3 - LEXICOGRAFÍA COMPUTACIONAL

ESTRUCTURA Y ESTRATEGIAS DE UNA BASE DE DATOS TEXTUAL: BDTLC-DCC

Josep Maria Domènech i Gibert
Àngels Egea i Puigventós
Joan Soler i Bou

Institut d'Estudis Catalans
Barcelona

1. Introducción: el DCC y el CTILC

El tratamiento automatizado de corpus textuales se ha convertido en los últimos años en un punto de referencia para cualquier estudio sobre aspectos muy diferentes de la lingüística. En especial, se ha visto claramente la necesidad de disponer de este tipo organizado de información como base fundamental para la realización de trabajos lexicográficos, en los cuales cada vez más se va afianzando la idea de que la descripción del significado está íntimamente ligada al análisis de los contextos en que aparecen los elementos léxicos.[1]

La constitución de corpus de referencia debe plantearse como objetivo metodológico prioritario la concepción polivalente desde el punto de vista de sus distintas posibilidades de aplicación. El hecho de que los corpus puedan -o deban- estar orientados hacia un objetivo principal no debe condicionar negativamente el resto de los aspectos de la investigación lingüística. En el caso de corpus de gran extensión esto adquiere una importancia muy especial, si no decisiva, ya que su elevado coste sólo puede justificarse en función de su utilidad general.

El Diccionari del català contemporani (DCC) es un proyecto del Institut d'Estudis Catalans (IEC) que se inició en el año 1985. Este proyecto, dirigido por Joaquim Rafel i Fontanals, se propone como finalidad fundamental la redacción de un diccionario descriptivo del catalán contemporáneo, para lo cual es preciso el establecimiento previo del Corpus Textual Informatitzat de la Llengua Catalana (CTILC), que se concibe como un corpus textual de grandes dimensiones (53.000.000 de ocurrencias) articulado en diversas secciones cronológicas y temáticas. El corpus integra información léxica procedente de textos catalanes escritos durante los últimos 150 años, y se estructura en 23 grupos cronológicos de segmentos de 15 años para los grupos más alejados y de 5 para los más próximos. En cuanto al eje de distribución tipológica, el CTILC divide los textos en dos grandes grupos (pertenecientes a la lengua literaria y no literaria respectivamente), los cuales se subdividen a su vez en subespecificaciones tipológico-temáticas (véase nota 10).

2. Procesos en la constitución del CTILC

La constitución del CTILC y todo el tratamiento de la información se desarrolla en tres fases que se superponen parcialmente y que son: a) Selección de los textos a introducir. b) Introducción y verificación de los textos. c) Lematización. La fase a) dio lugar al establecimiento de un repertorio de autores y obras (RAO) que consiste en una base de datos bibliográfica que contiene información sobre 6.273 autores y 25.880 referencias de obras pertenecientes al período temporal del corpus. Entre las

[1] Esta idea, en su formulación moderna, parte de las reflexiones sobre el significado lingüístico expuestas por Wittgenstein.

informaciones contenidas por este fichero, cabe destacar las referentes al año de publicación de la obra (que determina su inclusión dentro de un grupo cronológico), y la clasificación tipológica (que determina, en primer término, si la obra en cuestión se refiere a lengua literaria o no literaria; y, en segundo término, a qué grupo pertenece dentro de cada uno). La fase b) representa, en primer lugar, el traslado de los textos a soporte magnético y la ejecución de los primeros procesos conducentes a los tratamientos posteriores del corpus. La introducción del texto sobre soporte magnético supone la ejecución de dos operaciones simultáneas: la primera, de tipo transcriptivo, constituye el traslado de soporte propiamente dicho, e integra las operaciones de desambiguación de aquellos caracteres que tienen diversas funciones distintas[2], así como la incorporación de las informaciones tipográficas que afectan a cada segmento del texto (cursiva, negrilla, etc.); la segunda, de tipo indicativo, constituye la incorporación de la codificación lógica necesaria para la caracterización de la información. Este tipo de marcaje puede tener consecuencias de diversa naturaleza en los procesos automáticos posteriores. Por ejemplo:

Código	Significado	Función
¬...¬	Texto no analizable	Eliminar información
+...+	Nombre propio	Marcar información para procesos selectivos
$...$	Extranjerismo	Marcar información para procesos selectivos
¥...¥	Sigla o abrev.	Marcar información para procesos selectivos
<...>	Nombre común may.	Marcar información para procesos selectivos

Una vez introducido en un formato correspondiente al de la edición de referencia sobre papel, cada texto es procesado independientemente para obtener la agrupación de todas las ocurrencias que comparten una misma secuencia de caracteres. Gráficamente, esto representa el paso del formato del texto en su configuración lineal a un tipo de estructura mucho más adecuada para los tratamientos posteriores, pero en la cual no están presentes las informaciones no estrictamente léxicas: es decir, no se tienen en cuenta ni los caracteres prosódicos en general (signos de puntuación, comillas, etc.) ni los fragmentos que se han considerado no analizables.

La fase c) constituye la incorporación a cada una de las ocurrencias del texto de la información necesaria para caracterizar gramatical y morfológicamente a cada una de ellas. A través de esta operación se asigna a cada ocurrencia un lema (y, en consecuencia, también una categoría gramatical), así como la caracterización flexiva que le corresponde como forma de ese lema. Este es un proceso semiautomático en el que ahora no vamos a entrar, pero que constituye la compleción de la información resultante de la fase anterior y la obtención para cada texto de lo que denominamos AMLO (Archivos de palabras lematizadas por obra), que constituye un conjunto de ficheros informatizados cada uno de los cuales es una unidad independiente del resto.

Para las labores de lematización del corpus se ha contado con un Diccionario básico informatizado (DBI), que contiene las informaciones necesarias para la caracterización a que nos referíamos, esto es, los lemas con cada una de sus formas.

---

[2] Es necesario, por ejemplo, tratar de forma distinta los casos en los que un guión aparece en una palabra partida a final de línea, en cuyo caso no forma parte del contenido gráfico de la palabra y, por lo tanto, ésta deberá reconstruirse sin guión, de los casos en los que el guión es un carácter inherente a la palabra en cuestión.

En el momento actual (agosto de 1992), se han introducido unos 43.000.000 de ocurrencias, de las cuales más de 27.000.000 se encuentran ya lematizadas.

3. Sistema de explotación del CTILC: BDTLC-DCC[3]

Todo el software diseñado para la fase de recogida del material y constitución previa del CTILC permite, pues, la introducción, corrección y lematización de cada uno de los textos considerado como una unidad independiente. Este tipo de estructura, que resulta óptima para la fase de recogida del material, dificulta notablemente -por razón de las dimensiones del corpus- el tratamiento global de la información, es decir, la ejecución de procesos (de consultas o de otro tipo) referidos a la totalidad o a una parte significativa del corpus. Cualquiera de estas operaciones, por simple que fuese, representaría un tiempo de ejecución demasiado alto, aun con una configuración informática muy superior a la que actualmente cubre las necesidades del proyecto.[4] Con el fin de poder realizar operaciones cualesquiera que resulten en cierto modo no condicionadas por el volumen de datos a tratar, se constituyó una base de datos textual a partir de la información contenida en los ficheros individuales de cada obra. El paso de una estructura de información a la otra se realiza como parte integrada del tratamiento del texto, y las labores de creación de la BDTLC-DCC comprenden las siguientes fases:

1. Creación del AGML (Archivo General de Palabras Lematizadas) a partir de los AMLO.
2. Incorporación de la información del «entorno» textual al AGML, creación de la BDTLC (Base de Datos Textual de la Lengua Catalana).
3. Implementación de los sistemas de mantenimiento y consultas, etc.
4. Diseño e implementación de utilidades dirigidas a análisis complementarios de tipo sintáctico-semántico, y a la integración del BDTLC en una estación de trabajo lexicográfica.

Actualmente se ha completado la fase 1, de manera que se ha constituido la estructura de datos que permite incorporar la información que se va procesando en las fases de incorporación y tratamiento del CTILC. De los 27.000.000 de ocurrencias lematizadas ya se han incorporado a la estructura del AGML más de 26.000.000 (el 31 de agosto). En cuanto a las fases 2 y 3, se encuentran en un estado de desarrollo muy avanzado, y se terminarán en breve. La fase 4 está siendo objeto de estudios previos destinados, por una parte, a la definición y caracterización del sistema de formalización de este tipo de informaciones y, por otra, al diseño de su implementación en nuestro sistema.

Físicamente, la base de datos textual de la lengua catalana (BDTLC-DCC) se concibe como un sistema de ficheros relacionados a partir de dos criterios fundamentales: por una parte, la optimización del espacio en disco ocupado por los datos; y por otra, la previsión de las operaciones a realizar para poder contar con unos tiempos de respuesta relativamente bajos.

Por razón de optimizar el almacenamiento del material, se ha previsto que no se almacenará el texto en cuanto tal, sino que se reconstruirá a partir de la seriación de las sucesivas referencias de localización, con lo cual debemos incorporar a la BD la información correspondiente relativa a los caracteres que pueden encontrarse entre una ocurrencia y otra.

4. Estructura del AGML

El Archivo General de Palabras Lematizadas (AGML) es una base de datos que integra informaciones procedentes de los ficheros lematizados por obra (AMLO), del Repertorio de Autores y Obras (RAO) y del Diccionario Básico Informatizado. Las informaciones a las que tiene acceso el AGML son las siguientes:

---

[3] Véase figura 1.

[4] La configuración de hardware con que se lleva a cabo el proyecto es un ordenador AS/400 de IBM (modelo E-70), de 96 Mb y 17,8 Gb de memoria en disco.

Procedentes de los AMLO:

- Localización de la ocurrencia en el texto.[5]
- Contenido gráfico de la ocurrencia.
- Número del lema.[6]
- Código morfológico de la forma.[7]

Procedentes del DBI:

- Contenido gráfico del lema.
- Categoría gramatical del lema.
- Código de procedencia del lema.[8]
- Código de procedencia de la forma.
- Código de normalizado / no normalizado del lema.[9]
- Código de normalizado / no normalizado de la forma.

Procedentes del RAO:

- Autor.
- Título de la obra.
- Año de edición.
- Tipo de lengua (literaria / no literaria)
- Caracterización tipológica.[10]

Por razones de optimización de espacio, todas estas informaciones no se encuentran físicamente en un mismo fichero, sino que se ha configurado un sistema complejo de ficheros auxiliares que se interaccionan con un fichero principal (el AGML propiamente dicho). Concretamente, de todas las

---

[5] Número de obra, página, línea y orden dentro de la línea.

[6] Nos referimos al número de referencia en el fichero de lemas del DBI.

[7] El número de lema y el código morfológico de la forma son las informaciones suficientes y necesarias para la solución de lematización de la ocurrencia.

[8] El código DFA indica que la palabra ha entrado en el DBI a partir del *Diccionari general de la llengua catalana*. A las palabras que aparecen en el *Diccionari de la llengua catalana* y no en el anterior se les asigna el código DEC. Cuando las palabras representan una síntesis, ya sea por su significado o por su categorización gramatical, de los dos diccionarios, llevan el código DFE. Para acabar, las palabras que aparecen en el corpus y no hemos encontrado documentadas en ninguno de los dos diccionarios las codificamos como DCC (Diccionari del Català Contemporani).

[9] Entendemos que una palabra es normalizada cuando sigue las normas ortográficas establecidas por el Institut d'Estudis Catalans.

[10] Se refiere a las subdivisiones de los tipos literario y no literario. Las obras pertenecientes al género literario se dividen en 4 grupos (poesía, novela, teatro y ensayo). Para la división de las obras no literarias, se realizó una adaptación de la Clasificación Decimal Universal (CDU):

00 - Correspondencia	05 - Ciencias puras, naturales
01 - Filosofía	06 - Ciencias aplicadas
02 - Religión y teología	07 - Arte. Divertimentos. Deportes
03 - Ciencias sociales	08 - Lengua y literatura
04 - Prensa	09 - Historia y geografía. Biografía

informaciones que hemos citado, las únicas que podemos encontrar físicamente en el fichero principal del AGML son las que provienen de los archivos lematizados por obra. Con esto, se ha conseguido que este fichero contenga registros de poca longitud y que, por lo tanto, se reduzca de manera significativa la ocupación de disco. En cualquier caso, cabe decir que la extensión del fichero principal (que contendrá más de 53.000.000 de registros) no es reductible en lo que respecta al nivel de las localizaciones, ya que estas jamás se repiten. Desde el punto de vista del tratamiento que recibirán posteriormente, los registros del AGML pueden clasificarse en dos grandes grupos:

a) Ocurrencias propiamente analizables.
b) Ocurrencias marcadas como nombre propio.

Los nombres propios no tienen asignada ninguna solución de lematización y no se tienen en cuenta en las operaciones de consulta (al menos hasta que se diseñe un sistema de consultas específico para este caso). Aunque no se asigne solución de lematización a estas ocurrencias, la estructura física de los registros es exactamente igual que la de los analizables, a fin de mantener la posibilidad de pasar una ocurrencia del grupo b) al a), o viceversa.

5. Recuperación de la información en el AGML

El AGML permite la obtención de todo tipo de datos (léxicos, estadísticos, etc.) que no precisan de la reconstrucción textual. En especial, se han diseñado una serie de índices KWOC (Key Word Out of Context) que aportan información sobre lemas, lemas y formas, formas, y también sobre distribuciones frecuenciales.

5.1. Índices de lemas.

Los índices descritos en este apartado muestran los lemas, y los datos relacionados con éstos, ordenados desde perspectivas distintas según las posibilidades de ordenación de que disponemos en cada caso. El Índice General de Lemas del Corpus y el Índice de Lemas Secundarios[11] tienen un carácter básicamente descriptivo, mientras que el Índice Diferencial y el Listado de Control de Crecimiento del Corpus tienen un carácter más comparativo que permite una valoración del estado del corpus en función de la información procesada en un momento determinado y en relación a los objetivos finales.

5.1.1. Índice general de lemas del corpus.

El índice general de lemas puede corresponder a cinco tipos de listados de formato bastante similar y que vienen determinados por los criterios de ordenación de los lemas que se definen en el momento de hacer su petición. La información está distribuida en cinco columnas donde se especifica la etiqueta del lema, su código gramatical, su código de procedencia y su frecuencia absoluta dentro del corpus de trabajo definido por el usuario -que puede corresponder a la totalidad del corpus o a una parte- y, según el tipo de listado de que se trate, se pueden obtener otras informaciones como la frecuencia relativa, el tipo de lengua a que pertenece cada lema listado y el lema principal que le corresponde si es secundario. Al final de cada listado se especifica el número total de lemas listados y el de ocurrencias lematizadas, es decir, el número de ocurrencias que suman el total de obras con que se ha trabajado.

Al principio del listado se especifica, por orden de código de obra, las obras que pertenecen al corpus definido. Hay maneras distintas de delimitar el corpus sobre el que deseamos trabajar. El

[11] Un lema es secundario de otro lema ya existente, al que nos referiremos como principal, en el caso de que compartan un mismo contenido semántico y se asemejen formalmente, pero no sea posible identificar como formas del lema existente ninguna de las ocurrencias que han dado lugar a la creación del lema secundario.

resultado que se obtenga dependerá de los parámetros de selección que se hayan definido. Estos parámetros nos permiten seleccionar las obras que pertenecen a uno o más grupos cronológicos, o bien seleccionar las obras que pertenecen a un tipo de lengua determinado. Estos parámetros de selección cronológica y tipológica, además, pueden combinarse. También se puede seleccionar una obra o más de una obra de un mismo autor como corpus de trabajo.

Trabajando sobre unos mismos datos, obtendremos listados diferentes según como se ordene la información. Hay cinco criterios posibles de ordenación:

- Ordenación alfabética por lema.
- Ordenación alfabética inversa.
- Ordenación por frecuencia decreciente.
- Ordenación por código de procedencia. Los lemas se listan agrupados según el código de procedencia y ordenados alfabéticamente dentro de cada grupo. Cada conjunto de lemas con un mismo código de procedencia constituye una sección del listado. Al final de cada sección se especifica el número total de lemas listados y el número total de ocurrencias.
- Ordenación por tipo. Los lemas se agrupan por el tipo de lengua a que pertenecen y se ordenan alfabéticamente. Los lemas pueden ser de tipo GN (General), LT (Literario) y NL (No Literario). Los de tipo general pueden formar parte de la lengua literaria y de la no literaria, es decir, contienen formas que están asociadas a ocurrencias en obras que pueden pertenecer a uno u otro grupo. En cambio, los lemas de tipo literario o los de tipo no literario son exclusivos de cada grupo. El listado está estructurado en secciones, como en el caso anterior, y al final de cada sección se especifica el número de lemas listados y de ocurrencias lematizadas.

5.1.2. Índice de lemas secundarios.

Contiene los lemas secundarios que se encuentran en el AGML. Se puede delimitar el corpus de trabajo aplicando el mismo sistema de parámetros de listado descrito para el Índice General de Lemas del Corpus. Los lemas secundarios contenidos en el corpus de trabajo definido se listan por orden alfabético. Por cada lema secundario se indica el código gramatical, el código de procedencia, la frecuencia absoluta y el lema principal que le corresponde. Esta información está estructurada, como en los casos anteriores, en forma de columnas. Inicia el listado una descripción de los parámetros seleccionados y una lista de todas las obras que forman parte de la extensión de corpus delimitada.

5.1.3. Índice diferencial.

El objetivo de este índice es dar cuenta del número de lemas de procedencia DFA, DEC y DFE, incorporados al Diccionario Básico Informatizado (DBI), que no se encuentran aún en ninguna solución de lematización del AGML.

5.1.4. Listado de control de crecimiento del corpus.

Este listado nos informa de los lemas utilizados en soluciones de lematización entre dos fechas concretas y que no se habían utilizado con anterioridad a la primera de estas fechas, que reciben el nombre de fechas de cierre o simplemente cierres y que corresponden al momento de incorporación al AGML de datos procedentes de las últimas obras lematizadas. El hecho de que los lemas tengan asociada la fecha de su alta en el corpus, nos permite obtener esta información entre dos fechas de cierre cualesquiera. A partir de estos datos es posible hacer una valoración del ritmo de crecimiento del repertorio de lemas del corpus en relación con el número de ocurrencias lematizadas.

5.2. Índice de lemas y formas.

A partir del listado generado por esta opción, que hace uso de los mismos argumentos de selección que ya se han especificado para los listados del índice de Lemas, podemos obtener una

descripción, para cada lema, de todas las formas con ocurrencias en el corpus. Los lemas se listan por orden alfabético, indicando el código gramatical, el código de procedencia y las frecuencias absoluta y relativa. Las formas de un lema también se listan por orden alfabético. De cada forma, se indica el código morfológico, el código de procedencia y las frecuencias absoluta y relativa. Eso permite observar de qué modo se distribuye la frecuencia de un lema determinado entre sus formas. Un resumen final nos muestra el número de lemas principales y secundarios listados, el número total de formas listadas y el número de ocurrencias lematizadas.

5.3. Índice de formas.

El Índice de Formas se subdivide en varios índices:

- índice inverso de formas
- índice alfabético de formas D
- índice inverso de formas D
- índice alfabético de formas con localizaciones

Para la obtención de cualquiera de estos listados se define un corpus de trabajo de acuerdo con los parámetros de selección cronológica y tipológica ya indicados.

5.4. Distribución de frecuencias.

El objetivo de estos listados es dar una relación detallada de la distribución de los datos referentes a la frecuencia absoluta y relativa de cada lema en cada una de las subdivisiones de carácter cronológico y tipológico establecidas. Existe un listado de Distribución Cronológica de los lemas y uno de Distribución Tipológica. El primero consta de 23 columnas que corresponden a los 23 segmentos cronológicos en los que está subdividido el corpus textual. El segundo se subdivide en dos grandes grupos que se pueden listar conjuntamente o independientemente uno de otro según optemos por listar sólo los lemas del corpus literario, sólo los del no literario o ambos. Los lemas del literario se distribuyen en 4 columnas; los lemas del no literario se distribuyen en 10 columnas. Los parámetros de filtraje de la información a listar son en los dos casos los mismos:

a) listar sólo los lemas con una frecuencia absoluta superior a X
b) listar sólo los lemas distribuidos en más de X columnas
c) listar sólo los lemas pertenecientes al corpus literario, listar sólo los lemas pertenecientes al corpus no literario, listar los lemas de la totalidad del corpus.

Otro listado llamado Resto recoge los lemas que han quedado excluidos por los parámetros de filtraje.

Existe un tercer listado de distribución de frecuencias de naturaleza diferente a la de los anteriores. Se trata del Listado de Distribución de frecuencias por códigos gramaticales. Especifica la frecuencia absoluta y relativa que tiene cada código gramatical en el corpus.

6. Estrategia de reconstrucción textual: del AGML a la BDTLC[12]

Como ya se ha dicho, el AGML es una estructura de datos que sólo permite la ejecución de operaciones cuyos resultados no implican reconstrucción del contexto (KWOC). En el caso de algunos

[12] Véase figura 1.

listados de análisis de formato KWIC (Key Word in Context), como las concordancias, se precisa de la reconstrucción contextual, para lo cual se debe incorporar al AGML la información necesaria para garantizar su total autonomía en cuanto a esa capacidad.

Desde el punto de vista del resultado final que se quiere obtener, las informaciones a añadir en el AGML para llegar a la formación de la Base de Datos Textual de la Lengua Catalana (BDTLC) pueden dividirse básicamente en tres tipos diferentes:

- Fragmentos no analizables.
- Informaciones lógicas.
- Informaciones tipográfico-contextuales.

6.1. Fragmentos no analizables.

Durante el proceso de introducción de las obras sobre soporte magnético, se codificaron como no analizables los fragmentos de texto que quedaban fuera de los objetivos de nuestro corpus (fragmentos en lengua extranjera) o que no podían ser sometidos a un tratamiento lingüístico (cifras, fórmulas, etc.). Esta información, que no ha sido tenida en cuenta en ninguna de las operaciones realizadas hasta este momento, se hace imprescindible para la reconstrucción textual. Para la indización de esta información se decidió crear un registro en la BDTLC para cada cadena de caracteres alfanuméricos comprendida entre un separador y otro separador, fuese cual fuese su contenido. Estos registros, del mismo modo que los nombres propios, tienen idéntica estructura que los demás pero no se tienen en cuenta en las operaciones de explotación de la base de datos ni en los cálculos estadísticos. Por lo tanto, serán requeridos, únicamente, para la reconstrucción del texto.

6.2. Informaciones lógicas.

En este caso, se trata de incorporar a la Base de Datos las informaciones asociadas a las parejas de signos lógicos que han sido incorporados en el momento de la introducción de la obra sobre soporte magnético y que no han recibido ningún tratamiento específico hasta ahora. En el punto 2 se han expuesto algunos de los códigos lógicos usados en el CTILC.

Todos estos códigos pueden estar combinados entre ellos, en algunos casos, hasta un máximo de tres. La combinación de dos o tres códigos provoca que la incorporación de estas informaciones sea especialmente compleja ya que deben tenerse en cuenta cuales son las posibles combinaciones. Es posible, además, que algunos de estos códigos aparezcan combinados con el de no analizable (_). Existe la posibilidad de que la inversión del orden de dos códigos provoque informaciones diferentes. Por ejemplo, la combinación

= ¬ ¬ =

(un título no analizable) es admisible, pero la alteración del orden

¬ = = ¬

provoca que el signo = deje de tener valor lógico para convertirse en un literal.

La incorporación de todas estas informaciones de carácter lógico a la Base de Datos debe permitir, en el futuro, la realización de operaciones específicas sobre subconjuntos lógicos del corpus, como listados de siglas, títulos, etc.

6.3. Informaciones tipográfico-contextuales.

En este caso, se trata de recoger las informaciones necesarias para permitir la reconstrucción del texto a partir de la concatenación de localizaciones.

6.3.1. Informaciones tipográficas.

Por una parte, debe indicarse para cada ocurrencia si está afectada por un código de cursiva o negrilla. Existe la posibilidad de que los dos códigos aparezcan combinados en una misma ocurrencia. En este caso, deberán consignarse los dos en el registro correspondiente de la base de datos.

Para eliminar la posibilidad de duplicación de formas idénticas, las palabras que aparecían en el texto en mayúscula por su situación (por ejemplo, después de punto), se neutralizaron a minúscula. Para asegurar una reconstrucción fiel de los textos, se ha creado un campo en el que se indica, cuando es necesario, que la palabra aparece en mayúscula. La neutralización no se realizó en las formas codificadas como siglas o nombres propios, por lo que, en estos casos, la información de este campo será redundante.

6.3.2. Informaciones contextuales.

Para conseguir la reconstrucción total del contexto, no basta con saber si la palabra va en mayúscula o minúscula, en redondilla, cursiva o negrilla. Aparte de estas informaciones estrictamente tipográficas, hay que recoger, también, las secuencias no alfabéticas que se encuentran entre palabra y palabra (signos de puntuación, códigos de salto de párrafo, etc.). La indización de esta información se ha conseguido con la creación de un campo que archiva la cadena de caracteres que hay entre el último carácter de una ocurrencia y el primero de la siguiente. Por ejemplo, en el caso:

...treball). «Avui,

archivaríamos que la cadena alfanumérica de la derecha de la forma «treball» es:

). «

La última información que debe recogerse para la reconstrucción del texto es si entre el final de la cadena no alfabética de un registro y el primer carácter del siguiente hay un espacio en blanco. Por ejemplo:

havia d'ajudar-lo
1 2 3 4

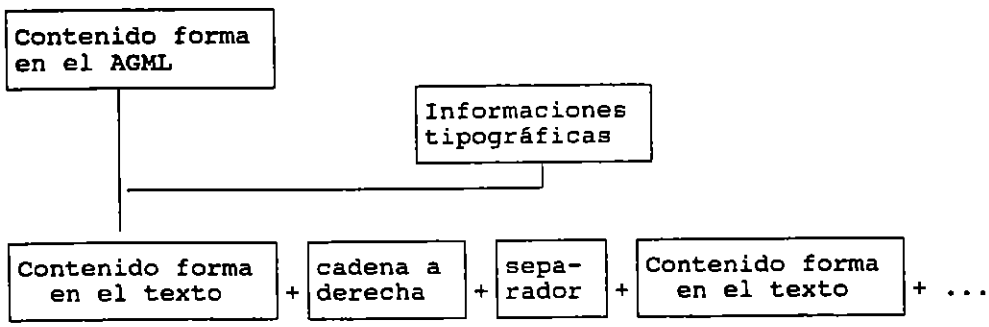
En este caso, el programa deberá indicar en el campo correspondiente la existencia de un espacio en blanco entre 1 y 2, pero no entre 2 y 3 o 3 y 4. Esta información no puede recogerse de manera generalizada (considerar, por ejemplo, que no habrá nunca un espacio en blanco después de apóstrofe) porque al lado de un caso como el que hemos visto podríamos encontrar:

havia d' ajudar-lo
1 2 3 4

en el que sí encontramos un espacio en blanco entre 2 y 3 (éste es un caso muy frecuente en textos escritos en lengua no normalizada).

6.4. Reconstrucción del texto.

Una vez incorporadas todas estas informaciones, la reconstrucción del texto se realizará a partir de la concatenación de los siguientes elementos:



7. Módulos de mantenimiento

El sistema de mantenimiento de la BDTLC ha tenido que crearse teniendo en cuenta dos aspectos importantes. Por una parte, debe haber un control de los eventuales cambios que pudieran hacerse en otros ficheros que puedan afectar a las informaciones que ya están archivadas en la Base de Datos. Por ejemplo, no se debe poder suprimir nunca del DBI un lema mientras en la Base de Datos haya ocurrencias que estén remitidas al mismo. Por otra parte, hay que tener presente que el cambio en un registro de la Base de Datos puede hacer variar los cálculos estadísticos que se hayan hecho y, por ello, el programa deberá regenerarlos en el momento de hacer la corrección.

A medio plazo, el sistema de mantenimiento de la BDTLC deberá permitir diversas operaciones de diferente naturaleza: desde el cambio de la solución de lematización hasta el cambio del contenido gráfico de la propia ocurrencia, pasando, por ejemplo, por la incorporación de una nueva línea de texto, que representaría el caso de más difícil solución.[13]

En estos momentos, disponemos de un sistema de mantenimiento que nos permite variar la solución de lematización de una ocurrencia. Es una tarea sencilla que representa, solamente, cambiar el contenido de los campos «código morfológico» y «número de lema». Disponemos, también, de una opción que nos permite realizar este cambio de manera global; es decir, cambiar una solución de lematización por otra en todos los registros de la Base de Datos que tengan el mismo contenido de forma, el mismo código morfológico y el mismo número de lema. Esta opción permite que, si es necesario realizar un cambio de este tipo, podamos hacerlo de una manera muy rápida y segura.

8. Sistema de consultas

8.1. Consultas al AGML.

El sistema actual de consultas al AGML es provisional y deberá ampliarse convenientemente en fases sucesivas de desarrollo del proyecto y, principalmente, de cara a la explotación de la BDTLC. En general, el tipo de consultas que se han desarrollado hasta ahora vienen condicionadas por las necesidades internas del proyecto, y se dirigen, por tanto, a la verificación de datos del AGML. Se ha puesto una atención especial en el hecho de que el sistema se corresponda totalmente con el que se aplicará sobre las consultas de la BDTLC que no requieran de la reconstrucción textual.

---

[13] Pensemos, por ejemplo, que, si añadimos una línea de texto en una página, nos cambiará la localización de todas las ocurrencias de las líneas posteriores.

El sistema de que disponemos en estos momentos permite hacer tres tipos de consultas:

- Consultas por forma: El objetivo de la consulta es conocer en qué localizaciones ha aparecido una forma determinada.
- Consultas por lema: El objetivo, en este caso, es conocer en qué localizaciones han aparecido las formas de un lema determinado. El usuario tiene la posibilidad de hacer la consulta sobre una selección de las obras o sobre todo el corpus.
- Consultas por localización: Esta opción permite al usuario saber cual es la solución de lematización que se ha dado a una ocurrencia determinada (contenido gráfico de la forma, código morfológico de la forma, contenido gráfico o etiqueta del lema, código gramatical del lema y número del lema).

8.2. Consultas a la BDTLC[14]

Para la futura explotación de la BDTLC se ha previsto la creación de un sistema complejo de consultas que, actualmente, se encuentra en fase de estudio y documentación.

8.2.1. Módulos de selección.

El sistema se concibe a partir de tres módulos de selección básicos:

- Selección de datos.
- Selección de subcorpus.
- Selección de contexto.

El módulo de selección de datos trabajará con las informaciones del Diccionario Básico Informatizado (DBI). El usuario podrá seleccionar con qué forma, o qué formas, desea trabajar.[15]

El módulo de selección de subcorpus trabajará con los datos archivados en el RAO. En este módulo, el usuario seleccionará estos parámetros:

- Parámetros cronológicos.
- Parámetros temáticos.
- Parámetros de autores.
- Parámetros de obras.

Las posibilidades de combinación de este tipo de parámetros no estarán limitadas sólo a ciertas combinaciones; es decir, el usuario podrá, por ejemplo, seleccionar un determinado período cronológico y, dentro de éste, elegir una temática concreta.

En cualquiera de estos dos primeros módulos (selección de datos y selección de subcorpus), el usuario debe poder pedir información referente al número de ocurrencias afectadas por la selección que está realizando. Es decir, si, por ejemplo, selecciona el lema ABANS AV o selecciona un grupo cronológico determinado, tiene que poder saber con cuántas ocurrencias trabajará.

Una vez seleccionados los datos que queremos consultar y el subcorpus con el que queremos trabajar, tendremos la posibilidad de delimitar la información que deseamos obtener añadiendo alguna restricción contextual, es decir, indicando si es indiferente el contexto en que aparezca o si debe llevar

---

[14] Véase figura 2.

[15] Prevemos la posibilidad de incorporar en este módulo caracteres virtuales de longitud variable o fija.

algún/unos elemento/s concreto/s a la derecha o a la izquierda. Así pues, tendríamos estas cuatro posibilidades:

```
    X
A   X
    X   A
A   X   B
```

donde X representa la palabra clave y A y B una/s forma/s o una/s categoría/s gramatical/es determinada/s.

8.2.2. Parámetros de salida.

Con la aplicación de estos tres módulos, el usuario habrá formulado ya una consulta compleja con los datos contextualizados. En este momento, deberá indicar los parámetros de salida que desea:

- Contexto: Si el usuario quiere los datos con contexto, deberá indicar su longitud (una línea, un párrafo, de punto y seguido a punto y seguido, un número determinado de palabras, etc.).
- Ordenación: El usuario seleccionará el tipo de ordenación que desea para los datos que ha solicitado (por frecuencias, alfabético de lemas, por obra, etc.).

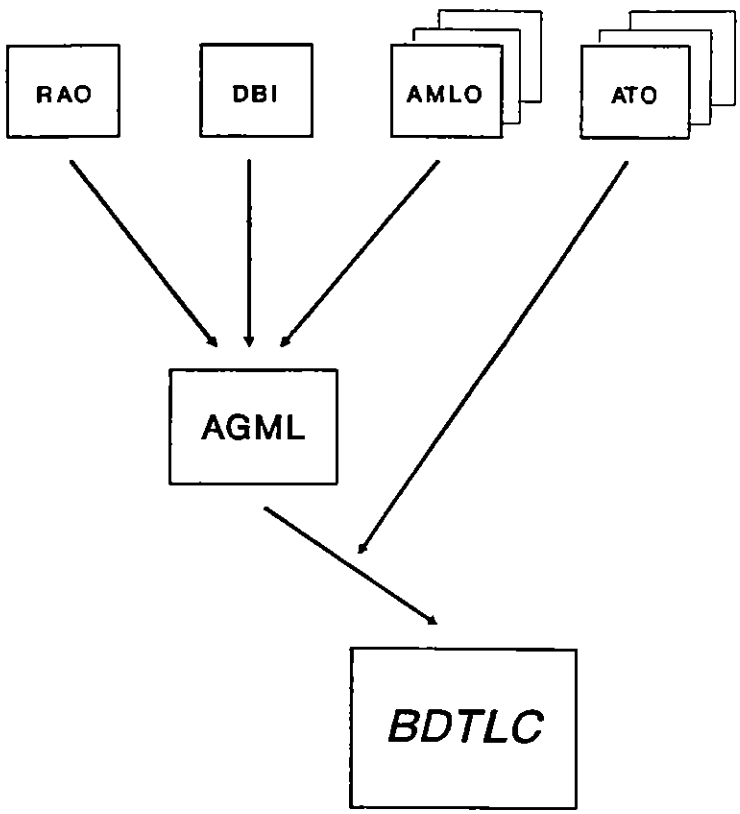
Para finalizar, el usuario indicará el tipo de salida que desea: un listado en papel, un fichero visualizable en pantalla, sobre soporte magnético, etc.

9. Ejemplo de registros de la BDTLC

Pág. Lín. Contenido de la línea

1 1 =Capítol ¬I¬: La $/madame\$=
1 2 A +França+, el "cas" del <Rei> (entre altres coses...),
1 3 va fer que /el dia ¬18¬\, ¥p. ex.¥, +París+ caigués...[F.P.]

Loc.	Cont. forma	Cód. Mor.	Cont. lema	Cód. gram.	May.	Cadena a la derec.	Separador	Códigos lógicos	Códig. tipog.
1,1,1	capítol	S	capítol	M	Sí		Sí	=	
1,1,2	i				Sí	:	Sí	= ¬	
1,1,3	la	FS	el	AR	Sí		Sí	=	
1,1,4	madame	S	madame	F	No		Sí	= $	/
1,2,1	a		a	PO	Sí		Sí		
1,2,2	França				Sí	,	Sí	+	
1,2,3	el	MS	el	AR	No	"	No		
1,2,4	cas	S	cas	M	No	"	Sí		
1,2,5	del	S	del	CT	No		Sí		
1,2,6	rei	S	rei	M	Sí	(	No	<>	
1,2,7	entre		entre	PO	No		Sí		
1,2,8	altres	P	altre	A	No		Sí		
1,2,9	coses	P	cosa	F	No	...),	Sí		
1,3,1	va	3PI	anar	VA	No		Sí		
1,3,2	fer	IF	fer	VVP	No		Sí		
1,3,3	que		que	C	No		Sí		
1,3,4	el	MS	el	AR	No		Sí		/
1,3,5	dia	S	dia	M	No		Sí		/
1,3,6	18				No	,	Sí	¬	/
1,3,7	p. ex.		pe	SIG	No	,	Sí	¥	
1,3,8	París				Sí		Sí	+	
1,3,9	caigués	3IS	caure	VI	No	...[F.P.]	Sí		



AMLO: Archivo de Palabras Lematizadas por Obra
RAO: Repertorio de Autores y Obras
DBI: Diccionario Básico Informatizado
ATO: Archivo de Textos por Obra
AGML: Archivo General de Palabras Lematizadas
BDTLC: Base de Datos Textual de la Lengua Catalana

Figura 1

SISTEMA DE CONSULTAS A LA BDTLC

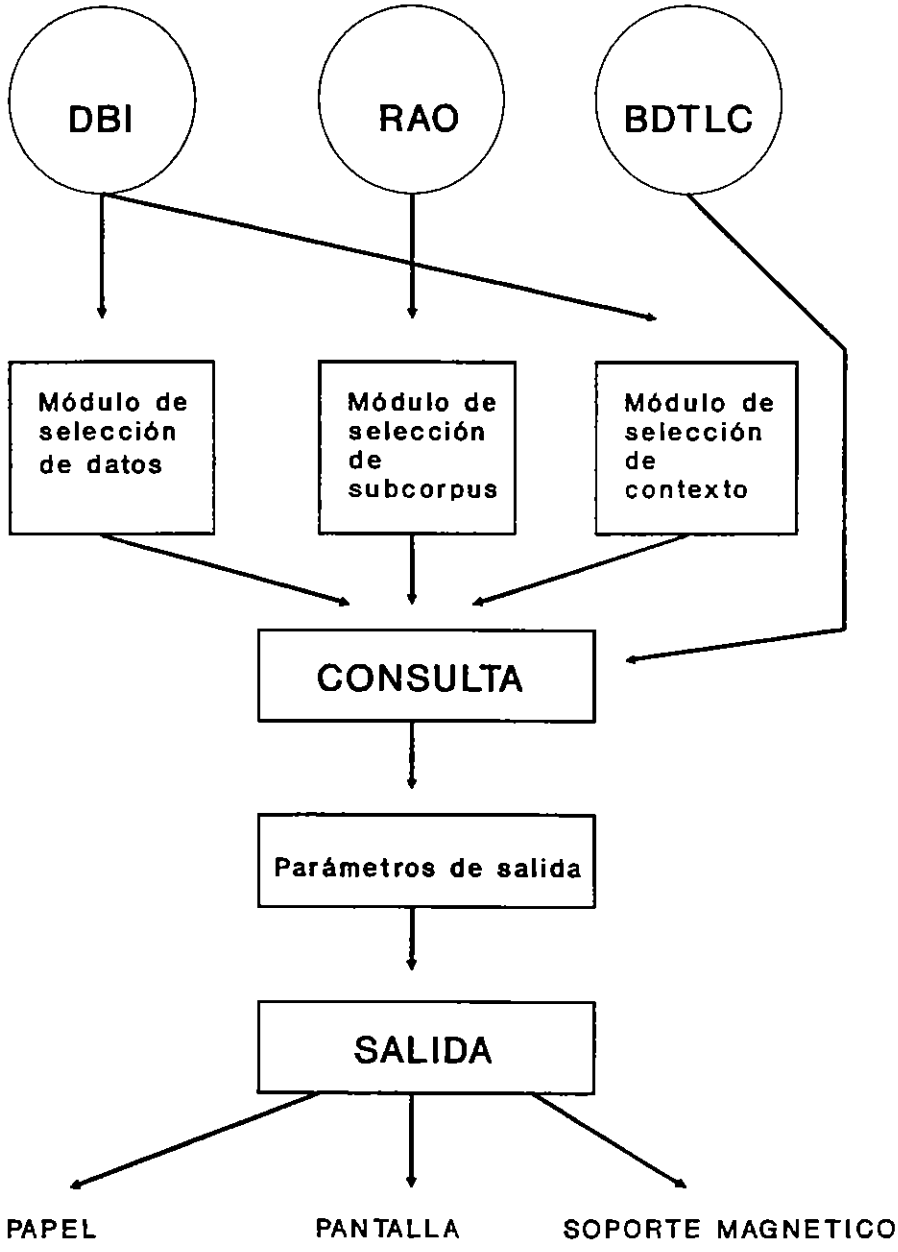


Figura 2