después de un dilatado período de interrupción por circunstancias de orden político. En aquel momento se acuerda la realización de un proyecto lexicográfico que responda a los principios propugnados por la lexicografía moderna en los años precedentes; el proyecto, denominado *Diccionari del català contemporani* (DCC) pretende asumir los avances consolidados de la lingüística y de la lexicografía y aplicar las posibilidades de la tecnología moderna en el tratamiento de la información con la utilización de ordenadores como estrategia de trabajo.

Asimismo la Sección Filológica se plantea cómo acometer las nuevas ediciones o actualizaciones del diccionario que tenía en aquel momento todavía carácter normativo, pero que estaba claramente desfasado, el *Diccionari general de la llengua catalana* de Pompeu Fabra. En esta línea decide elaborar un nuevo diccionario normativo sobre la base del mismo de Fabra, pero revisado y, sobre todo, actualizado. El resultado es el *Diccionari de la llengua catalana*, publicado en 1995, y una segunda edición nuevamente revisada y ampliada, publicada el pasado año 2007. No voy a entretenerme en el comentario de esta acción, porque desde el punto de vista de su realización responde al primero de los tipos enunciados por Knowles en el trabajo que les he citado hace un momento. Es decir, el uso de la informática en su realización tiene un carácter meramente auxiliar, puesto que se trata de una edición ampliada de un diccionario existente, aunque revisada y remozada.

Sí, en cambio voy a darles algún detalle del DCC, que responde al tercero de los tipos enunciados por Knowles, es decir, la elaboración de un diccionario partiendo de nuevos planteamientos lexicográficos. Con el nombre de DCC designamos un proyecto complejo, que, como ya he avanzado, responde al deseo de la Sección Filológica del IEC en el sentido de que los trabajos lexicográficos que había de emprender en un nuevo período de su historia no fueran insensibles a los avances científicos, metodológicos y tecnológicos que habían tenido lugar durante el largo tiempo de inactividad forzada.

El proyecto DCC se estructura en dos fases:

1. Creación de recursos lingüísticos:
   a) *Corpus textual informatitzat de la llengua catalana* (CTILC)
   b) *Base de dades lexicogràfica* (BDLex)
2. Descripción lexicográfica:

   Diccionari descriptiu de la llengua catalana (DDLC)

Primera fase: Creación de recursos lingüísticos

Características principales del CTILC

Desde el punto de vista cronológico, el CTILC se extiende desde 1832 hasta 1988; abarca, pues, textos de más de 150 años de la historia de la lengua catalana escrita entre los siglos XIX y XX. La fecha de origen viene determinada

por el inicio de la recuperación del uso literario de la lengua en la época contemporánea y la fecha final por el momento en que se ultimó la selección de los textos que iban a formar parte del corpus.

Desde el punto de vista tipológico, el CTILC incluye textos de carácter literario y textos de carácter no literario. Cada uno de estos dos tipos ha sido dividido en otros, que han permitido una selección equilibrada de los textos a tener en cuenta no solo para cada tipo de lengua (literaria o no literaria), sino para cada subtipo (los cuatro géneros tradicionales: Narrativa, Poesía, Teatro y Ensayo, para la lengua literaria, y diez grupos temáticos o funcionales distintos para la lengua no literaria: Filosofía, Religión y Teología, Ciencias Sociales, Prensa, Ciencias Puras y Naturales, Ciencias Aplicadas, Bellas Artes, Ocio y Deportes, Lengua y Literatura, Historia y Geografía y Correspondencia). Esta división tipológica da lugar a una jerarquía de componentes del corpus que es necesaria para proporcionar una información suficientemente representativa de la lengua, como ha sido reconocido en estudios posteriores sobre la naturaleza de los corpus de referencia.

La extensión total del corpus es de 52.371.944 ocurrencias o palabras del texto, que se reparten en 23.105.591 (44%) correspondientes a la lengua literaria y 29.266.353 (56%) correspondientes a la lengua no literaria. El número de obras o textos distintos, de extensión muy diversa, que corresponden a este volumen de texto es de 3.299, de las cuales 1.011 corresponden a textos literarios y 2.288 a textos no literarios.

Una de las preocupaciones principales a la hora de concebir el CTILC fue la máxima representatividad, es decir, que el conjunto de textos tomados en consideración reflejara de la mejor manera posible la lengua escrita utilizada en el período de tiempo que abarca; para conseguir este objetivo, en el momento de la selección se puso especial empeño en lograr el máximo equilibrio entre los distintos tipos de texto que las necesidades de la comunicación escrita —estética o funcional— habían producido a lo largo del ámbito cronológico del CTILC. Para ello, además de la división tipológica a que me he referido hace un momento, se establecieron también unos grupos cronológicos que tienen una extensión de diez años en la parte más antigua (hasta 1913) y de cinco años a partir de 1914. Con estas medidas se intentó lograr el máximo equilibrio entre las más variadas manifestaciones de la lengua escrita.

Una característica destacada del CTILC es que se trata de un corpus totalmente lematizado; en él se ha llevado a cabo una operación de análisis lingüístico en virtud de la cual se ha categorizado gramaticalmente cada una de las ocurren-

cias de cada forma gráfica y se ha asociado a una unidad léxica de referencia; con ello se cubren dos objetivos: por una parte se desambiguan formas gráficas que corresponden a formas gramaticales distintas (de una misma serie flexiva o de series flexivas correspondientes a lemas distintos) y por otra parte, como consecuencia de la misma operación, se relacionan entre ellas las distintas formas de una serie flexiva, las cuales quedan asociadas a un mismo lema. Como es sabido, en una buena parte de los corpus más o menos extensos la unidad de trabajo es la palabra entendida como cadena de caracteres gráficos, con lo cual estas unidades tienen un elevado grado de ambigüedad, porque una misma forma gráfica puede corresponder a diferentes unidades gramaticales o léxicas; la información que contiene un corpus sin lematizar sólo es accesible a través de las formas gráficas con todos los inconvenientes que ello comporta para la mayoría de consultas que se quieran realizar. El CTILC, por el hecho de estar completamente lematizado, tiene un elevado nivel de funcionalidad; realizando una consulta a partir de un lema obtenemos información sobre todas las formas de flexión correspondientes, con las posibles variantes gráficas o formales que aparecen en textos de diferentes épocas, y también a las eventuales formas derivadas mediante afijos apreciativos o intensivos, para las cuales, siguiendo criterios lingüísticos, no se ha creado un lema propio, sino que se han asociado al lema correspondiente a su base léxica. El número de lemas a que ha dado lugar esta operación es de 149.185, que corresponden a 678.386 formas gramaticales y a 51.253.680 ocurrencias del texto (el total de ocurrencias del corpus menos los nombres propios).

El CTILC fue constituido entre los años 1985 y 1997; podemos considerarlo, pues, un trabajo pionero en nuestras latitudes. Una vez terminado, dispone de un sistema de consulta y explotación concebido especialmente para su utilización con finalidades lexicográficas, pero que es también útil para cualquier tipo de estudio que se pretenda elaborar a partir de datos empíricos. En este momento el CTILC se puede consultar públicamente por Internet a través de la Web del IEC; el procedimiento de consulta permite obtener una serie de contextos relativos a un lema determinado. Como actividad vinculada a la elaboración del CTILC, fue publicado entre los años 1996 y 1998 un diccionario de frecuencias que incluye la totalidad de los datos léxicos y estadísticos del corpus (Rafel, dir. 1996-1998).

Características principales de la BDLex

La Base de Dades Lexicogràfica (BDLex) contiene debidamente informatizados y cargados en una base de datos los 13 diccionarios catalanes que han

sido considerados más significativos de los siglos XIX y XX; fue creada con el fin de facilitar el acceso rápido y sistemático a la información que contienen estas obras y obtener el máximo rendimiento de las consultas. Fue concebida, como ya he apuntado, como un recurso complementario dentro del proyecto general, y más adelante concretaré cual es la utilización específica que se le da en la elaboración del diccionario descriptivo. Fue constituida inicialmente, entre 1995 y 1997, con 11 diccionarios, y hasta los años 2001-2002 no ha podido ser completada con los dos restantes, uno de los cuales, el *Diccionari Català-Valencià-Balear*, de A. M. Alcover y F. de B. Moll, es de una gran extensión y de una extraordinaria complejidad estructural. Los diccionarios de que consta actualmente son los siguientes:

Corpus lexicográfico de la BDLex

DMFC Febrer i Cardona, A. "Diccionári menorquí espanól francês y llatí". [manuscrito, principios siglo XIX].

DEBJ Esteve, J.; Bellvitges, J.; Juglà, A. *Diccionario catalan-castellano-latino*. 1803-1805.

DLCL Labèrnia, P. *Diccionari de la llengua catalana: ab la correspondencia castellana y llatina*. 1839-1840.

DMCF Figuera, P. A. *Diccionari mallorqui-castella*. 1840.

DVCE Escrig, J. *Diccionario valenciano-castellano*. 1851.

NDMA Amengual, J. J. *Nuevo diccionario mallorquin-castellano-latin*. 1858.

DMCT [Tarongí i Cortès, J.]. *Diccionari mallorquí-castellá.* 1878. [Inacabado]

DGMG Martí i Gadea, J. *Novísimo diccionario general valenciano-castellano*. 1891.

DCVB Alcover, A. i Moll, F. de B. *Diccionari Català-Valencià-Balear*, 1926-1962.

DPCV Vallès, E. *Pal·las: diccionari català-castellà-francès: amb vocabularis castellà-català francès-català*. [1927].

DGLC Fabra, P. *Diccionari general de la llengua catalana*. Barcelona, 1932.

DGFP Ferrer Pastor, F. *Diccionari general,* 1985.

DIEC Institut d'Estudis Catalans. *Diccionari de la llengua catalana*, 1995.

La información que contienen estos diccionarios fue trasladada a soporte informático y el texto fue tratado de modo que permite la reproducción de los originales siguiendo los criterios tipográficos de cada uno. Los diferentes elementos que configuran la estructura de los diversos diccionarios han sido identificados sistemáticamente y codificados de manera adecuada de tal manera que una vez incorporados a la base de datos pueden ser objeto de consultas orientadas selectivamente y pueden ser relacionados entre ellos.

Segunda fase: Elaboración de un diccionario descriptivo

El objetivo de la segunda fase del proyecto DCC, que se encuentra en proceso de realización, es la elaboración de un diccionario descriptivo de la lengua catalana contemporánea a partir, principalmente, del análisis y de la explotación de CTILC. Entendemos aquí por *diccionario descriptivo* aquella obra lexicográfica que tiene por objeto la definición de las unidades léxicas de la lengua desde el punto de vista de su contenido y de su utilización real, sin restricciones basadas en criterios prescriptivos.

Una de las justificaciones de la elaboración de un diccionario descriptivo de estas características por parte de una academia de la lengua es la creencia de que las prescripciones lingüísticas estarán tanto mejor fundamentadas cuanto mejor conocida sea la lengua en su realidad fáctica; en el caso del IEC la justificación es doble por cuanto entre sus misiones estatutarias, como hemos visto, tiene no solo el establecimiento de la normativa lingüística, sino también "ocuparse del estudio de la lengua". Con la elaboración de un diccionario descriptivo de estas características el IEC no solo produce una obra concebida de acuerdo con los principios más generalmente asumidos por la lexicografía contemporánea, sino que se dota de un instrumento muy valioso a la hora de ejercer su actividad como institución académica encargada del establecimiento y de la actualización de la normativa de la lengua catalana.

Características principales del DDLC

Como sea que el término *descriptivo* ha sido utilizado en lexicografía con sentidos y valores diversos, que ahora no es el momento de glosar, conviene que precisemos que en el caso del DDLC se utiliza en un sentido estricto, un sentido que lo acerca al concepto de *diccionario de lengua puro* o *diccionario lingüístico* que propuso Dirk Geeraerts (1985) a fin de distinguirlo del llama-

do corrientemente *diccionario de lengua*; desde este punto de vista el *diccionario lingüístico* tiene una intención preferentemente reflexiva y un objetivo funcional de carácter científico, mientras que el diccionario de lengua tiene una intención comunicativa y hermenéutica y un objetivo funcional de carácter pragmático; los representantes conspicuos de los diccionarios propiamente lingüísticos son los llamados *diccionarios teóricos*, que pretenden la descripción de una lengua mediante la aplicación rigurosa de una teoría lingüística determinada.

El DDLC, sin pretender ser un diccionario teórico, comparte con estos alguna de sus características: no tiene un carácter pragmático ni una finalidad pedagógica, es concebido como una investigación *ex novo*, su realización persigue el máximo rigor científico, su formulación aspira a un alto grado de explicitud y utiliza una cierta formalización en su lenguaje y en su presentación. Como consecuencia de ello, sus usuarios ideales son los profesionales de la lengua; sin embargo, esta obra lexicográfica no pretende dirigirse solo a especialistas, sino que, además de ser útil para estos, aspira a estar al alcance de cualquier lector medianamente culto interesado por los problemas de la lengua en cuanto que usuario: se pretende, pues, presentar el contenido del diccionario combinando el rigor en el tratamiento de la información con la claridad expositiva y con la facilidad de interpretación.

Uno de los problemas que plantea la elaboración de un diccionario descriptivo por una academia de la lengua es que este diccionario contiene palabras o acepciones que no son reconocidas por la normativa vigente, a pesar de encontrarse documentadas en los textos; por una parte este diccionario puede ser considerado más científico que el normativo por cuanto intenta dar cuenta de una manera sistemática de la realidad de la lengua a partir de datos empíricos, pero por otra parte puede ser visto como un peligro para el uso lingüístico considerado correcto. En el caso que nos ocupa, este tema se debatió ampliamente en el seno de la Sección Filológica, la cual decidió que se identificara con una marca visible todos aquellos elementos (entrada, categoría, acepción, patrón sintáctico, etc.) que no gozaran de sanción normativa en el momento de divulgarlos.

En el aspecto material, tanto desde el punto de vista del proceso de redacción como en sus resultados finales, el proyecto de diccionario descriptivo pretende incorporar los últimos avances metodológicos y tecnológicos que se han producido en el campo de la lexicografía. Como consecuencia de ello tiene las características de un diccionario electrónico, en forma de base de datos multifuncional que permite utilizaciones diversas, tanto relacionadas

con la investigación como con la difusión de la obra; si nos centramos en la difusión, permite las más variadas posibilidades: por una parte puede ser divulgado por vía electrónica a través de Internet y por otra parte puede servir de base para diversidad de publicaciones, sea en papel, sea en cualquiera de los soportes electrónicos disponibles.

Una de las características más destacadas del DDLC, que conforma todo el diccionario, es el hecho de utilizar el CTILC como base fundamental y fuente exclusiva; se trata, pues, en sentido estricto, de un diccionario basado en corpus. No es este el momento ni el lugar de detallar las ventajas de este procedimiento; solo destacaré el hecho que de acuerdo con este principio metodológico, basado en datos empíricos, se establece no solo la nomenclatura del diccionario, sino los significados de las unidades léxicas a partir del uso que realmente se ha hecho de ellas, en vez de utilizar métodos apriorísticos (basados en los diccionarios preexistentes) o intuitivos (basados en la conciencia lingüística del lexicógrafo); la aplicación de este método permite la utilización de datos de carácter estadístico que pueden llegar a tener una gran importancia tanto en la fijación de la nomenclatura como en aspectos internos de la estructura del diccionario; permite también la ejemplificación a través de testimonios de uso refrendados por una cita y una referencia concretas, con lo cual podemos, si es necesario, situar en el tiempo un determinado uso o significado. Como consecuencia de la utilización de un corpus como punto de referencia para su elaboración, el DDLC incluye una información exhaustiva y sistemática sobre las estructuras sintácticas de que pueden formar parte, en sus distintas acepciones, las unidades léxicas tratadas y también sobre las coocurrencias más frecuentes.

Este método y este procedimiento, como a nadie se le oculta, tiene, sin embargo sus detractores, que defienden otros modos de proceder en la práctica lexicográfica, temiendo que un sistema como este desvirtúe los logros de la lexicografía tradicional. Justamente, saliendo en cierto modo al paso de este tipo de objeción, el DDLC incorpora todos los datos relativos al significado que aparecen en los principales diccionarios y que no se han verificado en el corpus; pero ello se hace fuera del cuerpo principal del artículo y dejando constancia de qué diccionario o qué diccionarios aportan la información correspondiente. No se entremezclan, pues, la información procedente del análisis del corpus, que determina el núcleo fundamental del artículo, con la que procede de los diccionarios y no tiene un reflejo en el corpus. Por lo que respecta a la nomenclatura, la del DDLC tiene un carácter complejo, en el sentido de que no todas las unidades léxicas que encabezan un artículo son descri-

tas en él, sino en otro artículo encabezado por un elemento distinto. Así, pues, la nomenclatura del DDLC está formada por unos elementos de rango general (caracterizados por una descripción lexicográfica propiamente dicha) y por otros elementos de rango subsidiario (caracterizados por remitir a otros artículos); estos elementos subsidiarios, controlados sistemáticamente, corresponden a dos tipos: a) los derivados formados a partir de alguno de los siete procesos de derivación sistemática establecidos en los criterios de redacción del diccionario, y b) las variantes formales documentadas en el corpus que están asociadas a una forma léxica principal, sin ser meras variantes gráficas de ella. Contiene también el diccionario un tercer tipo de elementos vinculados a unas entradas determinadas, que se encuentran integrados en los artículos correspondientes a ellas y no dan lugar a entradas propias; se trata de los derivados apreciativos, las variantes flexivas de la entrada no previstas en el modelo flexivo que la caracteriza y las conversiones sintácticas, que están vinculadas a las acepciones. Por último, las unidades léxicas plurinominales no constituyen tampoco entradas propias ni están vinculadas a acepciones o mezcladas con ellas, como ocurre en muchos diccionarios, sino que están agrupadas en una sección propia del artículo correspondiente a la entrada a la que se encuentran vinculadas de acuerdo con unos criterios explícitos.

El DDLC se organiza en artículos, que constituyen la unidad básica en que se estructura toda la información que contiene el diccionario. Los criterios para el establecimiento y la delimitación de los artículos se basan exclusivamente en principios de carácter formal y gramatical, y se aplican de una manera sistemática a lo largo de todo el diccionario sin ninguna excepción. Estos criterios consisten en la identidad o diferencia en cualquiera de las tres características que configuran la entrada: la grafía de la forma canónica, su categoría y sus propiedades flexivas. La diferencia en cualquiera de estas tres características determina la existencia de un artículo distinto; en caso contrario, toda la información se agrupa en un único artículo, independientemente de las diferencias de carácter meramente semántico o de consideraciones de carácter etimológico, que no intervienen como criterios diferenciadores de los artículos, contrariamente a lo que ocurre en la mayoría de los diccionarios tradicionales. En el DDLC no existen, pues, dos entradas que coincidan en su forma gráfica, en su categoría y en su modelo de flexión.

Otra de las características del DDLC que merece ser mencionada es el sistema de ordenación de las acepciones, que se basa exclusivamente en una combinación de criterios frecuenciales y lógicos. Resumidamente les diré que en un artículo que presenta varias acepciones puede ocurrir que estas tengan

una afinidad semántica estrecha, en cuyo caso se aplica el criterio frecuencial: la acepción que se muestra más frecuente en el corpus aparece en primer lugar y el resto se ordenan por orden decreciente de frecuencia; si, en cambio, pueden establecerse varios grupos a partir del criterio de afinidad de significado, aparece en primer lugar el grupo que contiene la acepción más frecuente y el resto de grupos por orden decreciente de la acepción más frecuente del grupo. El sistema de ordenación presenta, pues, una jerarquía de dos niveles.

De acuerdo con lo dicho más arriba sobre el tipo de diccionario a que aspiramos, las definiciones del DDLC evitan la información de carácter enciclopédico (basada en la descripción de la realidad) y se concentran en la información de carácter lingüístico (descripción del valor significativo, de las restricciones léxicas y de las propiedades sintácticas de las unidades). Desde el punto de vista del texto definidor, se tiene en cuenta la distinción entre los elementos propiamente definidores (constituyentes intrínsecos de la definición) y los que se refieren a condiciones o restricciones selectivas (constituyentes extrínsecos), codificándolos adecuadamente.

El DDLC contiene, para cada entrada, información de carácter cuantitativo relativa al uso observado en el corpus. Esta información aparece simplificada, con una representación gráfica, no numérica, que indica la pertenencia de la entrada a uno de los cinco rangos de uso establecidos. Para las entradas correspondientes a los tres primeros rangos de uso, el DDLC contiene, además, información sobre la repartición porcentual de sus categorías morfológicas en el corpus.

Contenido y estructura del DDLC

Como ya hemos comentado, el DDLC consta de una serie de elementos estructurales organizados en forma de base de datos; cada uno de ellos está vinculado a un elemento de referencia, que es la entrada, lo cual permite configurarlos en forma de artículo de diccionario, que es la manera más común de presentarlos ante los usuarios, pero no la única posible; por otra parte, esta es la forma más práctica de referirse organizadamente a los distintos elementos que conforman la obra, por lo cual, en este comentario sobre el contenido y la estructura del diccionario seguiremos la pauta organizativa del artículo estándar de DDLC.

El artículo así concebido consta de unos elementos que constituyen su cabecera (la *entrada*, la *categoría*, la *información flexiva*, la *información estadís-*

tica y el *perfil morfológico*) y de otros que constituyen su cuerpo (las *acepciones*, las *locuciones*, las *variantes*, los *derivados* y la *información complementaria*). Estos elementos tienen distintos grados de obligatoriedad y de dependencia unos de otros y, en consecuencia, puede haber artículos muy complejos —sobre todo si tenemos en cuenta la recurrencia potencial de algunos de sus componentes— y artículos muy simples. Hay, sin embargo, unos mínimos elementos obligatorios sin los cuales no es posible un artículo: la entrada, la categoría, la información flexiva (explícita o implícita) y la información estadística, por lo que se refiere a la cabecera, y como mínimo una acepción por lo que se refiere al cuerpo del artículo; esta acepción puede aparecer en su modalidad de *descripción*, que conlleva una serie de elementos asociados, o en su modalidad más simple de *remisión*. Conviene quizá aclarar que en el DDLC se entiende por remisión una referencia explícita iniciada por la indicación *Vegeu* ("Véase"); conviene, pues, no confundir este tipo de referencias explícitas con las referencias implícitas que se dan habitualmente en forma de definiciones sinonímicas o sintéticas, que frecuentemente se denominan también remisiones; unas y otras forman parte de la estructura referencial del diccionario pero tienen un valor muy distinto: las primeras, las remisiones propiamente dichas, se encuentran en lugar del bloque *descripción*, mientras que las segundas constituyen el elemento *definición* dentro del bloque *descripción*; unas y otras responden a dos funciones completamente distintas. La existencia de remisiones en el DDLC —en este sentido estricto— es un mero reflejo del carácter complejo de la nomenclatura a que me he referido al tratar de las características generales del diccionario.

Puesto que ya he aludido a ellos al hablar de las características generales del DDLC, no voy a entretenerme ahora en el comentario de los elementos de la cabecera del artículo: la *entrada* (la representación gráfica de la forma canónica de la unidad léxica que se describe), la *categoría*, la *información flexiva*, la *información estadística* y el *perfil morfológico* para aquellas entradas de los rangos 1, 2, o 3.

Dentro del cuerpo del artículo, el bloque de *acepciones* tiene una cierta complejidad. Como hemos visto ya, una acepción puede consistir meramente en una remisión, en el sentido que hemos detallado, pero lo más habitual es que consista en lo que llamamos *descripción*. La descripción de una acepción consta de los siguientes elementos: el *patrón sintáctico*, las *restricciones*, la *definición*, las *colocaciones*, el *ejemplo* o los *ejemplos* y las *conversiones sintácticas*. Los patrones sintácticos se expresan por cadenas de símbolos categoriales (N, V, Adj, etc.), que pueden incorporar modificadores en forma de subíndices (N_{COMPT}, V_{PRON},

etc.); también en forma de subíndices se indica el carácter correferencial o no de dos o más elementos del patrón; cada símbolo del patrón puede ir introducido por una palabra literal (una preposición, un artículo, etc.). Por otra parte, determinadas correspondencias sintácticas regulares se describen de forma relacionada, como por ejemplo, una construcción intransitiva y la transitiva causativa correspondiente. Se indican también las restricciones léxicas o semánticas que existen sobre los elementos no nucleares del patrón. Ya hemos hecho referencia más arriba a la naturaleza de la definición; solamente voy a añadir aquí que los elementos extrínsecos de la definición se corresponden con determinadas posiciones del patrón sintáctico y se utilizan para ajustar semánticamente un descriptor de sentido más general que el elemento definido o bien para ajustar sintácticamente la definición al comportamiento del definido en relación con sus argumentos. Otro apartado específico de la descripción de una acepción consiste en las *colocaciones*, es decir, los grupos de unidades que presentan una coocurrencia frecuente a causa de una cierta atracción léxica entre ellas; como sea que el significado de las colocaciones es función del de sus componentes, no conllevan una descripción semántica específica. En cuanto a los *ejemplos*, solo recordaré que son citas extraídas del CTILC, que se reproducen sin ninguna adaptación ni modificación; para la selección del ejemplo más adecuado para ilustrar una determinada acepción, se tienen en cuenta una serie compleja de criterios que ahora no podemos detallar; solo mencionaré que el ejemplo debe contener la información necesaria y suficiente para ilustrar la acepción de que se trata y que no debe ser ambiguo desde ningún punto de vista. Si la acepción tiene más de un patrón, debe haber un ejemplo por cada patrón. Los ejemplos se identifican con una referencia simplificada que aporta una primera información sobre la procedencia de la cita; en la versión electrónica del DDLC se pueden desarrollar interactivamente los elementos codificados de esta referencia simplificada y se puede acceder a la referencia bibliográfica completa de la obra en cuestión. Finalmente, el último elemento de la descripción de una acepción son las *conversiones sintácticas*, que no comportan ningún tipo de descripción, pero llevan asociado un ejemplo. Los tipos de conversiones que se tratan en este apartado son: infinitivo → sustantivo, participio → adjetivo, interjección → sustantivo y sustantivo → interjección.

Otro gran apartado del artículo está constituido por las *locuciones*. Entendemos por locución cualquier combinación de elementos léxicos de significado no directamente deducible del de sus componentes; tenemos, pues, locuciones nominales, adjetivales, verbales, adverbiales, prepositivas, conjuntivas e interjectivas. Estas unidades se tratan como elementos subordinados a una entrada y, por tanto, se describen dentro del artículo correspondiente a su entrada de referen-

cia; cada locución, por otra parte, puede tener una o más acepciones y cada acepción consta de los mismos elementos estructurales subordinados que las acepciones de una entrada, excepto las conversiones sintácticas.

El apartado de *variantes* contiene tres tipos de elementos distintos: las variantes formales de la entrada que no son consideradas meras variantes gráficas, los derivados apreciativos o intensivos y las variantes de flexión no previstas en el modelo flexivo asociado a la entrada.

El apartado de *derivación* incluye todos los derivados —formados sobre la misma base que la entrada o sobre una base culta con cierta similitud formal y de sentido equivalente— cuyo sentido sea máximamente predictible, es decir, que sean el resultado de un proceso derivativo transparente a todos los niveles (semántico, morfosintáctico y fonológico). A fin de evitar decisiones subjetivas de los redactores se han establecido y definido claramente siete procesos derivativos especificando los afijos que pueden actualizarlos, el tipo de transformación que suponen, la categoría y el significado del derivado en relación con los de la base a que está asociado y la definición formularia que le corresponde. Para cada derivado que se incluye en este apartado se indica el número del proceso derivativo a que corresponde, tantos ejemplos como acepciones de la entrada se encuentran actualizadas por este derivado en el corpus, indicando en cada ejemplo el número de la acepción de la entrada que actualiza. Este apartado contiene también colocaciones y conversiones.

Finalmente, el artículo incorpora en un apartado específico, titulado *información complementaria*, toda la información léxica relativa a las acepciones y locuciones de una entrada que se encuentran en alguno de los diccionarios de la BDLex pero que no se han documentado en el CTILC; en cada caso se indica de qué diccionario procede la información y la entrada bajo la que se encuentra, en aquellos casos en que esta no coincide con la entrada del artículo de DDLC.

Para completar esta visión solo nos queda hacer referencia a la marca de *no normativo* y a la de *no documentado*. La marca de *no normativo*, a la que he aludido más arriba, indica aquella información descriptiva, del nivel que sea, no aceptada o simplemente no recogida por la normativa vigente; esta marca puede afectar a varios elementos de un artículo, y, por otra parte, se proyecta siempre sobre los elementos de rango inferior de cualquier estructura; así, pues, si está colocada delante de la entrada, afecta a todo el artículo, si está colocada delante de una acepción, afecta a todos los elementos de la acepción; en cambio, si está colocada delante de un patrón sintáctico, afecta solo a este patrón, y, si está colocada ante una variante afecta solo a esta variante.

La marca de *no documentado* tiene su justificación en el hecho de que en algunos casos el DDLC incorpora información que se ha creído necesaria por coherencia descriptiva, pero que no se halla en el CTILC; con ello se deja constancia explícita de determinados elementos (generalmente una categoría o un patrón sintáctico alternante) que no se encuentran en el CTILC, pero que podríamos esperar a partir de cierta información que sí se verifica en él. El uso de esta marca es muy poco frecuente, porque son pocos los casos en que se da esta circunstancia concreta.

Ejecución y estado actual del DDLC

Fase preparatoria

En el año 1998, una vez ultimados los trabajos de constitución de los recursos lingüísticos previstos, se llevan a cabo una serie de actividades preparatorias para la elaboración del diccionario descriptivo. Estas actividades consistieron básicamente en a) el establecimiento de la nomenclatura del diccionario a partir de la evaluación del lemario del corpus; b) el establecimiento de los criterios lexicográficos que se habrían de aplicar y de las normas de redacción de la obra; c) la redacción de una serie de 600 artículos prototípicos, que se utilizó como banco de pruebas de los criterios que se iban estableciendo y de los instrumentos informáticos que se diseñaban paralelamente para la ejecución del proyecto; y d) el establecimiento de una estación de trabajo lexicográfico que había de permitir a los redactores disponer de todos los elementos necesarios para el desempeño de su labor, debidamente integrados y accesibles desde sus terminales.

Redacción sistemática

Después de esta fase preparatoria, en 1999 se inicia la redacción sistemática de la obra, en la cual nos encontramos todavía. En relación con esta nueva fase, que es de una complejidad notable, mencionaré solo unos pocos aspectos que puede resultar interesante destacar

Uno de ellos hace referencia al proceso de redacción; la redacción del DDLC no avanza siguiendo estrictamente el orden alfabético, sino que, a partir de una progresión alfabética básica, conjuntamente con el artículo de la palabra correspondiente al orden alfabético, se redactan también los artículos relativos a otras unidades léxicas que están relacionadas con ella por su forma

o por una relación semántica o sistemática, sea cual sea la letra del alfabeto a que pertenecen; cada redactor elabora, pues, paralelamente, una serie más o menos larga de artículos cuyas entradas están relacionadas de acuerdo con estos criterios; con este procedimiento se intenta evitar en lo posible la falta de coherencia estructural y descriptiva que afecta, en algunos casos de manera grave, a la mayoría de diccionarios existentes.

En otro orden de cosas, una de las preocupaciones principales de la dirección del proyecto durante el tiempo que llevamos de redacción ha sido el control de la producción, tanto desde el punto de vista cuantitativo como desde el punto de vista cualitativo, que se lleva a cabo básicamente a través de reuniones semanales del coordinador del proyecto y del secretario de redacción conjuntamente, con cada uno de los redactores para seguir de cerca el trabajo que tienen entre manos en cada momento; por otra parte, el ritmo de producción de artículos es actualizado diariamente, lo cual permite hacer un seguimiento pormenorizado del estado de los trabajos del diccionario en general y también un seguimiento personalizado del trabajo de cada redactor.

Una acción importante en los trabajos de ejecución de la obra es lo que llamamos *validación estructural*. Se trata de una operación que se realiza sobre los artículos ya redactados y tiene como objeto el control sistemático de la aplicación rigurosa de los criterios de redacción y de la coherencia estructural de la obra. No puedo entrar aquí en detalles sobre la aplicación de este procedimiento; solo diré que un protocolo de actuación describe las acciones o comprobaciones específicas que se llevan a cabo sobre cada uno de los aspectos de los distintos componentes de los artículos.

Estado actual

La extensión de la parte redactada del diccionario varía de día en día, porque se van incorporando los artículos a medida que se da por terminado su proceso de redacción. El volumen actual (en datos de 6 de octubre de 2008) es de 47.652 artículos;[1] por lo que respecta a la distribución alfabética, la tercera parte de este volumen corresponde a la parte redactada según la progresión alfabética rigurosa, mientras que las otras dos terceras partes correspon-

[1] En el momento de revisar las pruebas de este trabajo para la publicación (11 de abril de 2011), el número de artículos redactados es 71.000, sobre un total de 99.000.

den a entradas pertenecientes a otras letras del alfabeto, de acuerdo con el sistema particular de progresión en la redacción a que me he referido.

Edición electrónica

La idea de divulgar en el curso de su realización la parte redactada del diccionario estaba latente desde el inicio de la obra, dado que la estructura de datos que la constituye permite esta posibilidad sin grandes dificultades. Esta idea fue tomando cuerpo después de valorar positivamente la importancia de un procedimiento de publicación como este, que se aviene con el tono general del proyecto; la publicación electrónica dinámica y el acceso telemático a los datos del DDLC no solo se integran de una manera apropiada al plan de trabajo de la obra, sino que se ajustan plenamente a su planteamiento innovador. Así, pues, una vez el conjunto de artículos redactados había adquirido un cierto volumen, se decidió poner en marcha el proyecto de difusión de la obra por vía electrónica.

Esta edición electrónica —consultable en la web del IEC (http://www.iec.cat)— consta de un componente nuclear —constituido por el conjunto de artículos redactadados— y de unos componentes accesorios, que son una *Presentación*, muy breve y sintética, y una *Guía de utilización*, más larga y compleja. Dado que las consultas se realizan directamente sobre la base de datos del diccionario, la edición electrónica está continuamente actualizada: por una parte, el número de artículos consultables va aumentando a medida que se da por terminado el proceso de redacción de cada grupo, y, por otra parte, se incorpora automáticamente cualquier modificación de que sea objeto la parte ya redactada.

Con esta edición electrónica, el DDLC se pone a disposición del público mediante una estructura dinámica que permite la relación entre elementos de un mismo artículo o de distintos artículos del diccionario y también con elementos complementarios externos a los propios artículos a través de los vínculos activos adecuados.

Referencias bibliográficas

Geeraerts, D., "Les données stéréotypiques, prototypiques et encyclopédiques dans le dictionnaire", *Cahiers de Lexicologie*, 46, 1, 1985, 1, págs. 27-43.

Knowles, F. E., "The Computer in Lexicography", in Hausmann, F. J. & al. (eds.), *Wörterbücher. Ein internationales Handbuch zur Lexicographie. Dic-*

tionaries. An International Encyclopedia of Lexicography. Dictionnaires. Encyclopédie internationale de lexicographie, Walter de Gruyter, Berlin, New York, II, 1990, págs. 1645-1672.

Quemada, B., "La nouvelle lexicographie", in Cabré, M. T. & al., (eds.), *La lingüística aplicada*, Universitat de Barcelona-Fundació Caixa de Pensions, Barcelona, 1990, págs. 55-78.

Rafel, J. (dir.), *Diccionari de freqüències*, Barcelona, Institut d'Estudis Catalans, 1. *Llengua no literària*, 1996. 2. *Llengua literària*, 1998. 3. *Dades globals*, 1998.