Proyecto APLE2: Procesamiento de los documentos a incluir en el Corpus Tècnic del IULA

Autores: Jorge Vivaldi y Elisabet Llopart
Fecha última modificación: 27/03/2015

1 Introducción

Este documento presenta todos los aspectos a tener en cuenta para el procesamiento de los documentos utilizados por el proyecto APLE2 y que se incorporarán al Corpus Tècnic (CT) del IULA.

Recomendamos una lectura global de este texto (en particular, las interacciones entre la conversión a texto plano –sección 5.1– y las correcciones del texto para su procesamiento lingüístico –sección 7.1–) antes de procesar cualquier documento a incluir en el CT.

2 Esquema general de funcionamiento

De manera general, se puede decir que el procesamiento de los documentos que se utilizarán en el proyecto APLE2 y que se incorporarán en CT se inicia en entorno Windows y finaliza con la indexación de dichos documentos en entorno Unix y su visualización con bwanaNet.

A grandes rasgos, la secuencia de pasos a seguir para incorporar un documento en el CT será la siguiente:

1) Seleccionar el texto a introducir
2) Convertir el texto seleccionado a formato electrónico (texto plano con codificación iso-8859-1). Asignarle un nombre temporal y realizar todas las operaciones de limpieza necesarias (eliminación de figuras, tablas, pie de página, cabeceras, etc.).
3) Abrir el fichero de texto resultante y utilizando el programa Word (o similar), contar el número de palabras del documento. Si el documento está formado por más de 15.000 palabras aproximadamente podría optarse por dividir (temporalmente) el documento en varias partes. Esta división no es indispensable pero podría facilitar el procesamiento del texto. Si se decidiera mantener la división en muestras, este número será necesario en la etapa 4 al introducir los datos bibliográficos[1].
4) Introducir los datos bibliográficos en la Base de Datos DB_1. Al introducir la cantidad de muestras en que se divide el documento, el sistema asignará automáticamente los nombres de fichero de cada muestra.
5) Renombrar el texto según indica la Base de Datos.
6) Realizar el preproceso de las muestras del documento
7) Comprobar el marcado estructural de las muestras (nsgmls)
8) Realizar el procesamiento de las muestras del documento
9) Completar la información necesaria para generar las cabeceras de los documentos a incorporar.
10) Generar los ficheros auxiliares para procesar los documentos mediante el CWB. Verificar (fullBalanceig.pl y errores de indexación).
11) Indexación en entorno CWB. Actualización de la Base de Datos DB_2.

La DB_1 es una base de datos Access con información bibliográfica con la cual se gestionan todos los documentos del CT. Entre los datos que se recogen en DB_1, se recogerá la información sobre los números de páginas del documento completo (ya sea este un libro o un artículo de un libro).

La Figura 1 muestra gráficamente el esquema general de funcionamiento a seguir para el procesamiento de documentos del proyecto APLE2 que se incorporarán al CT.

[1] Esta manera de actuar permite al codificador decidir autónomamente el número de muestras en que se dividirá cualquier documento.

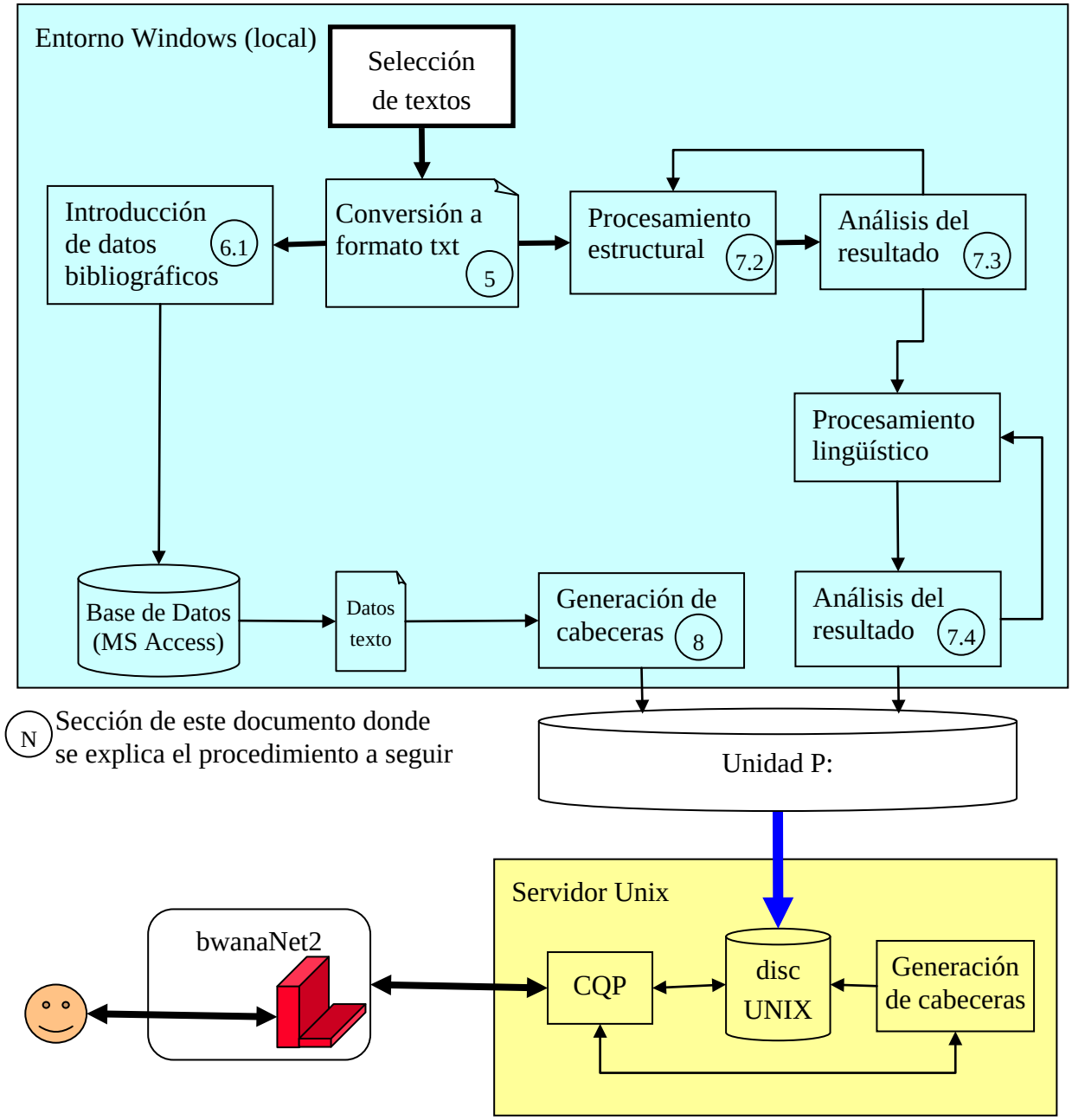


Figura 1. Esquema general de procesamiento

3 Estructura de directorios

El directorio raíz para los documentos varía en función del dominio según se indica en la tabla siguiente:

Dominio	Directorio de trabajo
Medicina	P:\IULA\CORPUS\CORPUS\TMPMEDIC
Economía	P:\IULA\CORPUS\CORPUS\TMPECON
Derecho	P:\IULA\CORPUS\CORPUS\TMPDRET
Medio Ambiente	P:\IULA\CORPUS\CORPUS\TMPMA
Informática	P:\IULA\CORPUS\CORPUS\TMPINFO

Dentro de cada uno de estos directorios se guardarán los textos a incorporar al CT tanto en su formato original como el de texto plano. En estos directorios existen otros subdirectorios cuya funcionalidad es la que se indica a continuación[2]:

- CTheaders: directorio donde se generarán, para cada documento, los tres ficheros de la cabecera
- mostres: ficheros de texto plano de cada documento
- mostres5: ficheros de texto verticalizados[3] de cada documento

4 Comprobaciones preliminares

Antes de empezar a preprocesar los textos se deben realizar las siguientes comprobaciones preliminares:

a) Acceso a la carpeta de Corpus (P:\IULA\CORPUS)

b) Acceso a la carpeta de SOFT (P:\IULA\SOFT)

   Si estas carpetas no aparecen en P:\IULA, pedir acceso a Jesús Carrasco.

c) Ventanas MS-DOS

   Para abrir una ventana MS-DOS ir a Inicio\Ejecutar y en el apartado “Abrir” escribir *cmd* y hacer clic en “Aceptar”.

d) Acceso a carpeta *tmp* en C:\ Si esta carpeta no existe en el directorio, crear una nueva carpeta en C:\ con este nombre

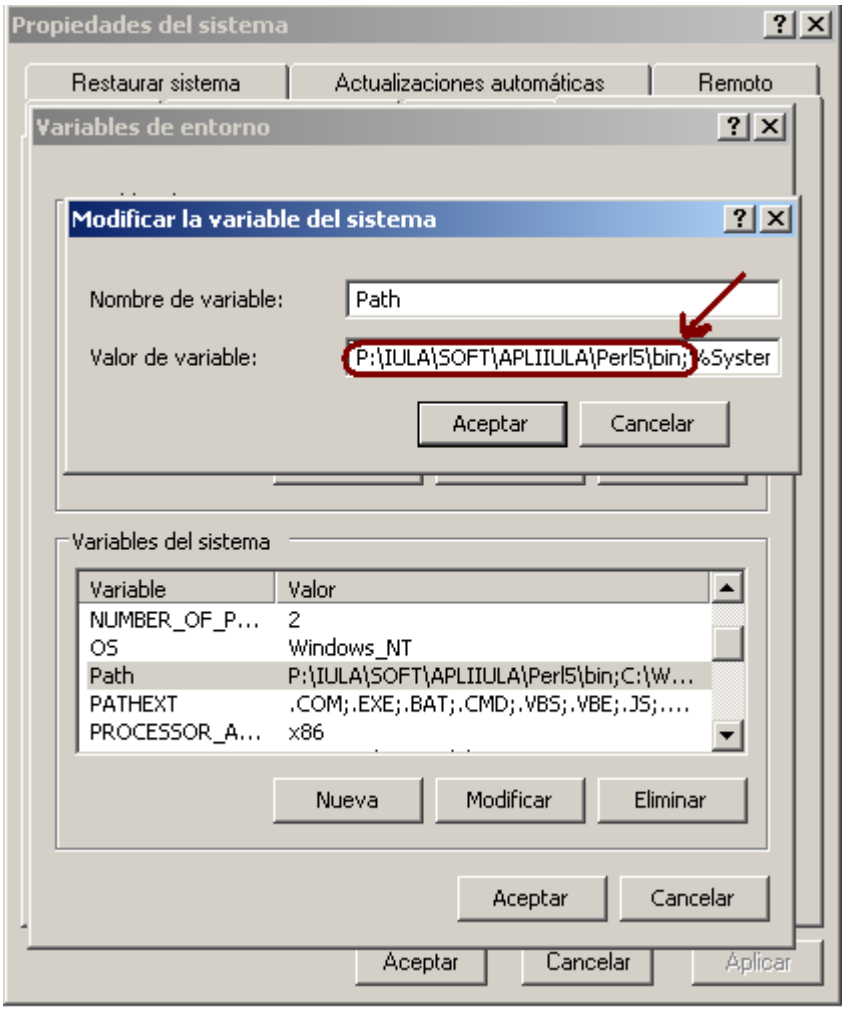
e) Acceso al intérprete Perl (P:\IULA\SOFT\APLIIULA\Perl5\bin)

   Para tener acceso a Perl seguir los pasos siguientes: Configuración/Panel de control/Sistema/Opciones avanzadas/Variables de entorno.

   En el apartado “Variables del sistema”, clicar “Path” y, entonces, clicar “Modificar”. En la nueva ventana de “Modificar la variable del sistema”, pegar la dirección de Perl (P:\IULA\SOFT\APLIIULA\Perl5\bin) a la izquierda del todo de la entrada “Valor de variable” y añadir un punto y coma (;) al final para separar del resto de texto. Atención! No introducir ningún espacio después del ‘;’. Clicar “Aceptar” en todos los cuadros de diálogo abiertos.

[2] Estos subdirectorios se generaran en una etapa posterior. Así pues, guardamos el documento en formato original y en texto plano directamente el directorio del dominio correspondiente.

[3] Es decir, el texto con un token por línea y con información lingüística (POS y lema).



Para comprobar si se tiene acceso al intérprete Perl, abrimos una ventana MS-DOS y ejecutamos el mandato: `perl –v` (para ejecutar el script debemos clicar la tecla *Intro* o *return*).

Si la respuesta del sistema es del tipo: `This is perl, v5.10.1 built for MSWin32-x86-multi-thread`… significa que se puede acceder intérprete Perl correctamente.

5 Selección de los textos y su conversión a texto plano

El primer paso es, obviamente, seleccionar textos pertinentes a cada uno de los dominios del CT. A continuación se deberán convertir los textos seleccionados de su formato original (probablemente PDF) a texto plano con codificación iso-8859-1 (es decir, legible desde la utilidad Notepad de Windows —verificar caracteres especiales—). Asignarle un nombre temporal y realizar todas las operaciones de limpieza necesarias (eliminación de figuras, tablas, pie de página, cabeceras, etc.). En la sección 7.1 hay más indicaciones sobre partes del texto a eliminar.

Para la conversión de un fichero en formato PDF a texto[4], existen al menos dos posibilidades para realizar esta tarea: utilizar el programa Adobe Acrobat Pro o bien el programa Terminus. Un aspecto importante a tener en cuenta en todos los casos es que si

[4] Para realizar la conversión a formato texto con cualquiera de las opciones de transformación que presentamos a continuación el texto a convertir no debe ser un texto protegido.

estamos procesando un texto a dos columnas es imprescindible pasar a una única columna para tener un resultado razonable[5]. Descartar el texto si no es posible convertirlo a una única columna.

De todas maneras, el procedimiento recomendado es una combinación de ambos programas que se pueden resumir como se indica a continuación:

a) Abrir el documento en PDF con el programa Adobe Acrobat Pro;
b) Ajustar los márgenes superior y inferior para eliminar cabeceras y pié de página;
c) Eliminar todas las partes del texto que no sean relevantes;
d) Guardar el documento PDF resultante (con otro nombre para no perder la versión original);
e) En el mismo Adobe Acrobat Pro, exportar de PDF a TXT (asignarle un nombre significativo[6]);
f) Abrir el TXT (para más comodidad, con el programa EditPlus*; si no, con el bloc de notas) y observar el texto;
g) Si el texto resultante no está bien exportado (ej. párrafos o líneas cortados, caracteres raros) utilizar Terminus para hacer la conversión PDF a TXT;
h) Descartar el texto si la conversión no es apropiada;
i) Utilizar EditPlus[7] para hacer las correcciones finales en el texto resultante.

En las secciones siguientes se proporciona más información para realizar los pasos indicados.

5.1 Conversión a PDF→texto con Acrobat

Esta es una de las tareas básicas a realizar con cualquier texto a incluir en el CT. En caso de fichero en formato PDF, puede ser de gran utilidad en la limpieza del texto el programa Adobe Acrobat Pro. A continuación se indica el procedimiento a seguir para convertir el texto PDF a TXT. En caso que este procedimiento no diera resultado, es conveniente experimentar con diferentes maneras de convertir el texto (PDF→texto, PDF→Word→texto, PDF→RTF→texto, etc.) y conservar aquel resultado que tenga mejor rendimiento (o sea el que introduzca el menor número de errores o éstos sean más fáciles de corregir).

5.1.1 Eliminar pies de página y cabeceras

Para ello, dentro del Acrobat Pro, se debe seleccionar: Documento→Recortar páginas y desde aquí debe aparecer la ventana "Controles de margen" (ver Figura 2). Desde aquí:

a) Modificar los valores Superior e Inferior hasta que no se vean la cabecera y el pie de página del documento;
b) Activar la casilla "Todas" en la opción de "Rango de páginas";
c) Seleccionarse "Aceptar" para confirmar las modificaciones anteriores.

[5] Otras herramientas que se podrían utilizar para hacer la conversión de n→1 columnas son i) "papercrop" (https://code.google.com/p/papercrop/) o ii) "calibre" (http://calibre-ebook.com).

[6] Este nombre se utilizará en todas las etapas hasta llegar al procesamiento lingüístico (sección 7.4).

[7] El programa de instalación se encuentra en P:\IULA\SOFT\instal\Edit+.

Este programa dispone de otras opciones (recorte diferente en páginas pares/impares, recortar a izquierda y/o derecha, etc.) que en ciertos casos podría ser necesario activar.

Tener en cuenta que esta acción no es destructiva. Por lo tanto, si vemos que los márgenes escogidos no son suficientes podemos variarlos a voluntad.

Figura 2. Recortar páginas

5.1.2 Eliminar figuras, tablas y otros elementos

El Acrobat Pro permite seleccionar porciones de texto que consideramos no relevantes (figuras, tablas, bibliografía, etc.) y eliminarlas simplificando de esta manera el trabajo posterior de conversión a TXT.

Para ello, dentro del Acrobat Pro, se deben combinar las siguientes opciones de menú (ver Figura 3):

a) Eliminar: objetos completos: Herramientas → Edición avanzada → Retocar objeto. Seleccionar con el cursor los objetos que se desean eliminar en el PDF y después presionar la tecla Suprimir (o Eliminar con el botón derecho del ratón). Es conveniente eliminar los elementos no deseados uno a uno y asegurarse de que se ha eliminado únicamente el objeto deseado. Con la secuencia Control + Z puede anularse la última acción de eliminar. La eliminación de un elemento complejo (p. ej. una tabla) puede requerir varios pasos seleccionando cada vez una porción de la misma.

b) Eliminar porciones de texto: Herramientas → Edición avanzada → Retocar texto. Esta opción es similar a la anterior con la diferencia de que una vez seleccionado un objeto el programa ofrece la posibilidad de seleccionar con el ratón los fragmentos de texto a eliminar.

Figura 3. Eliminar las partes del documento no relevantes

Una vez eliminada toda la información innecesaria en el PDF, guardar el documento modificado en PDF (con otro nombre si no se quiere perder el original) o directamente en TXT. Finalmente debemos generar el fichero en formato TXT para poder realizar el procesamiento lingüístico. Para ello tenemos dos posibilidades:

a) exportar directamente a formato TXT

   Para exportar de PDF a TXT seleccionar a: `Archivo → Exportar → Texto → Texto (normal)`, y guardar con el nombre deseado[8].

b) exportar a TXT utilizando un formato intermedio.

   En este caso se exporta primero al formato XML y a partir de aquí generamos el formato TXT. Para exportar de PDF a XML seleccionar: `Archivo → Exportar → XML 1.0`, y guardar con el nombre deseado. A continuación generar el formato TXT utilizando el siguiente mandato

[8] Es conveniente asignarle un nombre significativo que facilite su recuperación posterior, si hiciera falta.

```
C:\>P:<return>
P:>cd \iula\corpus\corpus\tmpmedic<return> (o al directorio donde se encuentre el fichero a comprobar)
P:\IULA\CORPUS\CORPUS\TMPMEDIC> P:\IULA\CORPUS\UTILS\pdfxml2txt.pl
-file m00956.xml
==> Parsing "m00956.xml" ...
.....................................................................
.....................................................
P:\IULA\CORPUS\CORPUS\TMPMEDIC>
```

El resultado será un fichero con el mismo nombre base pero con extensión ".txt".

El texto que se ha generado en cualquiera de las dos maneras resulta en líneas muy largas por que puede ser necesario activar la opción `Document → Word wrap` cuando se visualice con el EditPlus o un programa similar.

5.1.3 Correcciones finales

El fichero te texto obtenido puede contener errores (palabras enganchadas, errores de conversión, etc.). Para la detección rápida de estos problemas puede ser de utilidad utilizar un procesador de texto (Open Office o MS Word) y su corrector de texto[9] para detectar estos errores. En caso de utilizar una de estas herramientas es de suma importancia que el fichero se mantenga en formato TXT.

Para el resto de tareas de esta etapa es muy conveniente utilizar un editor de texto como por ejemplo el EditPlus [10].

Esta etapa se puede hacer en paralelo con lo que se indica en la sección 7.1. Aquí sólo mostramos indicaciones para correcciones específicas a la conversión PDF → TXT. Cosas a corregir:

A) guiones de separación de palabras.

En casos en que el texto sea del tipo: "... en pacientes con miocardiopatía is-quémica, ..." o bien "... con alteraciones li-geras-moderadas,...". Para ello seleccionamos la opción "Buscar y sustituir" del EditPlus, activamos la opción "Regular expression" y seleccionamos:

Find what: ([a-záéíóúñ])\-([a-záéíóúñ])

Replace with: \1\2

Atención! No hacer nunca una sustitución masiva.

Atención! Tener en cuenta que no siempre un guión indica división de palabra (ver ejemplo anterior: "ligeras-moderadas") y que la codificación del guión no siempre es la misma. Puede ser de utilidad hacer una comprobación con una palabra que veamos en PDF que está dividida y comprobar cómo ha sido convertida. Si resultara que no ha sido convertida es importante repetir la búsqueda/sustitución sustituyendo el guión con el carácter que realmente aparece en el texto. En este caso podría valorarse la posibilidad de hacer una sustitución masiva.

Atención! Según como se haya convertido el texto puede ser necesario adaptar la expresión regular a las necesidades reales del texto

[9] Utilizar la opción Herramientas→Ortografía y Gramática (Open Office) o bien Revisar→ Ortografía y Gramática (MS Word).

[10] Se aconseja utilizar cualquier editor de textos que permita utilizar expresiones regulares. Un programa válido (junto con el EditPlus ya mencionado) es NotePad+ (se puede descargar en http://notepad-plus-plus.org/).

Este proceso es importante porque una palabra mal dividida podría dar lugar a un término no reconocido en la fase de extracción de términos (ver, en el ejemplo anterior (secuencia "miocardiopatía is-quémica"), si no recuperamos el adjetivo "isquémica" éste no podrá ser reconocido como término).

B) Verificar que no haya porciones de texto erróneas o mal convertidas.
Por ejemplo en la secuencia:

```
Conclusiones. Cuando se analiza la reserva contráctil total
del VI con BDD en la gated-SPECT, debe tenerse en cuenta no
sólo el comportamiento de los segmentos con engrosamiento
basal muy deprimido, que es donde se plantea la viabilidad,
sino también el de los segmentos con alteraciones ligeras-
moderadas y el de aquellos en que el engrosamiento empeora.

Candell-Riera J et al. Análisis del engrosamiento
miocárdico con dobutamina mediante gated-SPECT
INTRODUCCIÓN
```

El fragmento indicado en fondo obscuro corresponde a una cabecera de página no eliminada. Por lo tanto deberá ser eliminado.

C) Tratamiento de títulos y párrafos.
En la sección 7.1 se indica que los títulos deben tener una línea en blanco antes y después para su correcta detección. Una manera sencilla de automatizar este proceso (sobre todo si se ha realizado el pro con Acrobat Pro y exportado como texto normal) es duplicar cada salto de línea. Para ello seleccionamos la opción "Buscar y sustituir" del EditPlus, activamos la opción "Regular expression" y seleccionamos:

Find what: \n
Replace with: \n\n

Esta acción tiene la ventaja principal de facilitar el marcaje de los títulos y una ventaja secundaria que es mantener la estructura de párrafos y frases del texto original.

De todas maneras es importante verificar que la acción se haya realizado correctamente sobre todo al pasar de una página a otra (del PDF). Si la segmentación en párrafos es la adecuada, eventuales saltos de línea adicionales no tienen importancia. A continuación se muestra un ejemplo de salto adicional de texto que es necesario corregir:

```
Después de analizar el engrosamiento de cada seg-mento
durante la infusión de BDD, se creo un índice para relacionar
en su conjunto los efectos del aumento y de la disminución
del engrosamiento en los 17 seg-mentos del VI en el
incremento de la FE = 5%. Este índice, denominado índice de
engrosamiento general (IEG), representa la relación entre el
número de seg-mentos con aumento del engrosamiento (A) y el
nú-mero de segmentos con disminución del engrosamien-to (B)
con respecto al área total del VI: (A – B / 17) ×

100. Se valoró el IEG positivo (predominio de los seg-mentos
con aumento del engrosamiento), el IEG nega-tivo (predominio
de los segmentos con disminución del engrosamiento) y el IEG
neutro (mismo número de segmentos con aumento y disminución
del engrosa-miento segmentario).
```

D) En la sección 7.1 (página 24) se indican otras porciones de texto que puede ser necesario eliminar. Debido a las facilidades que ofrece programa el Adobe Acrobat Pro puede suceder que sea conveniente realizar en esta etapa algunas de las operaciones de eliminación. Se recomienda revisar cuidadosamente dicha sección antes de avanzar.

5.2 Conversión a PDF→texto con Terminus

El programa Terminus (entre otras muchas cosas) permite pasar un texto en formato PDF a formato texto. Para esto es imprescindible ser usuario registrado en esta aplicación. Si no tenemos usuario, pedir acceso a Amor Montané.
El módulo *Documentos* ofrece tres posibilidades: la primera permite colgar un único documento al sistema a partir de un documento guardado en el ordenador y la segunda permite colgar varios documentos al mismo tiempo a través de un fichero comprimido.

a) Declarar un nuevo documento: para declarar un único documento, hay que atribuirle un nombre (que sirve para identificarlo posteriormente) e indicar la ubicación (archivo local) y la lengua.

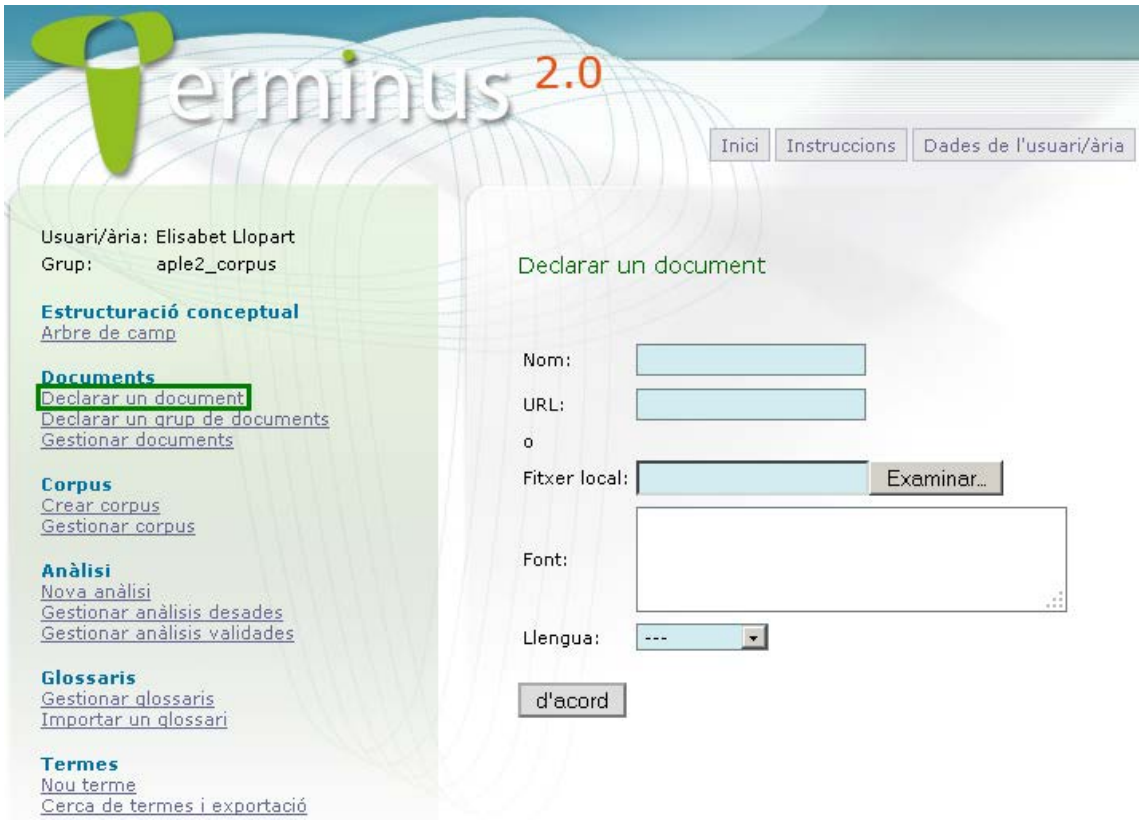


Figura 4. Declarar un nuevo documento con Terminus para su posterior conversión de PDF a texto

b) Declarar un nuevo grupo de documentos: para declarar un grupo de documentos, debemos seleccionar una de las tres opciones disponibles: subir un archivo comprimido en formato ZIP.

Figura 5. Declarar un grupo de documentos con Terminus para su posterior conversión de PDF a texto

Si elegimos la opción del archivo comprimido, Terminus pide que especifiquemos los datos comunes de los documentos que incluye (un nombre identificador y la lengua) y el fichero en cuestión. Hay que tener en cuenta que solo permite cargar archivos en formato ZIP, y es importante que contengan directamente los documentos y no estén organizados en varias carpetas en el interior del archivo comprimido.

Una vez hemos clicado "de acuerdo" para declarar un documento o un conjunto de documentos, para convertir el documento de PDF a TXT nos aparecerá el mensaje de la Figura 4 y debemos hacer clic en "bajar el documento en TXT". A continuación, podremos guardar el archivo en TXT.

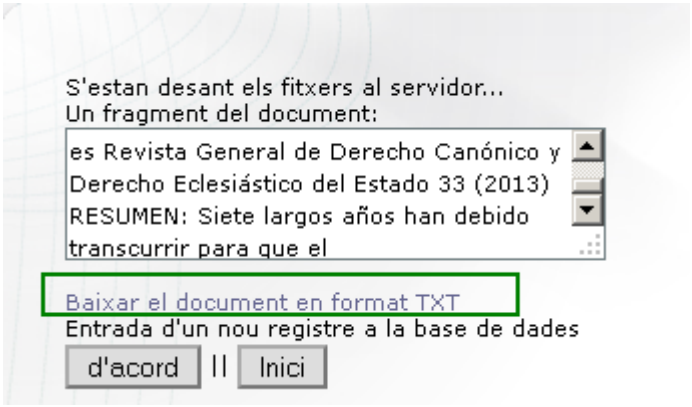


Figura 6. Bajar el documento en TXT

En el caso de declarar un nuevo grupo de documentos con un archivo ZIP, vamos a obtener una lista de los documentos y para descargarlos en formato TXT tendremos que seleccionar uno a uno cada uno de los documentos para bajarlos como se muestra en la Figura 4.

Figura 7. Documentos del archivo ZIP.

6 Base de datos

6.1 Base de datos bibliográfica (DB_1)

Todos los documentos del CT están incluidos en una base de datos que recoge toda la información disponible sobre los documentos incluidos en el CT y al mismo tiempo se encarga de asignar el nombre de fichero definitivo a la/s muestra/s. En este sentido, es importante recalcar que **antes de introducir la información bibliográfica a la base de datos, los textos se habrán procesado para convertirlos en formato .txt** para comprobar que estos textos son realmente utilizables y no presentan problemas que puedan llevar a descartarlos.

6.1.1 Contenido

La base de datos incluye información bibliográfica de cada publicación y, eventualmente, de sus partes. Se estructura en dos partes:

a) datos de la publicación
b) datos del documento

Si fuera necesario, esta organización permite, en el caso de una publicación que tenga varios capítulos, hacer un tratamiento diferenciado a cada uno de ellos. Este sería el caso, por ejemplo, de un libro que contenga una serie de artículos.

Tabla 1. Información asociada a cada uno de los campos de DB_1

#	Atribut	Valors possibles	Camp DB1	VALORS
4	autor		AUTOR/AUTORPART	
3	titulo		TITOL/TITOLPART	
7	lugarPub	Lloc de publicació del doc.	CIUTAT	
5	publisher		EDITORIAL	
6	fechaPub	Data de publicació del doc	ANY(P)	
	Pàgines		PAGINES	Pàgines de la part
	Volum i nombre		VOLUM I NUM	De la revista descrita
	Col·lecció	Nom i número de la col·lecció	SERIE	
18	idType	ISBN, ISSN	ISBN / ISSN	
	URL	Enllaç del recurs electrònic	HIPERVINCLE PUB/ HIPERVINCLE	
	Tipus de document		Tipus Document	ACT - actes de congrés ART - article de revista CACT - participació a congressos publicada en actes CL - capítol de llibre LL – llibre MON - monogràfic de revista PRO – pròleg RE - recurs electrònic RSS - ressenya o notícia bibliogràfica WP - working paper
	Menció de responsabilitat		Tipus Autoria	AUT – Autoria COEDI – Coedició COL – Col·laboració COORD – Coordinació DIR – Direcció EDI – Edició TR – Traducció

#	Atribut	Valors possibles	Camp DB1	VALORS
	Tipus físic	Format físic del document	Tipus Físic	cd: cd-rom e: electrònic (web) m: microfitxa p: paper
15	classTipus	parea, legal, …	ClassTipus	Propis de l'àrea Legal Professional Jurídic Instrumental Teòric Documentació d'usuari
25	Gènere textual	tesis, paper científic, revista divulgació científica, premsa, abstract, resumen, editorial, doc. d'usuari, …	Genere Textual	Tesi doctoral Tesi de llicenciatura Treball de recerca Tesina de màster Paper científic Revista divulgació científica Premsa Resum Editorial Doc. d'usuari
11	Idioma doc i variant lingüística	Codi ISO idioma	Idioma	"eng";"spa";"cat";"fre" Pendent validar codificació variant lingüística
10	status	orig, trad, ptrad, unknown	Estatus Traducció	Original Traducció Possiblement traduït Desconegut
2 14	Dominio i classDom	d, e, i, a, m, o, l / dominiSubdomini	Domini	Arbres de camps corresponents
1	codiCT		Codi_CT	Número Identificador al Corpus de l'IULA. Generació automàtica de l'identificador del document
21	Id mostres	Codi mostres	Id_mostres	Generació automàtica dels valors a partir del número de mostres separats per ","
20	Nro de mostres	#	Num_MOSTRES	
22	Tipus d'àmbit ref. bibliogràfica	full, sample	Text complet	Full Sample

#	Atribut	Valors possibles	Camp DB1	VALORS
17	Origen del doc	Escàner, suport electrònic (pdf, Word, PS, HTML, OO, …)	Format Original	Paper e-PDF e-Word e-PS e-HTML e-altres
24	Codificador		Operador	Rejane Viviana
16	dataInclude	Data inserció al CT	Data actualització	dd/mm/aa
8	numPals	#		Es calcula a UNIX en el moment de generar la capçalera
9	numBytes	#		Es calcula a UNIX en el moment de generar la capçalera
19	idCode	Valor de 18		Informació no codificada. Es treu el valor dels camps ISBN i ISSN
13	idioma paral·lels	En que altres idiomes el tenim (ex c----)		Es determina a part del número de mostra
CAMPS PENDENTS DE CONFIRMAR				
	Formalitat	Informal, neutre, formal		
	Destinatari	Especialista-lego, Especialista-especialista, …		
	Canal comunicatiu	escrit, escrit per ser llegit, oral, …		
	Status autor	Nadiu, no nadiu, desconegut		

6.1.2 Interficie

Para abrir este programa es necesario ejecutar el programa Access (Inicio → Menu → Microsoft Office Access 2007) y abrir la base de datos situada en P:\IULA\CORPUS\corpus_bd.mdb[11]. Para realizar este paso, seguir este procedimiento:

a) buscar la base de datos con el administrador de archivos;
b) hacer doble clic sobre ella;
c) habilitar las macros.

Es importante la acción de habilitar macros. En este sentido, observar el mensaje que se muestra al abrir la base de datos (ver Figura 8). El procedimiento a seguir es hacer clic en "Opciones", que dará lugar a la ventana mostrada en la Figura 9. La activación de la casilla

[11] La extensión mdb corresponde a un fichero Microsoft Access que en un administrador de Archivos (Windows XP o Windows 7) se muestra como un archivo del tipo: "Microsoft Office Access Database").

“Habilitar este contenido” y el posterior clic en “Aceptar” completa la etapa de habilitación de macros.

Figura 8. Habilitación de macros. Primer paso.

Figura 9. Habilitación de macros. Segundo paso

Después de habilitar las macros la interface mostrada debería ser la indicada en la Figura 10.

Figura 10. Interficie de acceso a la BD_1

Como se observa en la Figura 10, en el campo “operador” debemos seleccionar del desplegable el nombre correspondiente al autor de la ficha.

6.1.3 Introducción de nuevos registros

Para introducir los datos del nuevo registro se debe hacer clic en el botón de “nuevo registro” ubicado en la parte inferior de la ventada (ver Figura 10).

Si, por el contrario, introducimos un nuevo artículo de una revista o un capítulo de un libro que hemos registrado previamente en la base de datos debemos clicar el botón “nuevo registro” de la ventana “Dades del document”, no de la ventana “Dades de Publicació” (ver Figura 11). De este modo, la información del apartado “Dades de la publicació” relativo a un libro o revista se mantendrá (título publicación, autor o editor revista o libro, editorial lugar, ISBN-ISSN, etc.), y solo tendremos añadir la información sobre el artículo o capítulo específico (autor del artículo o capítulo, título, etc.).

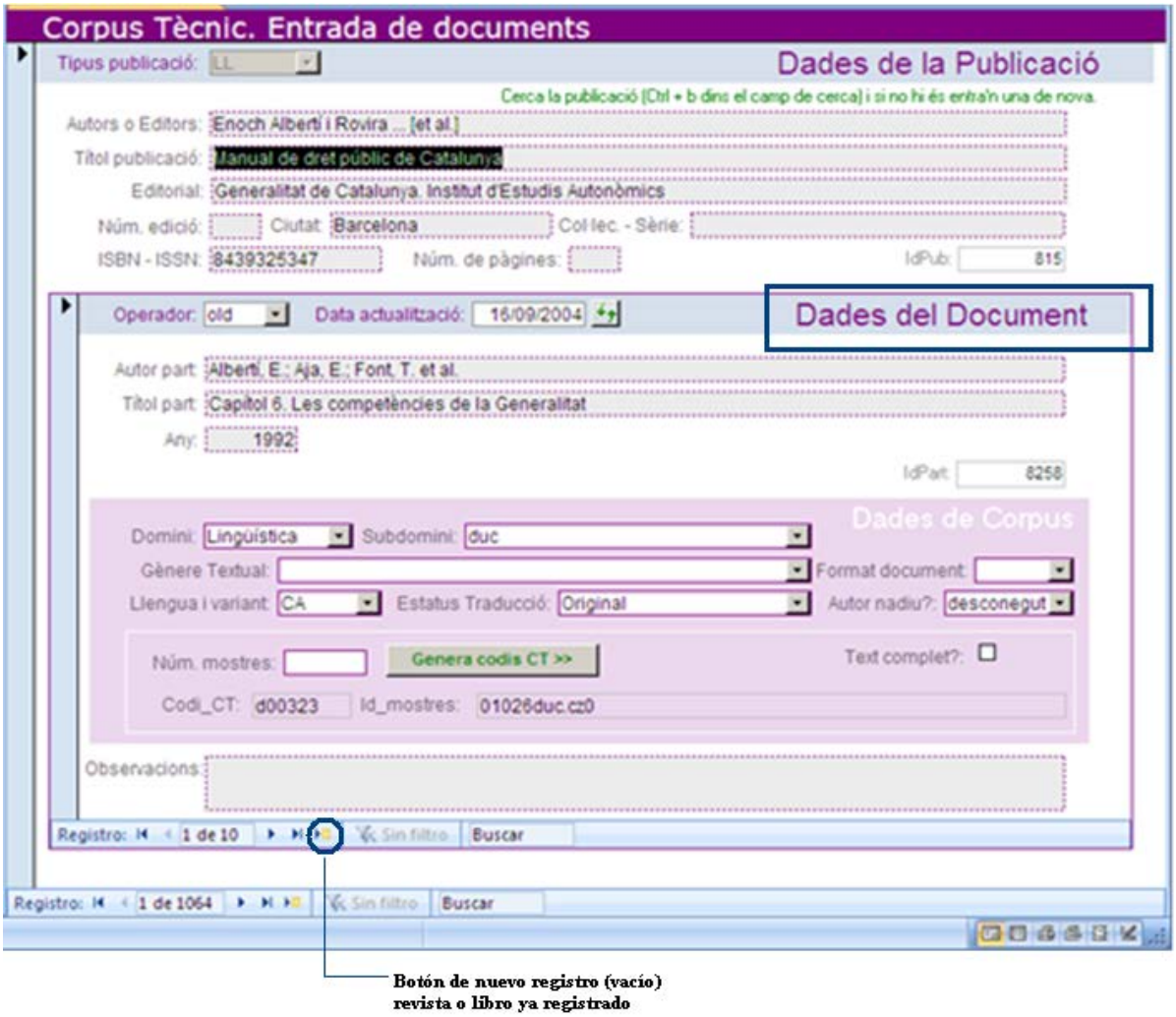


Figura 11. Interficie de acceso a la BD para insertar un artículo de revista o capítulo de libro registrado previamente

En primer lugar es necesario indicar el tipo de publicación. Por esta razón, una vez activado el botón de "nuevo registro" el usuario puede hacer la selección del tipo de publicación en una ventana como la mostrada en la Figura 12.

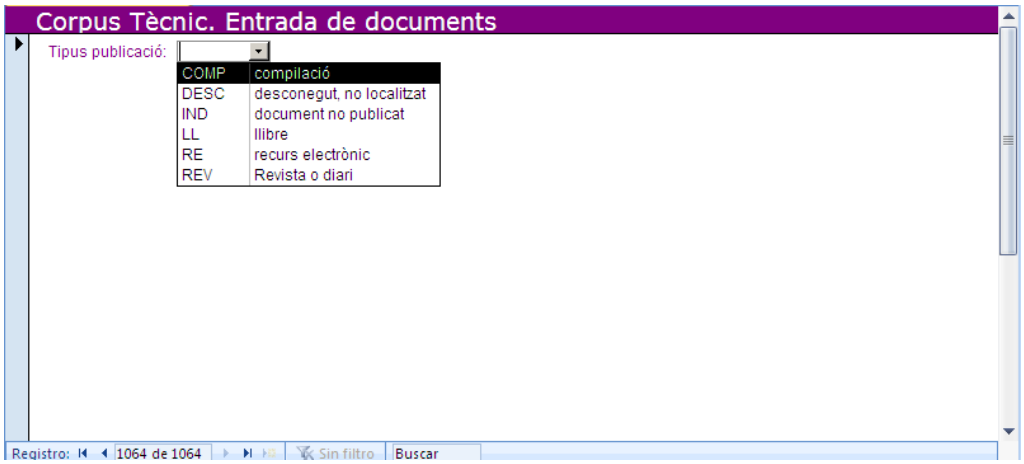


Figura 12. Selección del tipo de documento

Esta ventana ofrece las posibilidades que se indican a continuación:

a) Compilació
b) Desconegut, no localitzat
c) Document no publicat (tesis doctorales y otros documentos no publicados)
d) Llibre
e) Recurs electrònic (páginas web)
f) Revista o diari

En la práctica y teniendo en cuenta las necesidades del proyecto APLE2 las únicas posibilidades a tener en cuenta con c) y f). Las tesis doctorales se considerarán como "Document no publicat" mientras que los artículos de revista com "Revista o Diari". En las dos subsecciones siguientes se describen las peculiaridades de cada uno de estos tipos de texto.

En todos los casos el criterio de citación de autores es el siguiente:

`Apellido/s separadorNombres nombre/s (separadorAutores apellido/s separadorNombres nombre/s)*.`

Donde: `separadorNombres`="," y `separadorAutores` =";".

Los datos que se piden son en su mayoría autoexplicativos y es fundamental para la correcta incorporación del documento a la base de datos textual y posteriormente a "bwanaNet". En particular es importante tener en cuenta el apartado "Num. Mostres" donde normalmente se deberá poner 1 (es decir que todo el texto está incluído en un único fichero, es decir que no se divide en muestras).

También es muy importante el botón "Genera codis CT i dades capçalera >>". Este botón sólo funcionará cuando se haya indicado como mínimo las informaciones "Domini" y "Subdomini". La consecuencia de activar este botón es que se generarán los códigos identificativos de documento ("codi_CT"), de muestra/s ("Id_mostres") y un fichero de texto con todos los datos necesarios para la generación de los ficheros cabecera (ver Anexo I los datos que se recogen en este fichero). El primer código corresponde al identificador de documento de corpus (ej.: e00341, m00951,…) mientras que el segundo es el identificador de muestra/s (ej.: 04903sfi.ez0, 00234mcb.eu0, …). Ambos números son importantes, el primero para generar las cabeceras (Sección 8.1) y el segundo para dar nombre definitivo a los ficheros de la/s muestra/s (Sección 7.4).[12]

Si el documento no tiene muestras sólo será necesario completar la sección "Dades de la Publicació" i la parte de la sección "Dades del Document" con fondo de color ("Dades de Corpus").

Al introducir cualquier información textual (títulos y autores) es necesario tener en cuenta que esta información no puede incluir saltos de línea. También hay que verificar que los guiones i comillas sean las estándar[13]. Cuando se introducen autores hay que eliminar eventuales sub/super índices y asteriscos y otros caracteres especiales tales como: `, ´, “, ”, „, –, ˝, ...

[12] El sistema permite actualizar tantas veces como sea necesario el fichero auxiliar con los datos para la generación de las cabeceras. Una vez generado un código de documento este no se modificará aunque actualicemos los datos de cabecera.

[13] Una buena opción cuando el texto proviene de un pdf (y ocupa más una línea) es copiar dicho texto en un editor de texto (P.ej. EditPlus) y modificarlo para que no incluya saltos de línea.

6.1.3.1 Entrada de un documento del tipo Tesis Doctoral

Al seleccionar el documento del tipo “Document no publicat” la pantalla que se presenta al usuario es la indicada en la Figura 13.

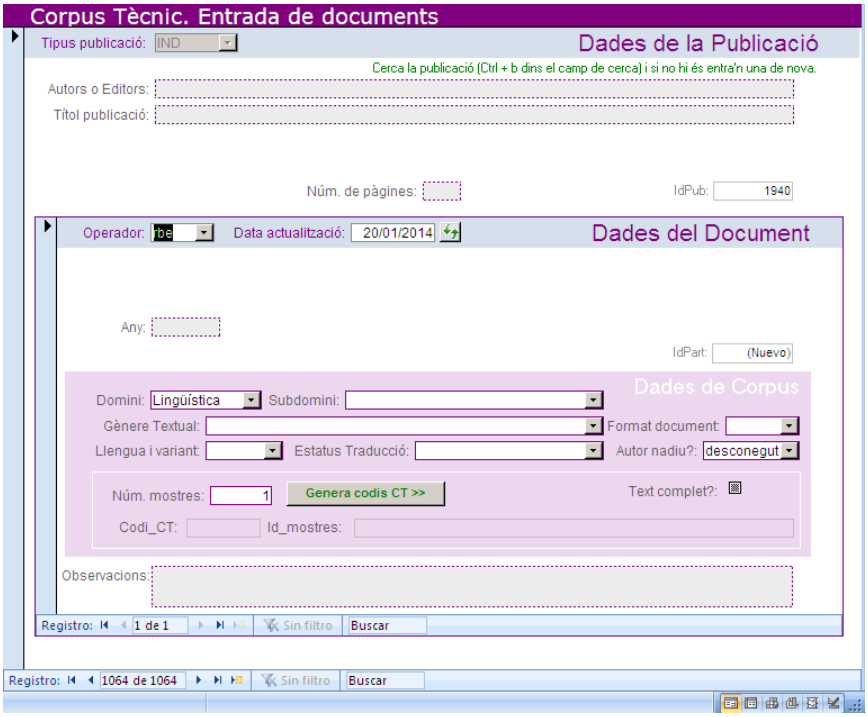


Figura 13. Petición de datos para un "Document no publicat"

6.1.3.2 Entrada de un artículo de revista

Al seleccionar el documento del tipo “Revista o diari” la pantalla que se presenta al usuario es la indicada en la Figura 14.

Figura 14. Petición de datos para una "Revista"

En esta figura se puede apreciar claramente cómo la entrada de datos para este documento se divide en dos partes muy diferenciadas:

a) datos de la publicación/revista como conjunto
b) datos particulares de cada documento/artículo de dicha revista que se incorpora el CT

En este tipo de documentos los datos generales de la revista se realiza una única vez mientras que los datos particulares de cada publicación de la revista se incorporan con cada artículo. Por lo tanto, el primer paso para introducir este tipo de documentos es averiguar si la revista como conjunto está codificada en la base de datos. Si no lo está, pulsar sobre "Insertar nueva revista" (ver Figura 14), introducir los datos de la revista y a continuación se introducir los datos del artículo. Si la revista ya existe en la Base de datos, es necesario buscarla (como se indica más adelante) y pulsar sobre a continuación sobre "Insertar nueva publicación" (ver Figura 14) e introducir los datos del nuevo documento.

Observar en la Figura 14 que el botón a activar para introducir una nueva revista no es lo mismo que introducir un nuevo artículo.

Para buscar una revista en la base de datos se debe proceder de la siguiente manera (ver Figura 15):

a) poner el cursor en el campo "Titol publicació",
b) hacer ^B (<Control> + B),
c) teclear el título de la revista y
d) pulsar la tecla <Return>.

Figura 15. Buscar una revista en la Base de datos

7 Marcaje estructural y preproceso

El marcaje estructural de los textos del IULACT se realiza de manera semiautomática con la intención de minimizar la intervención manual. Para maximizar la eficacia es importante hacer una preparación adecuada del texto (antes de insertar las marcas estructurales). Con este mismo objetivo se ha limitado el número de marcas estructurales a insertar; en la Tabla 2 se indica la lista de etiquetas que está previsto incorporar así como el mecanismo a seguir para su introducción. En algunos casos (`<foreign>` y `<loc>`) la introducción se limita a las que están incluidas en los ficheros de configuración del preproceso. En el caso de las listas (`<list>`) estás se marcarán correctamente sólo si el texto está formateado de una cierta manera. En la sección 7.1 se incluyen las indicaciones necesarias para la inserción de todas las marcas SGML.

Etiqueta	Introducción	Significado
`div1`	automática	Documento
`head`	automática	Título
`p`	automática/manual	Párrafo
`s`	automática/manual	Frase
`list`	automática	Lista
`item`	automática	Ítem (de un lista)

Etiqueta	Introducción	Significado
`foreign`	automática/manual	Palabra en otro idioma
`na`	automática/manual	Fragmento no analizable
`name`	automática/manual	Nombre propio
`abbr`	automática	Abreviatura
`loc`	automática	Locución
`num`	automática	Números arábigos y romanos
`date`	automática	Fechas en formatos diversos

Tabla 2. Etiquetas SGML del CT.

El preproceso así concebido se realiza en tres etapas:

a) Conversión del texto a procesar a texto plano
b) Marcaje estructural del texto convertido
c) Comprobación sintáctica del marcaje estructural
d) Procesamiento final

7.1 Correcciones en el texto a procesar

En esta etapa se trata de ajustar el formato del texto de tal manera que la etapa siguiente (preproceso) pueda operar sin dificultad.

Para preparar el documento TXT, abrimos el documento con el programa EditPlus[14] y tendremos en cuenta los aspectos que se indican a continuación:

- Introducir una marca manualmente implica añadir etiquetas SGML antes y después del texto que queremos marcar. El preproceso marca automáticamente todas estas unidades aunque puede suceder que por distintos motivos algunas de estas unidades no se marquen o no se marquen correctamente por lo puede ser necesario modificar el texto para que sean reconocidas o bien algunas se pueden añadir manualmente en el formato TXT.

  Por ejemplo para marcar como no analizable la fórmula en la frase siguiente: "La ecuación de una recta es y=ax+b." Debemos modificar dicha frase para que quede de la siguiente manera: "La ecuación de una recta es **<na>**y=ax+b**</na>**.". En este caso la marca que se ha introducido es 'na' que corresponde a un fragmento no analizable (ver en la Tabla 2 todas las etiquetas válidas). Más concretamente se han introducido dos marcas: inicio (<na>) y final (</na>). Todas las marcas SGML que se introducen son de este tipo con la única excepción de "…" que se substituirá por una única marca: <gap>.
- Para que un fragmento cualquiera sea marcado como **título** debe tener una **línea en blanco** antes y otra después.
- Para que una secuencia de una o más frases queden agrupadas en un único párrafo, el conjunto debe tener una **línea en blanco** antes y otra después.
- Tratamiento de **tablas y gráficos:** sólo se conserva el **título**[15]. Para ello deben tener el tratamiento correspondiente (indicado en el punto anterior), es decir deben incluir una línea en blanco antes y otra después. Si por razones de distribución del texto en

[14] En EditPlus, para ver el texto completo en la pantalla en vez de líneas de párrafo, en la barra de menú seleccionar Document→Word-wrap

[15] Eventualmente y sólo si el carácter del texto lo justifica (celdas de la tabla con interés terminológico), es posible tratar las columnas de una tabla como una lista. Para ello es necesario separar el texto de las celdas (con el Adobe Acrobat Pro) y formatearlo manualmente como una lista.

el texto original, el título se encontrara en medio de un párrafo se debe mover el fragmento al final de dicho párrafo. Si el texto está formado por dos frases debemos tratar cada una de ellas como si fueran un título.

- Si un **título u oración** está íntegramente en mayúsculas lo cambiamos a **minúscula** y solo mantenemos la mayúscula a inicio de oración y/o cuando sea conveniente (inicio de nombre propio, por ejemplo).
- En aquellos textos donde, después de la conversión, los **párrafos** se mantengan en una única línea es conveniente añadir una **línea en blanco** después de cada línea. De esta manera, aseguramos que el texto convertido tenga el mismo número de párrafos que el texto original. En el caso que el texto incluya secuencias susceptibles de ser marcadas como listas, se deberá tener en cuenta el punto siguiente.
- Para que los **ítems de una lista** puedan ser reconocidos como tales, cada uno de ellos debe estar siempre en **una única línea** (aunque incluyan más de una frase).
- Las **secciones** de bibliografía, anexos, sumario, agradecimientos, abstracts en otra lengua, palabras clave y cualquier otra sección **irrelevante** desde el punto de vista del discurso deben **eliminar**.
- Las **notas a pie de página** sólo se mantienen si contienen texto relevante desde el punto de vista del discurso. **Se deberán mover manualmente al final del texto** de tal manera que queden marcadas como párrafos. Para ello deben tener una línea en blanco antes y otra después. La llamada a pie de página debe ser eliminada. Los **números** de **notas al pie** que aparecen en los párrafos del texto principal, se deben **eliminar**.
- El preproceso marca automáticamente como "no analizables" (`<na>…</na>`) las direcciones de Internet y de correo electrónico. El resto debe marcarse manualmente. En textos de lingüística, las palabras y/o frases que constituyan **ejemplos** deben marcarse como **no analizables**.
- En los **títulos que incluyan** números o letras como enumeración, eliminaremos dichos elementos de numeración.
- El preproceso marca automáticamente **palabras en otra lengua**. Aquellas secuencias que no sean recononocidas como tales deben marcarse manualmente con la marca `<foreign lang="XX">…</foreign>`. Donde XX es el código ISO de la lengua (dos caracteres)[16].
- Las **frases completas en otra lengua** debe llevar la marca de fragmento en otra lengua: `<s lang="XX">…</s>`. Donde XX es el código ISO de la lengua.
- Los **párrafos en otra lengua** deben llevar la marca `<p lang="XX">…</p>`. Donde XX es el código ISO de la lengua. Nótese que no es necesario identificar las frases que componen el párrafo pues el preproceso se encargará de ello.
- El **preproceso** marca automáticamente los **nombres propios** siguiendo una estrategia propia[17]. De todas maneras, puede suceder que no incluya todos los

[16] El preproceso se encarga de marcar algunas palabras (o secuencias de palabras) en otro idioma. Éstas se incluyen en diferentes ficheros según sea la lengua origen: `foreign_DE.txt`, `foreign_EN.txt`, `foreign_ES.txt`, `foreign_FR.txt`, `foreign_IT.txt`, `foreign_LA.txt` que se encuentran en el directorio `P:\IULA\CORPUS\UTILS\Preproceso\conf\`. El formato de todos estos ficheros es siempre el mismo:

```
palabra/s[tabulador]ISOidioma[tabulador]categoría
```

[17] Los nombres propios se detectan en base a la mayúscula inicial y palabras de enlace o conectores. Estos últimos están listados en el fichero `P:\IULA\CORPUS\UTILS\Preproceso\conf\conector.txt`.

nombres propios presentes en el texto (especialmente si está a inicio de frase) o bien un nombre propio podría dividirse en dos o más nombres. Es posible **corregir manualmente** estos problemas añadiendo las marcas `<name>` y `</name>` donde proceda.

- El **preproceso** marca automáticamente todos los **nombres propios, incluso los que están en otra lengua** que reconozca como tales[18].
- Las **siglas** se marcan automáticamente como **nombre propio** (ej: ONU → `<name>ONU</name>`).
- El **preproceso** marca automáticamente las **locuciones**[19]**.**
- Cuando se tiene una referencia del tipo **"año: página"** (ej.: 2000: 12) el espacio interior debe eliminarse para permitir la marcación del conjunto resultante como número (es decir "2000:12" → `<num>2000:12</num>`.
- El **preproceso** marca automáticamente las secuencias de puntos (…) como `<gap>`. Si hay errores de detección de estas unidades debe corregirse manualmente.
- Cuando haya comillas (simples o dobles) el preproceso las tratará correctamente sólo si tienen la codificación correcta (ISO-8859-1 → Windows). Si la codificación fuera otra (ej.: `, ´, ", ", …) se deberán **remplazar manualmente**.

En la primera preparación del texto no hace falta marcar los nombres propios ni las locuciones porque en principio se marcaran automáticamente. Pero si después de alguno de los pasos de procesamiento (marcaje estructural automático (7.2), la comprobación sintáctica del marcaje (7.3) o el análisis morfológico (7.4)) detectamos secuencias no marcadas, corregimos el texto original en TXT y lo volveremos a procesar automáticamente de acuerdo con los puntos 7.2, 7.3 y 7.4.

7.2 Marcaje estructural automático

Esta tarea se desencadena mediante el fichero "preproceso.bat" (P:\IULA\SOFT\BATS\). Este script requiere tres parámetros:

```
preproceso lengua fichero_entrada fichero_salida[20]
```

Si se llama sin parámetros da un pequeño mensaje de ayuda. El parámetro lengua puede tomar los siguientes valores: ca=catalán, es=español, en=inglés. Este script no hace más que personalizar la llamada al programa efectivo de preproceso (que está en `P:\IULA\CORPUS\UTILS\Preproceso`)

Para ejecutar este script se debe abrir una ventana MS-DOS e ir al directorio donde está el texto a analizar.

- [llamada a "Símbolo del sistema"]: se abre una ventana MS-DOS con el prompt "C:\>"
- A continuación proceder de la siguiente manera[21]:

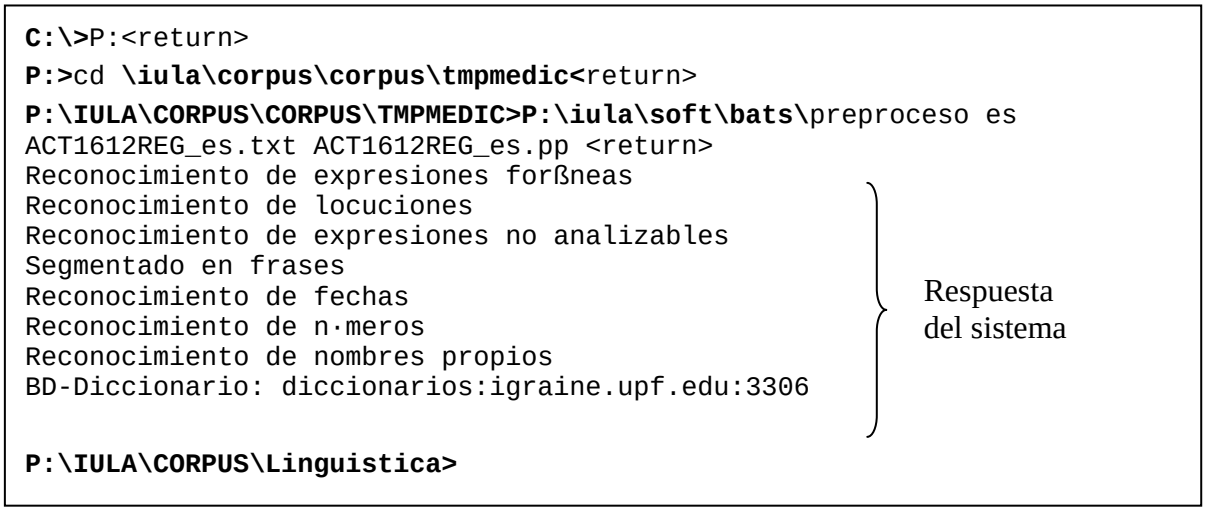
---

[18] El reconocimiento automático de nombres propios multipalabra depende en gran medida de las palabras de enlace o conectores. Estas unidades están incluidas en un fichero específico (`P:\IULA\CORPUS\UTILS\Preproceso\conf\conector.txt`).

[19] Las locuciones que son reconocidas automáticamente se encuentran en los ficheros `locucion_ca.txt`, `locucion_es.txt` y `locucion_en.txt` del directorio `P:\IULA\CORPUS\UTILS\Preproceso\conf\`.

[20] Este fichero se creará una vez se ejecute el script. Si el fichero ya existe lo sobrescribirá con los datos nuevos.

[21] No hacer caso de mensajes del tipo "`Deep recursion on subroutine...`" que puedan aparecer durante la ejecución de este mandato.



En este ejemplo, se aplica el preproceso al texto (en español) `ACT1612REG_es.txt` y el resultado se copia al fichero `ACT1612REG_es.pp`

A continuación se abre el fichero resultado del preproceso, es decir, el archivo con extensión ".pp" con un editor de texto (p. ej.: EditPlus) y se comprueba que el resultado es el esperado. Cosas a verificar:

- Que no haya frases no marcadas. Para ello buscar secuencias: "." + " " + [mayúsculas]. Estas son secuencias potencialmente problemáticas.
- Que todos los ítems de una lista están marcados. Entradas de listas cuyo label no empieza por "1" o "a" son indicadores de posibles ítems de lista no marcados o bien de listas partidas.
- Si detectamos un problema ocasionado por el formato del texto de entrada se debe corregir el texto de entrada con un editor de textos (Editplus o similar), es decir, debemos corregir el texto en formato `.txt`, no las copias para la revisión (formato `.pp`), repetir el preproceso y verificar que se haya resuelto el problema.

7.3 Comprobación sintáctica del marcaje estructural

Esta comprobación sirve para asegurar que la sintaxis de marcas SGML sea la correcta. Es decir, que las etiquetas se usen en la posición prevista en la gramática del documento, que haya texto sólo en los lugares en cuales es correcto que lo haya, etc. En ningún caso nos asegura que la semántica de las etiquetas sea la correcta (es decir que, por ejemplo, un fragmento marcado como título sea realmente un título) ya que esto requiere una comprobación humana.

Para ejecutar este script se debe abrir una ventana MS-DOS e ir al directorio donde está el texto a comprobar. Y, para realizar esta comprobación es necesario modificar el documento que se muestra a continuación (`documento.sgm`), que está guardado P:\IULA\CORPUS. Cada uno de los miembros del proyecto aple2 deberá copiar este documento `.sgm` en la carpeta de dominio correspondiente con su nombre (`nombrefichero.sgm`), para que no interfiera en el procesamiento de los textos. Así pues, cada uno tendrá una copia.

```
<!DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDocIULAb//EN" [
        <!ENTITY     header  SYSTEM  'P:\IULA\CORPUS\header.cmu'>
        <!ENTITY     sample1 SYSTEM  'directorio/muestra'>
]>
<cesDoc version='3.15'>
  &header;
  <!-- N O M   d e l   D O C U M E N T -->
  <text>
     <body>
        &sample1;
     </body>
  </text>
</cesDoc>
```

Este fichero se utilizará para la comprobación sintáctica de todos los textos preprocesados. El único cambio que se debe realizar con el EditPlus es en la línea:

```
<!ENTITY        sample1 SYSTEM  'directorio/muestra'>
```

Se sustituye en cada caso `'directorio/muestra'` por el nombre del fichero que se desea comprobar (p. ej.: “ACT16/ACT1601GUT_es.pp”).

El procedimiento a seguir para comprobar un texto cualquiera es el siguiente:

- Actualizar el documento de comprobación con el fichero a verificar
- [llamada a “Símbolo del sistema”]: se abre una ventana MS-DOS con el prompt “C:\>”
- A continuación proceder de la siguiente manera:

C:\>P:<return>

P:>cd **\iula\corpus\corpus\tmpmedic**<return> (o al directorio donde se encuentre el fichero a comprobar)

P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\corpus\utils\SP1_3_4\bin\nsgmls -s -c P:\iula\corpus\catalog.sgm nombrefichero.sgm<return>

…

A partir de este proceso, obtendremos unos resultados como los siguientes:

```
P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\corpus\utils\SP1_3_4\bin\nsgmls -s -c
P:\iula\corpus\catalog.sgm nombrefichero.sgm
nsgmls:ACT1601GUT_es.pp:599:14:E: there is no attribute "LANG"
nsgmls:ACT1601GUT_es.pp:599:23:E: there is no attribute "POS"
nsgmls:ACT1601GUT_es.pp:599:30:E: element "FOREGIN" undefined
nsgmls:ACT1601GUT_es.pp:599:48:E: end tag for element "FOREIGN" which
is not open
nsgmls:ACT1601GUT_es.pp:601:22:E: end tag for "FOREGIN" omitted, but
its declaration does not permit this
nsgmls:ACT1601GUT_es.pp:599:0: start tag was here
P:\IULA\CORPUS\Linguistica\Actividades 16>
```

Que correspondería a los siguientes errores del texto:

Línea	Contenido
595	No son, <loc pos="C">sin embargo</loc>, del todo ciertos.</s><s>Es

Línea	Contenido
596	absurdo creer que el discurso científico es, o debería ser, objetivo
597	simplemente porque eso sea lo que se le supone a la propia ciencia,
598	pues tal objetividad es el resultado de una decisión que se ha tomado
599	<foregin lang="LA" pos="N5-66">a priori</foreign> sobre las
600	características que tal discurso debe tener y no la consecuencia de su
601	análisis minucioso.</s><s>En el texto que presentamos a continuación,

Es importante tener en cuenta el efecto "cascada". Esto significa que un error en la estructura puede provocar más de un mensaje de error. En el ejemplo superior, observamos que las seis indicaciones de error están concentradas en un entorno pequeño de líneas (599-601) lo cual debe hacer pensar que se trata de un único error. Después de analizar las seis líneas de error parece claro que la más significativa es la tercera (`element "FOREGIN" undefined`) que nos debería hacer pensar que lo pasa es que el nombre de la etiqueta está equivocado (`foregin` en lugar de `foreign`). El resto de líneas de error son consecuencia de este error.

La información detallada sobre los errores se puede observar en el texto en formato `.pp` abierto con el Editplus porque la sintaxis de cada línea de error es siempre la siguiente:

Programa de control	**fichero**	**Nro. de línea**	**Nro. de columna**	**Tipo de error**	**Mensaje**
nsgmls	ACT1601GUT_es.pp	94	81	X	reference to non-existent ID CT

Cuando abrimos el documento generado en formato .pp con el Editplus, podemos detectar los errores si situamos el cursor en la línea y la columna correspondiente al error detectado. Para saber exactamente en qué línea y columna situamos el cursor debemos observar los números que aparecen la parte inferior derecha del programa. En el caso de la Figura 12, se encuentra en la línea 94 y la columna 81. Como vemos en el mensaje de error, el problema es que cuando manualmente hemos especificado que se trataba de una palabra extranjera `<foreign lang="">`, no hemos indicado un código de lengua existente (`CT`). Para corregir el error, tendremos que ir al documento en txt y cambiar el código de lengua por un código correcto, que en este caso sería el latín (`LA`).

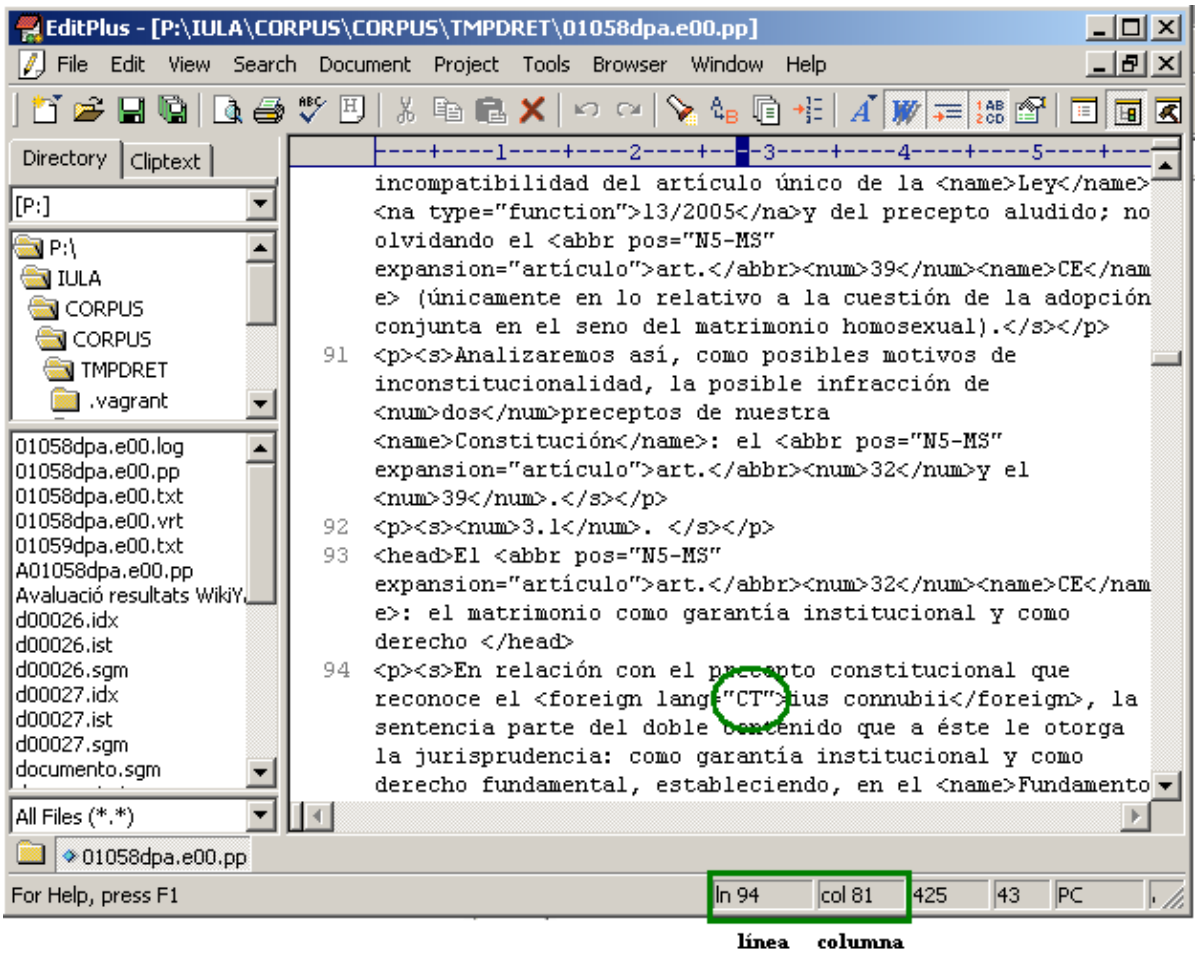


Figura 16. Detectar los errores en el documento en formato .pp con el Editplus.

Una vez detectados los errores con este procedimiento, no se corrigen directamente en el documento .pp, sino que debemos corregir los eventuales errores en el documento original en TXT con Editplus (resultado de la etapa 7.1) y repetir el preproceso (7.2) y esta comprobación. Es importante tener en cuenta los errores tipo cascada ya mencionados (un error en el texto puede desencadenar varios mensajes de error). Así pues, es importante prestar mucha atención a los mensajes de error.

Otra fuente de error habitual se produce cuando un título está formado por varias oraciones. Veamos un ejemplo donde el título que encabeza el documento es el siguiente fragmento:

> Factores de riesgo en pacientes con coccidioidomicosis diseminada fatal. Estudio de casos y controles.
>
> Resumen
>
> …..

Al hacer la comprobación sintáctica nos daría el siguiente mensaje:

```
P:\iula\corpus\utils\SP1_3_4\bin\nsgmls:pruebaJVP.pp:2:2:E:
document type does not allow element "P" here
```

Esto se debe a que el marcaje automático de este fragmento fue el siguiente:

```
<p><s>Factores de riesgo en pacientes con coccidioidomicosis
diseminada fatal.</s><s>Estudio de casos y controles </s></p>
<head>Resumen </head >
……
```

Observar que el texto empezaría con un párrafo, cosa que no está permitida por la DTD. La solución a este problema es simplemente dividir el título en varias frases de la siguiente manera:

```
Factores de riesgo en pacientes con coccidioidomicosis diseminada fatal

Estudio de casos y controles

Resumen

…..
```

Observar que se han eliminado los puntos de final de frase y que hay una línea en blanco entre las frases del título. De esta manera, el texto se procesará sin indicación de (este) error.

7.4 Análisis morfológico y desambiguación

Todos los textos que se integran el corpus IULA deben procesarse lingüísticamente. El propósito es doble: por un lado la detección de problemas remanentes de preproceso (sección 7.2) no detectables sin una lectura del texto y por otro la integración efectiva de los textos en el corpus.

El procedimiento a seguir para completar este análisis es el siguiente

- [llamada a "Símbolo del sistema"]: se abre una ventana MS-DOS con el prompt "C:\>"

- A continuación proceder de la siguiente manera:

```
C:\>P:<return>
P:>cd \iula\corpus\corpus\tmpmedic<return> (o al directorio donde se encuentre el fichero a comprobar)
P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\soft\bats\procesaFullES_inCT
standard unknown fichero.txt fichero.vrt iso-8859-1
        reading parameters ...
Segmentado en frases
BD-Diccionario: TreeTaggerDB:igraine.upf.edu:3306
Reconocimiento de expresiones forßneas
Reconocimiento de locuciones
Reconocimiento de expresiones no analizables
Reconocimiento de fechas
Reconocimiento de n·meros
Reconocimiento de nombres propios
BD-Diccionario: TreeTaggerDB:igraine.upf.edu:3306
        tagging ...
         finished.
P:\IULA\CORPUS\CORPUS\TMPMEDIC>
```

Es importante notar que fichero de entrada (parámetro 3) es el nombre del fichero de texto utilizado hasta ahora mientras que el parámetro 4 es el nombre de la muestra asignada por la base de datos documental (sección 6.1.3).

El fichero resultante será el texto de entrada pero verticalizado. Es decir, organizado con un token por línea y con la información lingüística asociada. La información se distribuye según el siguiente formato cuando se trata de un elemento textual:

```
Nro_Token    Tipo_Token    Forma    Inicio/Final/Continuación    lema\POS
```

y el siguiente formato cuando se trata de una marca estructural del texto:

```
##           TAG           Token_SGML
```

Siguiendo esta convención un fragmento de texto real tiene el siguiente aspecto:

##	TAG	<div1>		
##	TAG	<head>		
1	TOK	El	BOS	el\AMS
2	TOK	lenguaje		lenguaje\N5-MS
3	TOK	de		de\P
4	TOK	las		el\AFP
5	TOK	ciencias		ciencia\N5-FP
6	TOK	de		de\P
7	TOK	la		el\AFS
8	TOK	salud	EOS	salud\N5-FS
##	TAG	</head>		
##	TAG	<head>		
10	TOK	Introducción	BOS	introducción\N5-FS
tf	TOK		EOS	=\DELS
##	TAG	</head>		
##	TAG	<p>		
##	TAG	<s>		
13	TOK	En	BOS	en\P
14	TOK	la		el\AFS
15	TOK	recta		recta\N5-FS
16	TOK	final		final\JQ--6S
17	TOK	de		de\P
18	TOK	esta		este\ED--FS
…				
1915	TOK	yo		pr\REO-16S
1916	TOK	considere		considerar\VJR6S-
1917	TOK	que		que\C
1918	TOK	el		el\AMS
1919	TOK	lexicógrafo		lexicógrafo\N5-MS
1920	TOK	especializado		especializado\JQ--MS
1921	TOK	es		ser\VDR3S-
1922	TOK	una		uno\E6--FS
1923	TOK	mutación		mutación\N5-FS
1924	TOK	rarita		unknown\JQ--FS
---	DLD	,		=\DELIM
1925	TOK	o		o\C
1926	TOK	es		ser\VDR3S-
1927	TOK	una		uno\E6--FS

1928 TOK especie especie\N5-FS

…

El procedimiento de análisis funciona de tal manera que cuando una palabra no está incluida en el diccionario de procesamiento el lema queda indicado como ‘unknown’. Esto puede suceder tanto porque la palabra realmente no esté incluida en el diccionario como porque hay un error en el formato del texto (palabras enganchadas, error de tipeo, error en la conversión a txt, etc.) que produce una palabra desconocida.

Aprovecharemos esta circunstancia para realizar una última corrección al texto y producir una lista de palabras desconocidas y por lo tanto candidatas a ser incluidas en el diccionario de procesamiento.

El procedimiento a seguir para detectar estas palabras consiste en abrir el texto verticalizado con un editor de texto (ej.: EditPlus) y buscar la cadena ‘unknown’. Cuando nos encontramos con en uno de estos casos debemos decidir cuál es la situación en que nos encontramos y corregir el texto o incluirla en una lista de palabras candidatas a ser integradas en el diccionario de procesamiento. Si ha sido necesario corregir el texto es imprescindible repetir los pasos 7.2 a 7.4.

Algunos ejemplos de los errores que podemos detectar con la búsqueda de palabras marcadas como ‘unknown’ son:

- Unidades que forman parte de un título (head) en mayúsculas que aparece en dos líneas separadas. Para solucionar estos casos, situar todas las unidades del título en una misma línea.
- Las unidades que siguen a una enumeración. Después de un número de título o bien de una nota, la palabra que sigue puede que aparezca como “unknown”. Entonces, dejar una línea en blanco entre el número y la frase.

  `1.`

  `A MODO DE INTRODUCCIÓN: LA NO DISCRIMINACIÓN POR RAZÓN DE SEXO EN EL MATRIMONIO`
- Iniciales que acompañan al nombre de un autor: BERCOVITZ RODRÍGUEZ CANO, **R.** Si marcamos todo el nombre como nombre propio (<name>), ya no tendremos este error: `<name>BERCOVITZ RODRÍGUEZ CANO, R.</name>` Podemos seguir el mimo procedimiento en el caso de otros nombres propios (publicaciones, artículos, etc.) que de error.
- Palabras que estén mal escritas o bien que no estén en el diccionario de procesamiento. Veamos un ejemplo:

##	TAG	<s>		
3914	TOK	La	BOS	el\AFS
3915	TOK	pérdida		pérdida\N5-FS
3916	TOK	de		de\P
3917	TOK	la		el\AFS
3918	TOK	eumetría		unknown\N5-FS
---	DLD	,		=\DELIM
3919	TOK	la		el\AFS
3920	TOK	euergia		unknown\N5-FS
3921	TOK	y		y\C
3922	TOK	la		el\AFS
3923	TOK	eutaxia		unknown\N5-FS
3924	TOK	se		pr\R6EZZZZ
3925	TOK	manifiestan		manifestar\VDR3P-
3926	TOK	en		en\P

En este fragmento hay dos ocurrencias de palabras cuyo lema es 'unknown': "euergia" y "eutaxia". Probablemente en el primer caso se trate de un error de tipeo mientras que el segundo parece ser un palabra que falta en el diccionario de procesamiento.

Las palabras candidatas a ser introducidas en el diccionario de procesamiento deben incluirse en el fichero de texto que se encuentra en de cada uno de los directorios y recibe el nombre "afegirdicc.txt" con el formato siguiente:

```
forma(→POS lema)+
```

Ej.

```
Forma1    POS1 lema1
Forma2    POS1 lema1     POS2 lema2
```

Es decir deben indicarse todas las formas posibles y para cada forma deberán indicarse todas las parejas POS-lema. La información POS-lema debe separarse internamente con un espacio mientras que la forma y cada uno de los pares POS-lema debe separarse con un tabulador. Un ejemplo de este fichero es el siguiente:

subárea	N5-FS subárea
subáreas	N5-FP subárea
subespecialista	N5-6S subespecialista
subespecialistas	N5-6P subespecialista
subpoblación	N5-FS subpoblación
subpoblaciones	N5-FP subpoblación
subunidad	N5-FS subunidad
subunidades	N5-FP subunidad
sudoración	N5-FS sudoración
sudoraciones	N5-FP sudoración
tibial	JQ--6S tibial
tibiales	JQ--6P tibial

Una vez que se han resuelto todas las cuestiones relativas al léxico se debe realizar una ejecución con estos parámetros para eliminar la etiqueta "unknown" correspondiente a neologismos, no en los casos de errores y préstamos que ya habremos solventado. De esta manera, en vez de la etiqueta "unknown" indicará una propuesta de lema y así estas unidades se añadirán en el archivo con el nombre "afegirdicc.txt" para completar el diccionario interno.

Como se observa en el script, indicamos el nombre del archivo en .txt y, en cambio, el archivo que vamos a crear debe recibir el nombre correspondiente al ID de la muestra (ej.: 01078dpc.e00). En este caso, no se debe indicar la extensión del documento, es decir, no añadimos .vrt en el nombre.

La información sobre el nombre que recibe el ID de la muestra se puede comprobar en la base de datos en Acces, como también en el documento .conf que recibe el nombre del Codi CT (ej.: d00372.conf). El documento .conf se crea automáticamente en el directorio de nuestro dominio cuando registramos la referencia bibliográfica del artículo.

```
P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\soft\bats\procesaFullES_inCT
standard ct fichero.txt ID_muestra iso-8859-1
```

```
P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\soft\bats\procesaFullES_inCT
standard ct fichero.txt fichero.vrt iso-8859-1
```

Ejemplo:

```
P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\soft\bats\procesaFullES_inCT
standard ct d00372.txt 01078dpc.e00 iso-8859-1
```

```
P:\IULA\CORPUS\CORPUS\TMPMEDIC>P:\iula\soft\bats\procesaFullES_inCT
standard ct d00372.txt d00372.vrt iso-8859-1
```

```
1  date = 01/10/2014
2  operador = csh
3  <publicationInfo>
4          type = REV
5          titlePub = Academia Sevillana del Notariado
6          editorial = Editoriales de Derecho Reunidas. EDERSA
7          pubPlace = Sevilla
8  </publicationInfo>
9  <docInfo>
10         authorDoc = Rodríguez Benot, Andrés
11         titleDoc = La administración de la herencia en las sucesiones
12         year = 2009
13         pages = 253-304
14         volume = 19
15         domain = d
16         subDomain = dpc
17         genre = 104
18         langDoc = es-pe
19         docFormat = e-TXT
20         langStatus = orig
21         authorStatus = s
22         textComplet = 1
23         samplesNum = 1
24         samplesId = 01078dpc.e00
25 </docInfo>
```

Figura 17. Si abrimos con el EdiPlus el documento .conf correspondiente al documento en cuestión, podemos ver las referencias bibliográficas y el ID de la muestra.

En primer lugar se ejecuta el script para crear un documento con el nombre del ID_muestra. Y, en segundo lugar, procedemos con el siguiente script, que elimina la etiqueta “unknown” del documento verticalizado .vrt. Si abrimos de nueva la versión actualizada del documento .vrt con el Editplus y buscamos (search) los “unknown”, podremos comprobar que ya no aparecen porque ha sustituido la etiqueta por la unidad documentada en el texto como propuesta de lema.[22]

En este momento también es necesario eliminar el fichero temporal creado en la sección 7.2 para hacer la comprobación del marcaje estructural.

[22] En este punto se ejecutan dos scripts, una para crear un documento verticalizado con el nombre del ID_muestra y otro como fichero .vrt (por ejemplo, d00372.vrt) porque en el primer caso la etiqueta *unknown* se elimina del documento que va al Corpus IULA y en el otro caso se aplica para la extracción que va a la Base de datos.

8 Creación de la cabecera y otros ficheros auxiliares

Todos los documentos del CT incluyen unos ficheros de "cabecera" que contienen informaciones (bibliográficas y estadísticas) sobre el documento. El procedimiento a seguir para generar estos ficheros se basa en aprovechar el fichero auxiliar generado en la etapa de introducción de los datos bibliográficos en la base de datos (sección 6.1.3).

Además el documento antes de ser incorporado al CT pasa por una serie de procesos de verificación con el propósito de incorporar documento sólo documentos 100% correctos (tanto desde el punto de vista de la sintaxis de las marcas SGML como del fichero a ingresar en la Base de datos textual). Algunas de estas etapas ya se han realizado mientras que otras sólo se pueden realizar una vez que se hayan creado los ficheros de cabecera. Se describen a continuación estas etapas de comprobación y control.

8.1 Creación de cabeceras

Para generar los ficheros de cabecera de cualquier documento del CT se debe ir al directorio donde se encuentra este fichero e invocar el programa creaHeader.pl como se indica a continuación:

El nombre que debemos indicar para crear las cabeceras se corresponde con el nombre del codiCT que recibe cuando ingresamos la referencia bibliográfica en las base de datos.

```
P:\IULA\CORPUS\CORPUS\TMPMEDIC>perl P:\IULA\CORPUS\UTILS\creaHeader.pl
-d codiCT
```

```
C:\>P:<return>
P:>cd \iula\corpus\corpus\tmpmedic<return>  (o al directorio donde se encuentre el fichero a comprobar)
P:\IULA\CORPUS\CORPUS\TMPMEDIC>perl P:\IULA\CORPUS\UTILS\creaHeader.pl –d m00956
Procesa datos del documento en m00956.conf ...
Generación de las cabeceras
      - m00956.idx ...
      - m00956.isgm ...
      - m00956.ist ...
P:\IULA\CORPUS\CORPUS\TMPMEDIC>
```

Ejemplo:

Todos los ficheros de cabecera se generan en un directorio `CTheaders` dentro del directorio donde se encuentra el fichero con los datos de la cabecera. En el ejemplo mostrado más arriba para el documento m00956 se deberán generar los ficheros `m00956.idx`, `m00956.ist` y `m00956.sgm` en el directorio `P:\IULA\CORPUS\CORPUS\TMPMEDIC\CTheaders.`

Atención!

A partir de este punto,

la documentación de este apartado es incompleta y pendiente de actualización.

8.2 Entrada simulada a la Base de datos textual

8.3 Entrada efectiva a la Base de datos textual

Para cada documento a incorporar al CT el procedimiento a seguir es el siguiente:

- Generar el fichero para generar los datos en el formato final del CWB
- Generar el fichero para comprobar los datos en el formato final del CWB
- Generar el fichero para simular la incorporación al CWB

ANEXO I

Datos recogidos para la generación de los ficheros cabecera

Al pulsar sobre el botón “Genera codis CT i dades capçalera” se genera, en el directorio de trabajo, un fichero con el nombre `idDoc + .conf` (ej. `m00105.conf`) y el formato indicado más abajo. Si un dato no existe en la Base de Datos, se omite la línea correspondiente.

Este fichero es utilizado por el programa creaHeader.pl para generar los ficheros efectivos de cabecera del documento.

```
date = [data creació de la capçalera]
operador = [código en la base de datos del codificador]
<publicationInfo>
      ## Dades de la publicació
      type = [Tipus publicació]
      authorBook = [Autors o Editors]
      titlePub = [Titòl publicaciò]
      editorial = [Editorial]
      pubPlace = [Ciutat]
      isbn = [ISBN - ISSN]
</publicationInfo>
<docInfo>
      ## Dades del Document
      authorDoc = [Autor part]
      titleDoc = [Títol part]
      year = [Any]
      pages = [Pàgines]
      volume = [Volum i núm]
      domain = [Domini]
      subDomain = [Subdomini]
      genre = [Gènere Textual]
      langDoc = [Llengua i variant]
      langStatus = [Estatus Traducció]
      samplesNum = [Núm. mostres]
      samplesId = [id_mostre] ## lista de muestras separadas por ‘,’
</docInfo>
```

ANEXO II

Content for the file “contenido.txt”

	Col1[23]	Col2	Col3	Col4	Col5
Row1	user	name			
Row2	publicacion	IULA	Title of the book	ISBN	date
Row3	doc	File name	language[24]	Authors[25]	Title of the article
…	doc	File name	Language[1]	Authors[2]	Title of the article
…	doc	File name	Language[1]	Authors[2]	Title of the article
…	…	…	…	…	…

Temas pendientes del preproceso

- Problemas indicados en P:\IULA\CORPUS\:\Linguistica\problemasProcesamiento.doc
- “2000:12” → `<num>2000</num>:<num>12</num>`.
- Estrategia para las abreviaturas a final de frase
- Números con letras (ES, CA y EN)
- Fechas en inglés
- Nombres propios en inglés
- Convertir internamente todo a UTF-8 y exportar en …
- BD de nombres propios (actualización y utilización)
- Llamar a un Web service (preproceso, consulta de corpus) desde Perl

[23] El contenido de esta columna debe ser literal.

[24] Valid codes are ‘es’, ‘ca’ and ‘en’.

[25] Multiple authors must separated by a semicolon (;)