10 anys del Corpus de l'IULA

M. Teresa Cabré, Carme Bach, Jorge Vivaldi

Papers de l'IULA. Sèrie Informes, 44

Barcelona
Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
2006

BIBLIOTECA DE CATALUNYA. DADES CIP

Cabré i Castellví, M. Teresa (Maria Teresa)

10 anys del Corpus de l'IULA. - (Papers de l'IULA. Sèrie informes ; 44)
Text en català, resum en català i anglès
I. Bach, Carme II. Vivaldi, J. (Jorge) III. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada IV. Títol V. Títol: Deu anys del Corpus de l'IULA VI. Col·lecció: Papers de l'IULA. Sèrie informes ; 44
1. Corpus Tècnic Especialitzat (Projecte) - Història 2. Lexicografia - Investigació - Catalunya - Història 3. Llengües d'especialitat - Investigació - Catalunya - Història
800.3:681.3

Direcció de les Publicacions de l'IULA: Mercè Lorente Casafont

Coordinació de les Publicacions de l'IULA: Lluís Codina, Gemma Martínez

Primera edició: octubre de 2006 (versió electrònica)

Pl. de la Mercè, 10-12

08002 Barcelona

Dipòsit legal: B-48.882-2006

10 anys del Corpus de l'IULA

M. Teresa Cabré
teresa.cabre@upf.edu

Carme Bach
carme.bach@upf.edu

Jorge Vivaldi
jorge.vivaldi@upf.edu

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

En aquest paper, es presenten els criteris de treball que s'han seguit durant els 10 anys en què s'ha anat constituint el corpus de l'IULA. S'exposa l'estat de les dades del corpus, els recursos lèxics utilitzats per al tractament de les dades (diccionaris i etiquetaris) i les eines constituïdes o adaptades. Es dedica especial atenció a la documentació de la cadena de treball de processament del corpus, des de l'adquisició dels textos en format electrònic fins a la seva incorporació definitiva al corpus.

In this paper, we present the work criteria taken into account in the development of the IULA's corpus during the last 10 years. We show the state of the corpus data, the lexical resources used for the data treatment (dictionaries and taggers), as well as the tools we have developed or adapted. We pay special attention to the description of the whole corpus processing steps, from the electronic text acquisition to their definitive addition to our corpus.

Índex

1 Presentació

L'Institut Universitari de Lingüística Aplicada és un centre de la Universitat Pompeu Fabra de Barcelona destinat a la investigació i a la formació de postgrau. L'IULA s'organitza en 3 grups d'investigació: el Grup IULATERM (que compren tres unitats interrelacionades — la de lèxic, terminologia i discurs; la de lingüística computacional i enginyeria lingüística; i la de ciències de la documentació), el grup INFOLEX i el grup UVAL i tres laboratoris: OBNEO (Observatori de Neologia), LATEL (Laboratori de Tecnologies Lingüístiques) i Forensic Lab (Laboratori de Lingüística Forense).

Des de 1993 i fins al 2004, el projecte Corpus Tècnic (CT), dirigit per M. Teresa Cabré Castellví, ha estat el projecte comú d'investigació en què han participat tots els membres de l'IULA.[1] Recopila textos contemporanis escrits en cinc llengües diferents (català, castellà, anglès, francès i alemany) de les àrees d'especialitat de l'economia, el dret, el medi ambient, la medicina i la informàtica.[2] El corpus conté documents paral·lels recopilats per facilitar estudis de traducció. Alhora, el corpus multilingüe de l'IULA té un subcorpus de llengua general en català i castellà, extret de la premsa de gran difusió, constituït com a corpus contrastiu.

Els principals objectius d'aquest corpus són, d'una banda, facilitar l'anàlisi de les dades lingüístiques per tal de poder establir les lleis que regeixen el comportament de cada llengua en cada àrea i, d'altra banda, desenvolupar eines d'interès per al processament del llenguatge natural.

Els seus principals destinataris són els investigadors i tots aquells usuaris que vulguin fer consultes sobre els àmbits d'especialitat que recull.

Sobre la base del corpus, s'han fet estudis de caràcter terminològic, discursiu, morfològic, sintàctic, computacional, neològic o traductològic. A partir del CT, l'IULA ha desenvolupat una sèrie d'eines d'explotació. Una mostra d'aquestes eines són l'extractor automàtic de neologia (*SEXTAN*), el detector automàtic de terminologia (*MERCEDES*), l'extractor automàtic de terminologia (*YATE*) i l'alineador de textos (*ALINEA*).

De fet, aquest corpus ha estat i és encara el suport principal de las activitats d'investigació i docència del nostre institut.

L'eina que permet interrogar les dades el corpus mitjançant internet és *bwanaNet*, a la qual es pot accedir des de la pàgina principal de la web de l'IULA (www.iula.upf.edu) a l'apartat anomenat "recursos IULA".

[1] El projecte "Llenguatges d'Especialitat. Corpus Multilingüe" va ser finançat per l'Agència Catalana de la Recerca (CIRIT), projecte nún, CS93-4009 i pel Contracte programa del Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya 1996-2000.

[2] El corpus conté també textos especialitzats d'altres matèries utilitzats per a tesis i treballs de recerca dels estudiants, que van alimentant el corpus de forma més o menys continua.

2 Estat de la qüestió, dades

L'estat de les dades a data de tancament del CT (2004) és el següent:

A.1. Les ocurrències en milers per àmbit temàtic i per llengua que actualment componen el corpus de l'IULA es sintetitzen en el quadre següent:

Àrea	Català	Castellà	Anglès	Francès	Alemany
Dret	1463	2085	431	44	16
Economia	1776	1091	274	78	27
Medi ambient	1506	1062	599	230	429
Informàtica	655	1227	338	194	83
Medicina*	2619	4077	1555	27	198
Total . . .	8019	9542	3197	573	753

1. Nombre d'ocurrències per llengua i per àmbit

*El corpus de medicina inclou també el corpus de genòmica, constituït pel grup IULATERM:

	Català	**Castellà**	**Anglès**
Genòmica	945	1447	1119

A.2. D'acord amb els objectius inicials, una part d'aquest corpus està integrat per textos paral·lels. Actualment les dades pel que fa al corpus paral·lel de les parelles lingüístiques més significatives català-castellà, català-anglès, castellà-anglès són les que es recullen en el quadre 2.

Àrea	**Català-Castellà**	**Català-Anglès**	**Castellà-Anglès**
Dret	460	12	57
Economia	600	250	283
Medi ambient	214	213	144
Medicina	118	40	640
Informàtica	28	-	300
Total . . .	1.420	515	1424

Àrea	**Català-Castellà**	**Castellà-Anglès**
Genòmica	10	515

2. Ocurrències de corpus paral·lels per àmbit i parelles de llengües

A.3. Finalment, el corpus multilingüe de l'IULA compta també amb un subcorpus de llengua general extret de la premsa de gran difusió. Pel que fa al català, les dades s'han extret del diari electrònic *Avui*. Per al castellà, s'ha utilitzat *El País* digital, *La Vanguardia* digital, els textos del projecte Corpus-92 (constituït per exàmens de selectivitat) i els textos del Corpus de Xile Les dades actuals referents a aquest corpus de contrast es mostren en el quadre 3.

Àrea	**Català**	**Castellà**	Total
general	1526	3230	4756

3. Nombre d'ocurrències de llengua general

3 Recursos lèxics

3.1 Diccionaris

Per al processament del CT s'utilitzen dos diccionaris modulars: un per al català i un per al castellà. Aquests diccionaris interactuen amb els analitzadors, de manera que per a cada text, cada paraula rep un lema i una etiqueta gramatical, d'acord amb els etiquetaris de l'IULA.

3.1.1 Diccionari electrònic català

El diccionari català de l'IULA inclou dades del *Diccionari de la llengua catalana* (1983) i del *Diccionari de la llengua catalana* (1993). Aquestes dades s'han obtingut a través de convenis amb *Enciclopèdia Catalana*. També inclou dades del *Diccionari de la llengua catalana* (1995) de l'*Institut d'Estudis Catalans,* amb qui s'ha firmat un conveni per a la llicència d'ús, i de l'Observatori de Neologia de l'IULA-UPF.

Tot i haver partit d'aquests tres diccionaris, el diccionari català de l'IULA s'ha modificat en funció dels criteris de lematització i processament del corpus de l'IULA.

S'ha complementat també amb diccionaris de neologia del mateix CT i de l'Observatori de Neologia de l'IULA-UPF, és a dir, conté dades del lèxic general i especialitzat.

Consta de 97.700 entrades amb informació morfològica. Concretament per a cada entrada s'indica la categoria gramatical i el paradigma de flexió. El seu volum actual és de 43 Mbytes. El codi de caràcters del diccionari és el proposat en la norma ISO8859.

El diccionari és al servidor morgana.upf.es a la ruta:

/usr3/iula/soft/CATMORF/LEXIC/Dlc+Lx93+Lx95/Pl+Ql

i es compon de vuit submòduls en forma d'un conjunt de fets Prolog:

a) un diccionari base d'entrades nominals: nmenys.pl
b)un diccionari amb els lemes verbals de la primera conjugació: verbs_1a.pl:[3]
c) un diccionari amb els lemes verbals de la segona conjugació: verbs_2a.pl
d) un diccionari amb els lemes verbals de la tercera conjugació: verbs_3a.pl
e) un diccionari de neologia nominal trobada al CT de l'IULA: neologia_ct_noms_iula.pl
f) un diccionari de neologia verbal trobada al CT de l'IULA: neologia_ct_verbs_iula.pl
g) un diccionari de neologia nominal trobada al corpus de premsa de l'IULA: neologia_pr_noms_iula.pl

[3]Al directori Dlc+Lx93+Lx95/Pl+Ql només hi ha l'arxiu compilat. Els arxius font són a /usr3/iula/soft/CATMORF/LEXIC/Diec/Pl+Ql

h) un diccionari de neologia verbal trobada al corpus de premsa de l'IULA: neologia_pr_verbs_iula.pl

En el manual d'usuari de *CATMORF* (analitzador morfològic del català de baix nivell), s'inclou un apartat en què s'especifica el procés de manteniment i actualització de la base de dades del diccionari, que reproduïm a l'annex 6.

Estat de la qüestió de l'actualització del diccionari de neologia:

Neologia del CT	
Nombre d'entrades introduïdes	
nominals:	9.936
verbals:	406
Neologia del corpus de contrast (premsa)	
Nombre d'entrades introduïdes	
nominals:	665
verbals:	34

3.1.2 Diccionari electrònic castellà

El diccionari castellà de l'IULA inclou dades del *Diccionario Actual de la Lengua Española* (Vox- Bibliograf) amb qui s'ha firmat un conveni per a la llicència d'ús.

Tot i haver partit d'aquest diccionari, el diccionari castellà de l'IULA s'ha modificat en funció dels criteris de lematització i processament del corpus de l'IULA. A més a més, s'ha complementat amb el diccionari de neologia recollida en el mateix corpus.

Consta de 106.000 entrades en el lemari i 843.000 en el formari. El seu volum actual és de 13,4 Mbytes per al lemari i de 30,6 Mbytes per al formari. El codi de caràcters del diccionari és el proposat en la norma ISO8859.

Per a cada entrada s'indica la categoria gramatical, el paradigma de flexió i les característiques morfosintàctiques de les formes completes emmagatzemades a una base de dades de tipus Dbase.

El diccionari és a j:\privat\dbf_es:

- dale.dbf: lemari bàsic
  - dale06.dbf:
  - dale08.dbf:
  - dale10.dbf:
  - dale12.dbf
  - dale14.dbf
  - dale30.dbf
- dnle.dbf: lemari de neologia
  - dnle_f.dbf: formari de neologia

Estat de la qüestió de l'actualització del diccionari de neologia:

Actualment, el DNLE té 16.417 entrades:
nominals: - 15.580
verbals: - 537

3.1.3 Manteniment dels diccionaris

El manteniment dels diccionaris electrònics de l'IULA s'ha fet de forma periòdica amb la introducció de la nova neologia detectada en el CT i de forma molt més esporàdica amb la correcció d'errors que es localitzen durant el procés de documents. Per a ambdós tipus de tasques cal tenir en compte, en primer lloc, una sèrie de criteris bàsics establerts prèviament, i, en segon lloc, un conjunt de procediments per a la introducció de la neologia detectada de forma definitiva en els diccionaris.

3.1.3.1 Manteniment del diccionari català

Els criteris bàsics per al manteniment del diccionari català per al processament del CT són de dos tipus:

a) manteniment del diccionari base,

a) introducció de la neologia que apareix en els textos que s'incorporen al CT en el diccionari de processament.

A l'annex 1 poden consultar-se els criteris bàsics utilitzats per al manteniment del diccionari del català.

El procediment per a la incorporació de la neologia catalana del CT als diccionaris de l'IULA pot consultar-se a l'annex 2.

3.1.3.2 Manteniment del diccionari castellà

Els criteris bàsics per al manteniment del diccionari castellà per al processament del corpus de l'IULA són de dos tipus:

a) manteniment del diccionari base,

b) introducció de la neologia que apareix en els textos que s'incorporen al CT en el diccionari de processament.

A l'annex 3 poden consultar-se els criteris bàsics utilitzats per al manteniment del diccionari del castellà.

El procediment per a la incorporació de la neologia castellana del CT als diccionaris de l'IULA pot consultar-se a l'annex 4.

3.2 Etiquetaris

Amb el propòsit de proveir el CT d'una eina de referència per marcar gramaticalment els textos introduïts en català, castellà i anglès, s'ha proposat un llistat d'etiquetes per a cada llengua.

Per consultar els etiquetaris per al català i el castellà pot consultar-se el *Working Paper* 18 de la col·lecció de l'IULA: Morel, J. *et al.* (1998).

Per confeccionar l'etiquetari de l'anglès, es va fer una adaptació de l'etiquetari proposat per la *Constraint Grammar* al format del CT. Per consultar l'etiquetari per a l'anglès vegeu l'annex 5.

4 Eines

En el processament del CT s'utilitzen diferents eines:

a) desenvolupades per l'IULA,

b) de lliure distribució adaptades per l'IULA als interessos del projecte,

c) eines adquirides.

a) Eines desenvolupades per l'IULA

- Conjunt d'eines per a la incorporació de marcatge SGML als textos del CT:
  - marques de paràgraf i frase,
  - marcatge automàtic de pàgines HTML.
- Eina de preprocés per al català i castellà.
- Programes que faciliten la gestió i automatització de la cadena de processament en castellà:
  - recuperació automàtica de les mostres i anàlisi d'un document en castellà,
  - obtenció de la neologia,
  - devolució de les mostres analitzades a morgana.
- Eines que agiliten la introducció de neologia del Corpus en català, implementades en forma de pàgina web.
- *CATMORF*: analitzador morfològic del català, vegeu annex 6.
- *PALIC*: analitzador morfològic del castellà, vegeu annex 3.

Catmorf i *Palic* analitzen els mots dels textos del corpus a partir dels diccionaris de referència per a cada llengua (vegeu apartat 3), els lematitzen a partir dels criteris de lematització establerts per al processament del CT (vegeu annex 7), i els donen una categoria gramatical que necessàriament correspon a alguna de les etiquetes proposades en els etiquetaris de l'IULA (vegeu el *Working Paper* 18: Morel, J. *et al*. (1998) per als etiquetaris català i castellà o l'annex 5 per a l'etiquetari anglès).

- *AMBILIC*: desambiguador lingüístic per al castellà. *Ambilic* és un programa de desambiguació que aplica les regles lingüístiques redactades dins del projecte corpus.

Ubicació: j:\public\mat_aux\regles\espanyol.rdl

Comentaris a: j:\public\mat_aux\regles\espanyol.rem (fitxer per a comentaris de les regles creat el 2004).

- *bwanaNet*: eina d'interrogació del CT.

- *Sextan*: conjunt d'eines d'extracció i gestió de neologia per al català i el castellà. Vegeu working paper sèrie monografies 8, .
- *ALINEA*, alineador de textos.
- *bwana*: extractor d'informació del corpus, en desús.
- Eina d'accés a *CATMORF* via Internet, en desús.
- *Mercedes*: eina que permet detectar en els textos processats del CT els termes inclosos en uns diccionaris de referència.
- *YATE*: Extractor de candidats a termes que funciona a partir dels textos processats en el CT.

b) Eines de lliure distribució adaptades:

En el marc del projecte Corpus s'ha adaptat a la cadena de processament un desambiguador estadístic de lliure distribució per al català i per al castellà.

c) Eines adquirides:

- *Constraint Grammar*:
  - entorn per al desenvolupament de l'analitzador sintàctic de baix nivell per al català,
  - analitzador sintàctic de baix nivell per a l'anglès,
- *Dynatext*: base de dades textual per l'emmagatzematge del corpus, en desús.
- Entorn de desenvolupament per al llenguatge Java.

5 Cadena de treball del corpus

5.1 Fase de selecció de textos

El CT conté tres tipus de textos:

a) Textos especialitzats: Textos que un especialista de la matèria considera pertinents i que pot classificar temàticament dins d'una estructuració del domini prèviament consensuada per especialistes de l'àmbit.

b) Textos de premsa: Textos que utilitza l'Observatori de Neologia per fer el buidatge de neologismes.

c) Textos proposats i processats per estudiants del doctorat i visitants, que són al calaix de general, cadascun dins la carpeta que identifica el domini al qual pertany (lingüística, minerals...).

5.2 Fase de marcatge estructural i confecció de capçalera

Vegeu annex 8

5.3 Fase de processament lingüístic

5.3.1 Processament lingüístic de documents en català

Vegeu annex 9.

5.3.2 Processament lingüístic de documents en castellà

Vegeu annex 10.

5.3.3 Processament lingüístic de documents en anglès

Vegeu annex 11.

5.4 Fase d'incorporació a *bwanaNet*

Vegeu annex 12.

6 Publicacions

La constitució del CT de l'IULA, així com el desenvolupament de les eines d'exploració han donat lloc a diverses publicacions (http://www.iula.upf.edu/corpus/corpubca.htm)

7 Recursos humans

En la creació i desenvolupament del corpus de l'IULA hi han participat moltes persones que, agrupades per tasques específiques, figuren a continuació.

Direcció: M. Teresa Cabré

Consell científic: Toni Badia, M. Paz Battaner, M. Teresa Cabré, Lluís de Yzaguirre, Mercè Lorente i M. Teresa Turell.

Especialistes en cada àrea temàtica:

Dret: Jordi Argenter, Xavier Bernardí, Carles Duarte, Elena Ferran, Rolf Gaser.
Economia: Màxim Borrell, Miquel Centelles, Richard Gross, Juanjo Hernández,Vicente Ortun.
Medi ambient: Richard Gross, Xavier Mas, Pau Serra.
Medicina: Xavier Mas, Toni Valero.
Genòmica**:** Fernando Nápoles, Maria Roura.
Informàtica: Horacio Rodríguez, Jordi Vivaldi.

Informàtics: Marta Busquets, Jesús Carrasco, Manel Pujol, David Solé, Toni Tuells i Jordi Vivaldi

Documentalistes: Lídia Fosalba, Walter Llorach, Gemma Martínez i Mireia Ribera

Coordinació de tasques:

Coordinació tècnica: Jordi Vivaldi
Coordinació de les diferents etapes de processament: Carme Bach, Elisenda Bernal, Rosa Estopà, Roser Saurí i Xavier Solé.

Col·laboradors tècnics: Marta Carulla, Josep M. Fontana, Montserrat Forcadell, Louise Mcnally i Enric Vallduví

Becaris i professors ajudants: Araceli Alonso, Gemma Andújar, Judit Aumatell, Claudia Baez, Laura Borràs, Núria Castillo, Cristina Corcoll, Montserrat Cunillera, Iria da Cunha, Laura de la Fuente, Judit Feliu, Noemí Fluixà, Rosanna Folguerà, Yannick Garcia, Àngel Gil, John Jairo Giraldo, Marisa González, Montserrat González, Anna Grau, Ricardo Guantiva, Sònia Jiménez, Marta Juncà, Irina Kostina, Judit Lafuente, Melva Márquez, Roser Martínez, Jordi Morel, Núria Oliva, Rogelio Nazar, Helena Pàmpols, Sthephane Patin, Mar Pongilupi, Marina Polymeridou, Gabriela Resnik, Àlex Ribera, Elisabeth Ricart, John Roberto, Carlos Rodríguez, Ester Rosàs, Judit Sadurní, Miquel Sánchez, Marisa Santiago, Maria Stephanova, Mercedes Suárez, Òscar Talamino, Sergi Torner, Teresa Vallès, Patricia Wilches, Alba Xandri i Helena Xirau.

Finalment volem esmentar algunes entitats que hi han col·laborat:

Bibliograf	Consorci per a la Normalització Lingüística
Edicions UPC	Edicions Zinco SA
Editorial Ariel	Editorial Barcanova
Editoral Marcombo	Empresa Fotocopiadora Grafilia
Generalitat de Catalunya	Caixa de Pensions,
OPS/OMS	Promociones y Publicaciones Universitarias S.A.
Tribunal Eclesiàstic	Universitat Autònoma
Universitat d'Alacant	Universitat Politècnica de Catalunya (FIB)
Universitat Pompeu Fabra	

8 Annexos

8.1 Annex 1: Criteris establerts per al manteniment del diccionari del català

8.2 Annex 2: Incorporació de la neologia del català als diccionaris de l'IULA

8.3 Annex 3: Criteris establerts per al manteniment del diccionari electrònic del castellà

8.4 Annex 4: Incorporació de la neologia del castellà als diccionaris de l'IULA

8.5 Annex 5: Etiquetari anglès per al tractament del textos del CT

8.6 Annex 6: *Catmorf*: analitzador morfològic de català

8.7 Annex 7: Criteris de lematització del CT

8.8 Annex 8: Procediment per a l'adquisició de textos amb l'escàner i posterior etiquetatge estructural

8.9 Annex 9: Processament lingüístic de documents en català

8.10 Annex 10: Processament lingüístic de documents en castellà

8.11 Annex 11: Processament lingüístic de documents en anglès

8.12 Annex 12: Incorporació dels documents del CT a *bwanaNet*

8.1 Annex 1: Criteris establerts per al manteniment del diccionari electrònic del català

Intentem tenir un diccionari modular fidel als tres diccionaris a partir dels quals s'ha confeccionat i tractem alhora que es reconeguin el màxim nombre d'entrades possibles, sense considerar les diferències semàntiques, seguint un criteri d'economia. Això implica una sèrie de decisions que hem aplicat fins aquest moment:

a) Si trobem una mateixa entrada que en un diccionari és només (masculí) i en un altre només (femení), l'entrada en el diccionari màquina queda com (masculí i femení)

exemple:

Entrada	Lema	DLC3	DIEC	Diccionari Màquina
balboa	balboa	m	f	mf
banasta	banasta	f	mf	mf

b) Si trobem una entrada que és una variant ortogràfica d'una altra entrada ja existent, l'entrem:

DLC3	DIEC	Diccionari Màquina
vice-president		vice-president
	vicepresident	vicepresident

c) Si en una mateixa entrada hi ha subentrades que comparteixen les mateixes característiques morfològiques, no es té en compte aquesta distinció sempre que no afecti un canvi de lema:

exemple:

balear (Dlc3 i Diec) (adjectiu) (nom masculí i femení) i (nom masculí)

diccionari màquina *balear* (nom-adj) (invariable (mf)) no flexió de gènere i sí de nombre

Entrada	Lema	Dlc3 i Diec	Diccionari màquina
balear	balear	adj.	(nom-adj) (m i f)
		m i f	
		m	

ln('balear',[],morf(lema('balear',cat(nmenys),class(nom-adj),gen(mf),nom(sing),flex(gen(no),nom(si))),3,0).[1]

[1]Aquesta és la formalització de les entrades nominals en el diccionari màquina:

ln(entrada,[ ],morf,n1,n2)

entrada
[ ]..... Espai reservat a les regles de flexió/derivació
morf..... Informació pròpiament morfològica de l'entrada que es divideix en:
morf((lema,cat,class,gen,nom,flex((gen)(nom)))
lema
cat (que en el nostre cas és sempre nmenys, per tal com només revisem la classe major dels noms)
class (nom) o bé (adj) o bé (nom-adj)
gen (informació sobre el gènere de l'entrada)
(m) o (f) o ((mf) per a noms i/o adjectius invariables)
nom (informació sobre el nombre de l'entrada)
(sing) (p) o ((s-p) per a noms i/o adjectius invariables pel que fa la nombre)
flex (informació sobre la flexió de l'entrada):
gen (s'hi indica si l'entrada flexiona o no pel que fa al gènere)
nom (s'hi indica si l'entrada flexiona o no pel que fa al nombre)
n1 (nombre de vocals de l'entrada)
n2 (nombre de vocals accentuades gràficament)

bacil·làcies (Dlc3 i Diec) (femení plural) i (femení singular)

diccionari màquina *bacil·làcia* (nom) (f) no flexió de gènere i sí de nombre

	Dlc3 i Diec	Diccionari màquina
Entrada	bacil·làcies	bacil·làcia
	f. pl.	f. sing. flexió de nombre
	f. sing	

```
ln('bacil·làcia',[],morf(lema('bacil·làcia',cat(nmenys),class(nom),gen(f)
,nom(sing),flex(gen(no),nom(si))),5,1).
```

Si la subentrada implica un canvi de lema (bàsicament en noms femenins) aleshores aquesta subentrada té una entrada pròpia en el diccionari màquina:

informàtic (adj.) (m i f) (f)

diccionari màquina dues entrades:

informàtic (nom-adj) lema informàtic (m) flexió de gènere i de nombre

informàtica (nom) lema informàtica (f) flexió de gènere no i de nombre sí

	Dlc3 i Diec	Diccionari màquina entrada 1	Diccionari màquina entrada 2
Entrada	informàtic	informàtic	informàtica
	adj.	nom-adj	nom
	nom m i f	m (sí flexió de gènere i nombre)	f (no flexió de gènere, sí de nombre)
	nom f		

```
ln('informàtic',[],morf(lema('informàtic',cat(nmenys),class(nom-
adj),gen(m),nom(sing),flex(gen(si),nom(si))),4,1).
ln('informàtica',[],morf(lema('informàtica',cat(nmenys),class(nom),ge
n(f),nom(sing),flex(gen(no),nom(si))),5,1).
```

d) Les entrades dels nombres les deixem en la revisió tal com estan en els diccionaris malgrat que el tractament que en fem en el CT sigui diferent.[2]

Detecció d'errors del diccionari suport paper:

En principi corregim els errors que detectem:

exemple:

cadella (Dlc3) té una entrada pròpia com a nom femení, femella del cadell, però no és coherent amb la filosofia d'aquest diccionari, aleshores no li donem una entrada com a femení amb lema *cadella* sinó que apareix en màquina sota l'entrada *cadell* que sí flexiona en gènere.

Sintagmes travats:

En el diccionari en suport paper apareixen en algunes ocasions sintagmes travats que tenen una entrada pròpia. El diccionari màquina no està previst que tracti unitats separades per un espai en blanc.

exemple:

delirium tremens (Dlc3), *delírium trèmens* (Diec)

No entrem aquestes unitats en el diccionari perquè no està previst que hi hagi entrades amb espais en blanc. De totes maneres fem el comentari pertinent en la llista de comentaris i introduïm aquestes entrades com a locucions nominals per tal que siguin reconegudes pel preprocés. Cal fer un llistat d'aquests sintagmes a j:\public\mat_aux\dbf_asc\probleca\ travats.doc.

Col·locacions adjectivals

Hi ha una sèrie d'adjectius que tenen una entrada pròpia com a adjectius invariables però dels quals sabem que només poden aparèixer en determinades combinacions clarament marcades pel que fa al gènere. El criteri establert és deixar l'entrada com a adjectiu invariable per tal de ser fidels al diccionari paper i si hi ha alguna altra informació també posar-la.

[2] Si bé en els diccionaris els nombres apareixen com a noms i adjectius pel tractament del corpus els considerem especificadors (adj i pronoms). Quan es trobi un nombre en la revisió del diccionari cal informar-ne a la persona responsable del manteniment de diccionaris per tal que inclogui la informació a un fitxer que s'activa abans que el diccionari, en el qual es recullen les epecificitats que apliquem en el processament del CT. Aquest fitxer s'anomena *gramesca*. La formalització que correspondria a l'entrada d'un nombre en el fitxer *gramesca* és la següent:

gram(dos,"#dos\\dos:EC--M6#").

exemples:

camosa (Dlc3) **adj. i f** Dit d'una varietat de poma de mida mitjana o grossa de forma arrodonida i una mica aplatada. (Diec) **adj.** ***poma camosa*** V. poma.

(màquina) 2 entrades per *camosa* una com a nom femení que no flexiona en gènere però sí en nombre i una altra com a adjectiu invariable (m i f) que no presenta flexió ni de gènere ni de nombre.

Entrada	Lema	DLC3	DIEC	Diccionari màquina 1	Diccionari Màquina 2
camosa	camosa	adj.	adj		adj. (mf)
		nom f.		nom f.	

accelerat (Dlc3) **adj. i m** Dit de l'efecte produït per una acceleració.

(màquina) 2 entrades, una com a nom masculí que no flexiona en gènere i una altra com a adjectiu invariable (mf) que tampoc presenta flexió.

Entrada	Lema	DLC3	Diccionari Màquina 1	Diccionari Màquina 2
accelerat	accelerat	adj.		adj. (mf)
		nom m	nom m	

acceleratriu (Dlc3) ja ens diu clarament que és un adjectiu femení **adj f**, però el (Diec) sembla que l'entrada consta com a invariable **adj** Acceleradora. *Força acceleratriu.*

(màquina) una única entrada, com a adjectiu invariable que no flexiona en gènere.

Entrada	Lema	DLC3	Diec	Diccionari Màquina
acceleratriu	acceleratriu	adj. f.	adj.	adj. (mf)

Àcids

En relació directa amb les col·locacions adjectivals tenim l'entrada d'alguns àcids.

1. Quan els diccionaris de què partim coincideixen, els tracten com a adjectius invariables. Els considerem com a adjectius d'una sola terminació (mf) que no flexionen en gènere i sí en nombre.

exemple: *nítric*

2. Hi ha casos però en què els dos diccionaris no coincideixen. Mentre que el Diec indica sempre la categoria gramatical de l'entrada com a adjectiu, el Dlc3 no dóna informació sobre la categoria de l'entrada. En el diccionari màquina indiquem que és un adjectiu invariable pel que fa al gènere (mf) seguint el criteri anterior.

exemple:

acetilsalicílic (Dlc3) **acetilsalicílic, àcid**, definició

(Diec) **adj, acetilsalicílic,** V. àcid

Col·locacions nominals

En el diccionari en suport paper apareixen una sèrie de col·locacions nominals sense indicació de gènere. Els entrem com a noms amb el gènere que els pertoqui.

exemple:

hidrastina, clorur d' (DLC) gènere femení

hidrazo, compost (DLC) gènere masculí

Entrada	Lema	DLC3	Diccionari Màquina
hidrastina	hidrastina	-	nom. (f), flex gen (no), nombre (no)
hidrazo	hidrazo	-	nom. (m), flex gen (no), nombre (no)

Entrades que tenen un caràcter alfabètic estrany:

S'ha detectat alguna entrada en els diccionaris de què partim en la qual apareixen grafies que no pertanyen al nostre alfabet. El diccionari màquina no accepta aquestes entrades.

exemple: *α-acrosa* (Dlc3), aquest mot apareix ordenat per acrosa.

2-aminoantraquinona (Dlc3), aquest mot apareix ordenat per aminoantraquinona

En aquests casos no entrem les entrades però en fem el comentari.

Una cosa semblant succeeix amb entrades que tenen un caràcter marcat en cursiva:

exemple: ***d***-*al·losa* (Dlc3).

Entrarem l'entrada sota la lletra ***d*** tot i que estigui en l'ordre del diccionari sota la lletra ***a***.

8.2 Annex 2: Incorporació de la neologia del català als diccionaris de l'IULA

L'organització "conceptual" interna del diccionari del CT preveu tenir-lo dividit en tres grups principals:

- diccionari bàsic
- neologia de corpus
- neologia de premsa

Per tal de facilitar el tractament de les dades, cadascun d'aquests diccionaris es divideix en dos grans grups, un per als noms i un altre per als verbs. En el cas del diccionari bàsic, a més a més, hi ha un tercer mòdul per a les paraules tancades (preposicions, pronoms, etc.).

En el "Manual d'usuari de CATMORF" es troba tota la informació necessària per incorporar noves entrades al diccionari del CT (vegeu annex 6). De totes maneres, per facilitar aquesta tasca s'ha creat un mecanisme que es divideix en tres etapes:

a) detecció de la neologia del corpus tècnic i introducció a través de la pàgina web dels neologismes i de la informació morfològica que duen associats.

b) incorporació de les paraules (i la informació associada) a un fitxer temporal i verificació de les dades,

c) incorporació definitiva de las dades del fitxer temporal al diccionari de neologia del CT.

8.2.1 Detecció de la neologia del corpus tècnic i introducció a través de la pàgina web dels neologismes i de la informació morfològica que duen associats

8.2.1.1 Selecció del document del CT del qual es vol introduir la neologia

Introduirem la neologia dels documents que hem marcat.

A la base de dades de documents seleccionarem els documents que estiguin entrats al corpus, (aquells que en la columna *Base de Dades* tinguin assignat un número).

Un cop seleccionat el document a revisar (un dels que s'hagin marcat) se li assigna una **r** a la columna *NC*. D'aquesta manera s'indica que el document està en procés de revisió. Un cop el document ja està completament revisat i hem entrat la neologia trobada cal canviar la **r** del document per una **R.**

Ex:

Suposem que volem extreure la neologia del document de medi ambient núm. 45 *Empresaris verds per a un planeta blau*. El nom d'entrada d'aquest document a la base de dades és a00038.sgm. A la casella *NC* posarem una *r* per tal d'indicar que el document ja està en procés de revisió.

8.2.1.2 Processament del document escollit

Es tracta ara de processar lingüísticament el document que hem escollit, de tal manera que s'assignin tots els lemes i les etiquetes gramaticals als mots inclosos en el diccionari. Aquest és un pas previ per detectar les paraules desconegudes.

Per tal de processar el document, s'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el pasword de l'usuari.

Des de *morgana*, escriurem la següent ordre a la línia de comandes:

pretag5.pl -i nom_del_document.sgm -bd -q

Si volem analitzar el document 45 de medi ambient escriurem:

pretag5.pl -i a00038.sgm -bd -q

8.2.1.3 Consulta dels noms de les mostres en què haurem de comprovar el context d'aparició dels neologismes

Un cop finalitzat el marcatge d'un document del corpus, aquest passa a incorporar-se a una base de dades definitiva en entorn Unix. Per aquest motiu les mostres canvien de nom.

Ex:

Les 11 mostres del document 45 de medi ambient, a45.1, a 45.2, a45.3, han canviat de nom en l'entorn UNIX.

Per tal de poder consultar el context d'aparició dels neologismes del CT cal saber necessàriament el nom de les mostres on les paraules no reconegudes apareixen.

Per tal de saber el nom de cadascuna de les mostres, des d'una finestra UNIX i des del servidor Morgana cal executar la següent comanda:

dirCT -(lletra inicial de l'àrea a què pertany el document que estem revisant: a, e, i, d, m)

Ex:

Per saber el nom actual de les mostres del document de medi ambient 45 haurem de donar l'ordre següent:

dirCT -a

Ho consultarem a partir del nom .sgm del document.

Ex:

El document de medi ambient 45 té a la base de dades definitiva el nom a00038.sgm.

En el fitxer de sortida buscarem el nom de les onze mostres del document a00038.sgm:

```
a00038.idx: Empresaris verds per a un planeta blau
                1  00177ams.cz0
                2  00178ams.cz0
                3  00179ams.cz0
                4  00180ams.cz0
```

```
5  00181ams.cz0
6  00182ams.cz0
7  00183ams.cz0
8  00184ams.cz0
9  00185ams.cz0
10  00186ams.cz0
11  00187ams.cz0
```

8.2.1.4 Transport d'aquestes mostres des de l'entorn UNIX a NOVELL: via FTP

El pas següent és el transport de les mostres a revisar de l'entorn Unix a l'entorn Novell. Això ens permetrà modificar els originals quan sigui necessari amb més facilitat i també consultar còmodament el context d'aparició dels neologismes.

Per això es recomana obrir una carpeta a k:\tmpcorr\nom de l'usuari que fa la revisió, en la qual es transportaran les mostres des de UNIX.

Ex:

Suposem que l'usuari *bach* ha de fer la revisió del document a00038.sgm. A l'entorn Novell, dins de k:\tmpcorr cal obrir una carpeta nova *carme* on s'abocaran les mostres originals des de UNIX.

Si s'està treballant sobre més d'un document es recomana també fer carpetes diferents per a cada document per tal que no es barregin les mostres.

Ex:

Suposem que l'usuari bach revisa la neologia dels documents a00038.sgm i m00047.sgm. A k:\tmpcorr\carme obrirem dues carpetes a00038 i m00047 on abocarem les mostres corresponents de manera que no quedin desordenades.

8.2.1.5 Preparació del fitxer on quedaran recollits els mots no reconeguts pel diccionari

El següent pas és la preparació del fitxer on quedaran recollides les paraules no reconegudes.

En l'entorn UNIX i des del servidor morgana cal donar l'ordre següent:

apilado1xl.pl -i nom_del_document.sgm -des -o noves -q >&nom del fitxer

Ex:

Per crear un fitxer amb la neologia del document a00038.sgm de medi ambient s'ha d'executar la comanda següent:

apilado1xl.pl -i a00038.sgm -des -o noves -q >&a00038

Amb l'edició del fitxer a00038, es comproven quines són les paraules desconegudes que hi ha en aquest document.

Cal tenir en compte que si aquest procés s'ha de repetir diverses vegades cal esborrar el fitxer creat cada cop.

Aquest fitxer està localitzat a cd/usr2/iula/nom de l'usuari. És convenient traslladar-lo via FTP a la carpeta que està a k:\tmpcorr\nom de l'usuari\nom del document, per tal que es pugui imprimir des de word. En el cas del document a00038.sgm, el fitxer que conté la neologia està a k:\tmpcorr\carme\a00038.

8.2.1.6 Consulta del context d'aparició dels neologismes

Per tal de comprovar en quin context apareixen les paraules que no estan reconegudes des de morgana pot consultar-se amb exactitud el context d'aquestes paraules si s'executen les ordres següents:

apilado1xl.pl -i nom del document.sgm -bd -q

cd /usr3/iula/corpus/àrea/mostres4

vi nom de les mostres una a una

/? (aquesta ordre ens ensenyarà el context de totes les paraules desconegudes)

Per sortir d'aquest fitxer: :q!

Exemple:

apilado1xl.pl -i a00038.sgm -bd -q

cd /usr3/iula/corpus/mediamb/mostres4

vi 00177ams.cz0

Un cop tenim obert el document, demanem que ens trobi les seqüències desconegudes:

/?

Un cop ja tenim localitzats tots els contextos sortim del document amb la instrucció:

:q!

Aquesta feina de comprovació dels contextos es pot fer també des d'editors de textos que permeten l'accés via ftp com ara el Crimson Editor o EditPlus, fet que facilita molt la feina.

8.2.1.7 Comprovació dels mots no reconeguts en les mostres corresponents

Els mots no reconeguts no són sempre neologia sinó que pot tractar-se també d'algun tipus d'error.

8.2.1.7.1 Error de l'original

Succeeix en moltes ocasions que les paraules no reconegudes són un error d'escàner o un error ortogràfic ja present a l'original. En aquests casos es tracta de corregir la mostra des de l'entorn Novell.

8.2.1.7.2 Error de marcatge

En d'altres ocasions les paraules no són reconegudes perquè hi ha algun error de marcatge en les mostres que n'impedeix el processament correcte. En aquests casos, com en l'anterior, cal corregir l'original en entorn Novell.

Un cop hem corregit els errors de totes les mostres d'un mateix document transportarem de nou via FTP els documents de Novell a Unix.

D'aquesta manera agafarem els documents que teníem a k: /tmpcorr/nom de l'usuari i els traslladarem de nou a UNIX: /usr3/iula/corpus/àrea/mostres

Cal comprovar que en modificar els originals no s'hagi afegit algun nou error de marcatge. Per aquest motiu, a l'entorn Unix, al servidor morgana passarem un nou *parser* al document sencer.

Des de UNIX, en la línia de comandes cal escriure l'ordre següent:

nsgmls -s nom del document.sgm

Si no hi ha cap nou error de marcatge ens sortirà de nou la línia de comandes. Si hi ha algun problema, ens indicarà la mostra i la línia en què s'ha trobat el problema, que evidentment haurem de corregir, (tornar a passar de nou la mostra de Unix a Novell, corregir els errors, tornar a traspassar el document i tornar a passar el parser).

8.2.1.7.3 Error de diccionari

El diccionari utilitzat per processar els documents del CT és un diccionari modular format per tres diccionaris que es complementen: DLC83, DLC93 i DIEC.

Pot succeir que es presenti com a desconeguda una paraula que en realitat és en un dels tres diccionaris. Aleshores ho apuntarem en un fitxer de word on es recullen tots els errors de diccionari per tal que es puguin corregir de forma sistemàtica.

El fitxer es localitza a j:\public\mat_aux\dbf_asc\probleca i es diu analica.

Ex:

En un document sortia com a desconegut el plural *rebuigs*. Com que la paraula *rebuig* és al diccionari en el fitxer `analica` cal escriure:

Nom de la persona que entra la informació

Nom de la mostra on ha trobat el problema

Data en què n'informa

- no reconeix el plural *rebuigs* malgrat tenir en el diccionari *rebuig*

8.2.1.7.4 Error de preprocés

En algunes ocasions surten com a desconegudes paraules que o bé haurien d'haver estar reconegudes automàticament pel preprocés o bé que són al diccionari i que en algunes ocasions es reconeixen i en d'altres no.

Cal analitzar aleshores si es tracta d'alguna limitació del preprocés o bé d'algun defecte. S'apunta l'error en un fitxer de word preproca que es localitza a: j:\public\mat_aux\dbf_asc\probleca.

8.2.1.7.5 Neologisme

Finalment, un cop s'ha descartat que la paraula desconeguda no sigui cap dels quatre tipus d'error anteriors, s'està realment davant d'un neologisme a incorporar a la base de dades del CT.

Per entrar la neologia en català del CT de l'IULA es disposa d'una pàgina WWW que facilita la tasca.

La següent pàgina es troba a:

http://www.iula.upf.es/corpus/afegirlex/afegirlex.htm

En aquesta pàgina WWW hi ha una pantalla d'ajuda que explica com cal utilitzar-la.

A continuació destaquem una sèrie de punts molt importants que fan referència a certes restriccions que s'han de tenir en compte a l'hora d'entrar la neologia del CT a la pàgina WWW:

a) No s'ha d'incloure com a neologia cap verb que no sigui de la primera conjugació. Si en trobem algun de neològic, n'haurem d'informar al fitxer analica localitzat a j:\public\mat_aux\dbf_asc\probleca.

b) Tampoc no s'ha d'incloure cap paraula formada pels prefixos que tot seguit especifiquem:

agro-
andro-
anti-
auto-
avant-
bi-
bio-
co-
con-
contra-
cripto-
crono-
deci-
des-
dis-
eco-
electro-
euro-
ex-
extra-
filo-
fisio-
fito-
germano-
hidro-
hiper-
hispano-
im-
in-
inter-
macro-
micro-
mini-
mono-
multi-
neo-
para-
pluri-
poli-
politico-
post-
pre-
pro-
pseudo-
psico-
quadri-
radio-
re-
semi-
sero-
sobre-
sub-
super-
supra-
tele-
trans-
tri-
turbo-
ultra-
vice-
video

Si alguna de les paraules que apareixen com a desconegudes conté aquest prefix, cal informar-ne al fitxer analica, que es troba a j:\public\mat_aux\dbf_asc\probleca.

Un cop fets tots aquests passos, la neologia introduïda s'incorporarà als diccionaris del CT cada divendres.

A la base de dades en què es controla el processament dels documents, haurem de canviar la *r* per una *R*.

Incorporació dels neologismes i la informació que duen associada a un fitxer temporal i verificació de les dades

Quan la incorporació de les paraules s'ha realitzat utilitzant la pàgina Web específica, les dades corresponents queden registrades en un fitxer situat al directori cd ~httpd/logs/afegirlex amb un nom que es correspon a l'adreça Internet de l'ordinador des d'on s'han introduït les paraules (p.ex. 193.145.43.23).

En aquest fitxer es troben les dades en el format definitiu. Abans de la incorporació a la base de dades de neologia s'han de realitzar una sèrie de passos destinats a:

- separar els noms del verbs (vegeu secció 2.1),
- verificar que no es dupliquin entrades (vegeu secció 2.2),

Totes les operacions que es realitzin per a l'ampliació de les bases de dades de neologia s'han de realitzar des de el grup catmorf. Per entrar en aquest grup s'ha d'emetre la següent comanda:

morgana%: newgrp catmorf

Al finalitzar la tasca d'actualització s'ha de retornar al grup primari corresponent amb la comanda

morgana%: exit

8.2.1.8 Separació de les entrades nominals de les entrades verbals

Com ja s'ha mencionat, els diccionaris del CT es divideixen en dos grans grups de entrades: nominals i verbals. Aquesta divisió s'ha de fer des de:

/usr3/iula/soft/CATMORF/Afegir_Lexic/Dades

Abocarem totes les entrades nominals en un fitxer (tmpnoms.pl) i en un altre totes les entrades verbals (tmpverbs.pl). Abans de fer aquesta operació, cal obrir els fitxers existents tmpnoms.pl i tmpverbs.pl per veure si contenen dades. Si és així, esborrarem les dades.

Per abocar el nous noms en tmpnoms.pl executarem la comanda:[1]

morgana%: fgrep "ln(" ~httpd/logs/afegirlex/nom_*fitxer_dades* >! *nom_fitxer_temporal*

D'aquesta manera, si estem processant les dades que s'han incorporat des de 193.145.43.23 la comanda a executar és la següent:

[1] fgrep és una utilitat estàndard de UNIX per seleccionar les línies d'un text que compleixen algun criteri. En aquest cas són les línies que contenen la seqüència ln(

morgana%: fgrep "ln(" ~httpd/logs/afegirlex/193.145.43.23 >! tmpnoms.pl

De manera similar s'ha de procedir amb els verbs:

morgana%: fgrep "lv(" ~httpd/logs/afegirlex/193.145.43.23 >! tmpverbs.pl

Aquest procés s'ha de repetir per a tots els fitxers que tinguin dades amb la diferència que per al segon fitxer i els successius la comanda ha de ser una altra per tal que les dades es vagin afegint al final del fitxer.

morgana%: fgrep "ln(" ~ httpd/logs/afegirlex/*nom_fitxer_dades* >> *nom_fitxer_temporal*

A mode d'exemple, si tenim tres fitxers amb dades (193.145.43.23, 193.145.43.39 i 193.145.43.24) s'han d'executar les comandes següents:

fgrep "ln(" 193.145.43.23 >! tmpnoms.pl
fgrep "ln(" 193.145.43.39 >> tmpnoms.pl
fgrep "ln(" 193.145.43.24 >> tmpnoms.pl
fgrep "lv(" 193.145.43.23 >! tmpverbs.pl
fgrep "lv(" 193.145.43.39 >> tmpverbs.pl
fgrep "lv(" 193.145.43.24 >> tmpverbs.pl

S'ha de tenir present que els fitxers de noms i verbs es creen amb permisos d'escriptura només per a la persona que els ha creat. Per tant si la persona que ha de revisar les entrades és una altra cal canviar els permisos amb la comanda.

morgana%: chmod 664 tmp*.pl

És important que immediatament després d'executar aquestes comandes s'esborrin els fitxers amb les dades originals per permetre la introducció de més dades sense duplicar innecessàriament el procés.

8.2.1.9 Verificació de les dades

Les dades entrades (noms i verbs) s'han de verificar per tal d'evitar duplicacions i/o errors. Per aquest motiu s'ha d'executar un procés específic de comprovació d'errors. El procediment a seguir és el següent:

a) canvi al directori on resideix el programa de comprovació:

morgana%: cd /usr3/iula/soft/CATMORF/Afegir_Lexic/Comprovacions

b) crida de l'intèrpret i càrrega del programa de comprovació:

morgana%: /usr/local/sicstus3.2/bin/sicstus
| ?- [comprova].
{consulting /usr3/iula/soft/CATMORF/Afegir_Lexic/.....
.....
missatges de l'intèrpret Prolog
.....

```
yes
```

c) càrrega de les dades a comprovar:

```
| ?-['/usr3/iula/soft/CATMORF/Afegir_Lexic/Dades/
tmpnoms.pl'].
{consulting /usr3/iula/soft/CATMORF/Afegir_Lexic/.....
            missatges de l'intèrpret Prolog
yes
```

d) comprovació del noms:

```
| ?- comprova_noms.

yes
| ?- halt.
morgana%:
```

En aquest moment s'ha creat un fitxer de text amb el nom "after_comprovacions.txt-pl" que conté els missatges producte de la comprovació de cada paraula.

El significat dels missatges que apareixen en el fitxer "after_comprovacions.txt-pl" és el següent:

Lema xxxxxx guardat

El lema xxxxxx és correcte. Deixar l'entrada sense modificar.

Entrada xxxxxx apareix vàries vegades en aquest fitxer

És una entrada repetida és a dir s'ha trobat més d'una entrada per aquest mateix lema en el nou fitxer de neologia. Esborrar totes les entrades addicionals deixant-ne només una.

Cal comprovar anteriorment que la informació de les entrades que esborrem sigui la mateixa. Hi ha casos en què la informació morfològica de las entrades no és del tot compartida, com és el cas de l'entrada per *poliesportiu*, que per tant ocupa dues línies:

```
ln('poliesportiu',[],morf(lema('poliesportiu'),cat(nmenys),
   class(nom),gen(m),nom(sing),flex(gen(no),nom(si))),5,0).
ln('poliesportiu',[],morf(lema('poliesportiu'),cat(nmenys),
   class(adj),gen(m),nom(sing),flex(gen(si),nom(si))),5,0).
```

Si aquesta entrada estigués repetida ocuparia 4 línies i només se n'han d'esborrar dues.

Entrada xxxxxx ja la teníem registrada al diccionari original o als fitxers de neologia

Aquesta entrada ja era en el diccionari general (nmenys.pl) o als diccionaris de neologia (neologia_ct_noms.pl o neologia_pr_noms.pl) amb la mateixa informació morfològica. Aquests fitxers es troben a /usr3/iula/soft/CATMORF/LEXIC/Dlc+Lx93+Lx95/Pl+Ql/

Cal esborrar totes les ocurrències d'aquestes entrades repetides als fitxers temporals.

ATENCIÓ! lema xxxxxx ja és present al diccionari original o als fitxers de neologia

Aquesta entrada ja era en el diccionari general (nmenys.pl) o als diccionaris de neologia (neologia_ct_noms.pl o neologia_pr_noms.pl) encara que existeix una discrepància en les dades específiques. Verificar el tipus de discrepància i eliminar o no en funció del resultat de l'anàlisi.

8.2.2 Incorporació definitiva de les dades

Una vegada s'ha completat la verificació d'aquest fitxer, s'ha d'incloure[2] al final del diccionari de neologia corresponent.[3]
Un cop finalitzada la introducció de neologia s'han de compilar els fitxers de neologia.
Per als noms cal seguir el procediment següent:

```
morgana%: /usr/local/sicstus3.2/bin/sicstus
|?-fcompile('neologia_ct_noms_iula').
```

i pels verbs:

```
morgana%: /usr/local/sicstus3.2/bin/sicstus
|?-fcompile('neologia_ct_verbs_iula').
```

Un cop ajuntats el nous mots al fitxer de neologia cal regenerar el diccionari seguint el procediment indicat en el "Manual d'usuari de CATMORF", consulteu annex 6.

El mateix procediment s'ha de repetir per als verbs amb els canvis en les comandes (comprova_verbs en lloc de comprova_noms) i fitxers involucrats (els fitxers del diccionari de neologia ara són, en aquest cas, neologia_ct_verbs.pl o neologia_pr_verbs.pl).

A continuació, ja tant sols ens queda fer l'índex pels verbs i pels noms.

[2]En cas que es faci servir l'editor vi, situar-se al directori:

cd /usr3/iula/soft/CATMORF/LEXIC/Dlc+Lx93+Lx95/Pl+Ql. Les comandes a executar són les següents:

```
vi neologia_ct_noms_iula.pl
:$
:r /usr3/iula/soft/CATMORF/Afegir_Lexic/Dades/tmpnoms.pl
:wq
```

[3]Pel que fa als noms la neologia de corpus s'inclou al fitxer neologia_ct_noms.pl mentre que la neologia de premsa s'afegeix a neologia_pr_noms.pl

8.3 Annex 3 : Criteris establerts per al manteniment del diccionari electrònic del castellà

8.3.1 Introducció

En una gran part del procés, l'assignació automàtica d'etiquetes morfosintàctiques dels textos en castellà del CT s'efectua a partir d'un conjunt de diccionaris en format electrònic, que s'estructura en forma modular. A mode d'introducció i de manera simplificada, podem dir que el diccionari està constituït per tres grans mòduls:

- Una versió electrònica del *Diccionario Actual de la Lengua Española* (DALE) de l'editorial Vox-Biblograf.
- Un fitxer en què s'introdueix la neologia no documentada al DALE.
- Un fitxer en què es recullen les paraules gramaticals.

Els dos primers són al directori J:\PRIVAT\DBF_ES. Ambdós diccionaris estan formats per un fitxer que en serveix de font, on només hi ha els lemes (no les formes flexionades), i per diversos fitxers relacionats entre ells, on hi ha les formes flexionades. Aquests darrers, que es generen de forma automàtica a partir del fitxer de lemes, són els que es fan servir en el procés d'anàlisi dels documents del corpus.

Els fitxers que componen aquests diccionaris són els següents:

a) DALE:
- Fitxer de lemes: DALE.DBF
- Fitxers de formes flexionades (les formes s'agrupen segons la seva longitud):
  DALE06.DBF
  DALE08.DBF
  DALE10.DBF
  DALE12.DBF
  DALE14.DBF
  DALE30.DBF

b) Diccionari de neologia:
- Fitxer de lemes: DNLE.DBF
- Fitxer de formes flexionades: DNLE_F.DBF

Per tal que les tasques de manteniment d'aquests diccionaris no facin aturar el processament de documents del Corpus Tècnic, hi ha sengles còpies dels fitxers de lemes, que són les que es manipularan i actualitzaran en les tasques de manteniment. Aquestes còpies, situades al mateix directori, tenen el nom següent:
- DALE_BO.DBF
- DNLE_BO.DBF

D'aquesta manera, quan es detectin errors en qualsevol d'aquestes dues fonts o es vulgui introduir neologia nova, els canvis s'introduiran en les còpies corresponents. A partir de les còpies, s'obtindran els nous fitxers de formes

flexionades, i, un cop s'hagi comprovat que no hi ha errors, se substituiran els fitxer vells pels nous.

Al seu torn, el diccionari de paraules gramaticals, anomenat GRAMEMAS.DBF, està al directori J:\PUBLIC\MAT_AUX\DBF_ASC. En principi, és un diccionari tancat, i per tant, no s'espera que calgui fer-hi modificacions.

En aquest document s'explica, en un primer moment, l'estructura dels dos diccionaris castellans de formes no flexionades (el *DALE* i el *DNLE*), i, en un segon apartat, com es duu a terme l'expansió del diccionari de formes flexionades.

8.3.2 Estructura del DALE

La versió electrònica del DALE que es fa servir en el procés d'assignació d'etiquetes morfosintàctiques dels documents del CT està constituït per quatre components diferents:

a) Un diccionari de lemes (formes no flexionades): *DALE.DBF*

b) Un fitxer de desinències: *DESI_MAN.DBF*

c) Un fitxer d'etiquetes: *CODIS_D.DBF*

d) Un conjunt de fitxers amb les formes flexionades

DALE.DBF conté un llistat de paraules sense flexionar (lemes), per a cadascuna de les quals es proporciona la categoria gramatical i el paradigma de flexió que seguirà la paraula quan faci la flexió. Aquest primer component es relaciona amb el segon, el fitxer de desinències (DESIMAN.DBF), en el qual s'especifica quines desinències flexives corresponen a cadascun dels models de flexió. La combinació de la informació d'aquests dos components genera totes les formes flexives que conformaran el diccionari final, però no aporta cap informació sobre el valor que té una determinada forma. Per conèixer aquest valor, el programa d'expansió ha de consultar el tercer dels components del diccionari, el fitxer d'etiquetes (CODIS_D.DBF), que s'encarregarà d'assignar una etiqueta TEI a cadascuna de les formes resultants del procés d'expansió.

El resultat d'aquest procés d'expansió és un diccionari de formes flexionades que assigna a cada forma una etiqueta morfosintàctica i un lema. Perquè aquest procés es realitzi d'una manera efectiva cal relacionar la informació que contenen els tres primers fitxers que constitueixen el diccionari. Aquesta relació entre la informació dels diversos fitxers s'aconsegueix gràcies al fet que en l'estructura de cadascun d'ells s'han previst uns camps que vinculen cada registre d'un fitxer amb un o diversos registres dels altres components del diccionari.

A continuació explicarem de forma més detallada com funciona aquest procés. La figura 1 (pàgina següent) resumeix de manera gràfica quina estructura té cadascun dels tres primers components del diccionari, i quins són els camps de la base de dades que contenen informació pertinent per al procés d'expansió. Així mateix, s'especifica quins camps són els que permeten establir els vincles entre els diferents fitxers.

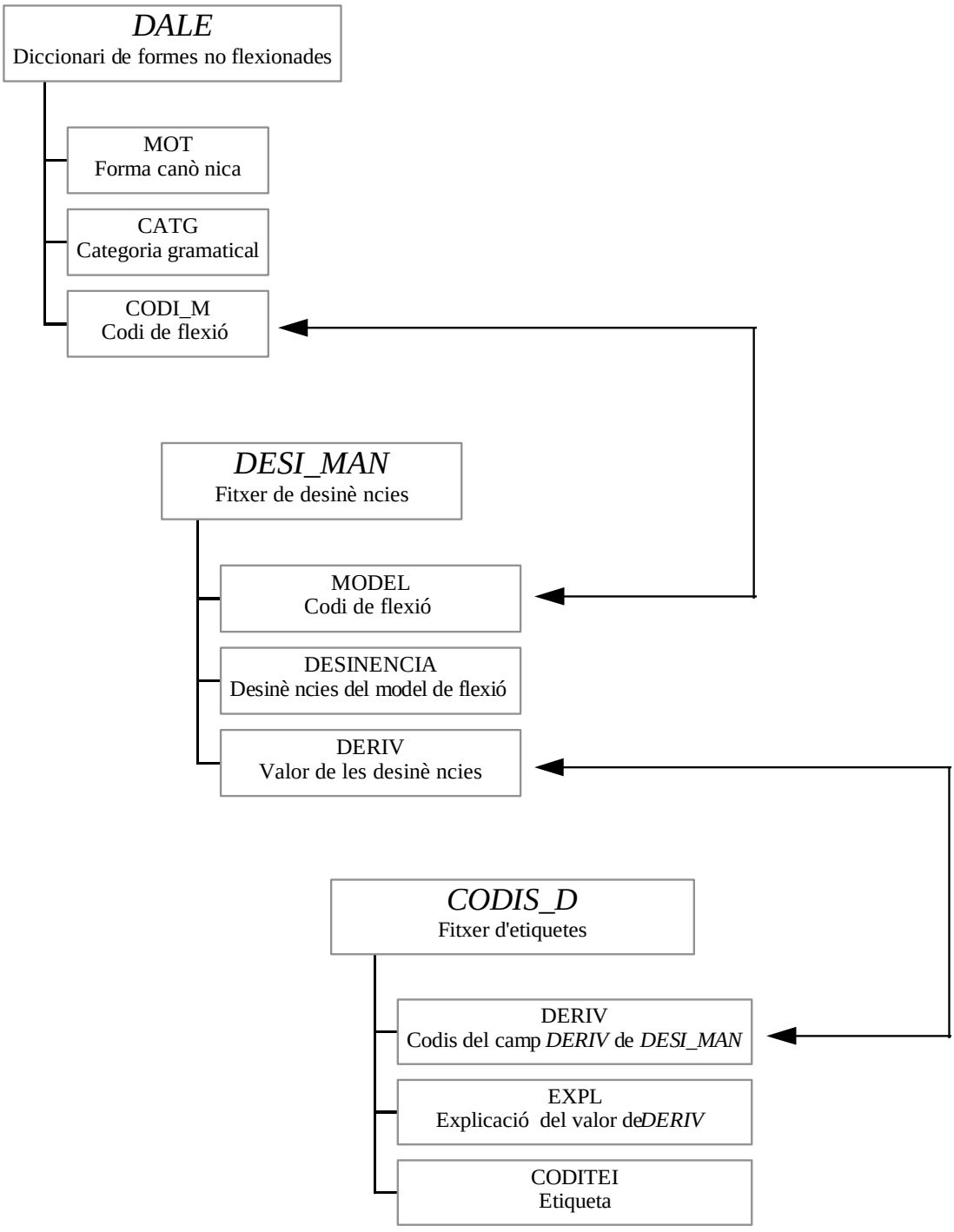


Figura 1

8.3.2.1 Diccionari de formes no flexionades

El primer component del diccionari castellà el constitueix una versió en format electrònic del *Diccionario Actual de la Lengua Española* (Barcelona: Biblograf, 1995), que es guarda al directori J:\PRIVAT\DBF_ES i que s'anomena DALE.DBF. Aquesta versió del diccionari en format electrònic no conté tota la informació que hi ha en el diccionari en format paper sinó que només inclou la macroestructura (i. e. el lemari) i la informació categorial i flexiva que s'associa a cada lema. No hi ha, per tant, cap altra informació pertanyent a la micorestructura (definició, ús, etc.).

Aquesta base de dades té la informació repartida en diversos camps. Per dur a terme l'expansió del diccionari és pertinent la que s'inclou en els camps següents:

a) MOT: forma canònica de la paraula (lema).

S'assumeix que la forma canònica és el masculí singular en els adjectius i en els noms de gènere comú, el singular en els altres noms (masculí o femení segons correspongui), i l'infinitiu en els verbs. Per tant, serà aquesta la forma que es recollirà en aquest camp.

Tanmateix, en algun cas, el *DALE* recull sota una sola entrada variants formals que en la versió electrònica han generat entrades diferents. Per exemple,

asunceno, -na, asunceño, -ña

s'ha desdoblat en el diccionari electrònic en dues entrades diferents, la informació de les quals en el camp *MOT* és la següent:

MOT asunceno

MOT asunceño

És important tenir en compte que les entrades s'introdueixen sempre amb tots els caràcters en minúscules.

b) CATG: categoria gramatical.

En aquest camp es conserva, sense modificar-la, la informació categorial del *DALE*. En el diccionari original, la informació categorial es dóna per mitjà d'abreviatures, i a una mateixa entrada li poden correspondre diferents categories gramaticals. En la nostra versió electrònica del diccionari hem canviat les abreviatures per uns codis que fan equivaler a cada categoria una única lletra en majúscules. Les categories gramaticals que s'han contemplat són les següents:

M nom masculí
F nom femení
C nom comú
A adjectiu
V verb[1]

[1]En la seva versió en paper, el *DALE* classifica els verbs atenent a una tipologia de base morfosintàctica. Tot i que s'ignora en el procés d'expansió, a la versió electrònica hem conservat aquesta informació afegint darrera del codi categorial *V* un seguit de codis que es corresponen amb

Quan una entrada aglutina diferents categories gramaticals, es reuneixen totes en el camp *CATG*. Així, per exemple, l'entrada *general* del diccionari en format paper queda en el diccionari electrònic com segueix:

MOT general
CATG AM

On A indica l'ús com a adjectiu i *M* l'ús com a nom masculí.

Pel que fa referència a les etiquetes categorials dels noms cal fer alguna puntualització. El *DALE* estableix tres tipus de noms segons el gènere: masculí, femení i comú. Un nom de gènere comú és, segons aquest diccionari, un nom que té la mateixa forma per al masculí i el femení (per exemple, *acróbata*). En aquest sentit, els noms de gènere comú són diferents dels noms que presenten flexió de gènere (per exemple, *autor, -ra*), que en la versió impresa del *DALE* porten les marques *m. f.* (i. e. masculí i femení). La diferència entre uns i altres noms queda exemplificada a la taula següent:

ENTRADA AL DALE	*GÈNERE*	*FORMA FLEXIONADA*
acróbata	comú	acróbata acróbatas
autor, -ra	masculí i femení	autor autores autora autoras

En la versió electrònica del diccionari, tanmateix, s'han canviat els codis categorials dels noms comuns i dels noms que poden ser masculins i/o femenins. Efectivament, en la versió electrònica del *DALE* es consideren els dos tipus de noms següents:

a) Noms que existeixen tant en femení com en masculí, presentin o no diferències degudes al gènere: nom comú (codi categorial C).

b) Noms que només poden ser masculins o femenins: seran, segons correspongui, noms masculins (codi categorial M) o *femenins* (codi categorial F).

Aquest canvi s'ha fet amb la finalitat que l'assignació de lemes a les diferents formes flexives del diccionari expandit respecti els criteris de lematització que s'han establert per al CT. Per consultar els criteris de lematització vegeu annex 7.

De forma coherent amb els criteris de lematització establerts, en el diccionari electrònic es diferencien, d'una banda els noms que poden ser masculins o femenins (codi categorial M o F, respectivament), i, d'altra banda, els noms que poden ser tant masculins com femenins, tinguin o no variacions degudes al gènere (codi categorial C). A la taula següent proporcionem exemples de codificació dels diversos tipus de noms:

---

els diversos tipus verbals. Els tipus que hem tingut en compte —que són els que contempla el *DALE*— són *transitiu* (T), *intransitiu* (I) i *pronominal* (P).

DICCIONARI DE LEMES		*DICCIONARI EXPANDIT*	
MOT	*CATG*	*FORMES FLEXIVES*	*LEMA*
acróbata	C	acróbata acróbatas	acróbata
autor, ra	C	autor autores autora autoras	autor
deseo	M	deseo deseos	deseo
abertura	F	abertura aberturas	abertura

Cal tenir en compte, però, que és possible que una forma correspongui a lemes diferents, un de masculí (d'una paraula que té flexió de gènere) i un de femení (de una paraula que no té flexió de gènere). Aquest és, per exemple, el cas de *física*, que pot ser una aparició del lema *física*, femení (la disciplina científica), o de la forma femenina del lema *físico*, amb variació de gènere (la persona que es dedica a la física). En aquest cas, s'hauran de fer dues entrades diferents al diccionari electrònic:

- MOT: *física*
  CATG: F
  Formes flexives: *física, físicas*.
- MOT: *físico*
  CATG: C
  Formes flexives: *físico, física, físicos, físicas*.

Semblantment, és possible que un mateix mot existeixi com a masculí i com a femení (això és, que en la lematització s'hagin d'assignar lemes diferents per al masculí i per al femení). Per exemple, *aroma* significa 'olor' en masculí, i en femení es el nom d'una flor. Quan això sigui així, en el camp CATG s'hauran de consignar les dues possibilitats, per la qual cosa s'hi inclouran les marques F i M (i no la marca C, que indicaria que tant al femení com al masculí li corresponen el mateix lema). És a dir, l'anàlisi d'un mot marcat com a MF donarà dos lemes diferents, un masculí i l'altre femení; en canvi, l'anàlisi d'un mot marcat com a C donarà un únic lema amb possibilitat de tenir una forma masculina i una altra femenina (o subespecificat de gènere).

Finalment, pel que fa referència a la informació que es codifica en el camp CATG, s'ha de fer esment del fet que en els mots gramaticals[2] aquest camp resta en blanc, ja que en el procés d'etiquetatge dels textos la informació relativa a aquests mots no s'extreu del *DALE* sinó del fitxer de GRAMEMAS. Així mateix, tampoc no es

[2]Es considera que són mots gramaticals els adverbis, les conjuncions, els determinants, les preposicions, els pronoms i els adjectius numerals, indefinits i possessius. Totes les paraules d'aquestes categories s'han recollit al fitxer de *GRAMEMAS*, que inclou també algunes paraules que, tot i no poder-se considerar pròpiament mots gramaticals, tenen una freqüència d'ús molt alta (com, per exemple, els verbs *ser* i *estar*).

fa constar la informació sobre les locucions en què intervé un mot, perquè les locucions es tracten en el preprocés.[3]

c) CODI_M: model de flexió.

Com que a DALE.DBF només es recull la forma canònica de cada paraula, les diferents formes flexives de cada mot s'han de generar automàticament. A cada paraula se li assigna un codi que especifica quin és el model flexiu que li correspon.

8.3.2.2 Fitxer de desinències

En un fitxer independent es desenvolupen els diferents models de flexió que s'han utilitzat al camp CODI_M de la versió electrònica del *DALE*. Aquest fitxer s'anomena DESI_MAN.DBF i es troba al mateix directori que el DALE, és a dir, a J:\PRIVAT\DBF_ES. Entre la diversa informació que conté aquest fitxer, és necessària per al procés d'expansió del diccionari la que es recull als següents camps:

a) MODEL: codi de flexió.

Aquests codis coincideixen amb els del camp CODI_M del fitxer DALE.DBF. Cadascuna de les desinències necessàries per a generar les diferents formes flexives d'un determinat model constitueix un registre diferent en aquesta base de dades, de manera que hi haurà tants registres com formes flexives es generin amb aquest model amb la mateixa informació al camp MODEL. Per exemple, el model de flexió 10 fa el plural dels noms acabats en *-ón*, i té, per tant, dues desinències diferents. Segons hem dit, a DESI_MAN hi hauria dos registres per a aquest model de flexió:

MODEL	DESINENCIA
10	ón
10	ones

b) DESINENCIA: desinències de cada model de flexió.

Entenem *desinència* en un sentit lax, no en el sentit tècnic de la morfologia: en aquest camp, es recull aquell segment del mot que varia en flexionar (i no només els morfemes de flexió *strictu sensu*). Així, per exemple, una paraula que és regular en la seva flexió pot pertànyer a un model de flexió diferent del regular pel fet que en alguna de les variants flexives té canvis en l'accentuació[4]:

- *alemán, alemanes*
- *examen, exámenes*
- *retén, retenes*

c) DERIV: valor de les desinències.

[3] Hi ha un fitxer de locucions anomenat *LOCUCIES.DBF* al directori *J:\PUBLIC\MAT_AUX/DBF_ASC*.

[4] Als exemples, subratllem els segments que s'haurien d'incloure al camp *DESINENCIA*.

Cadascuna de les desinències d'un determinat model genera una forma flexiva amb un valor gramatical concret. Així, el model de flexió *1*, que s'aplica a alguns noms i adjectius, té dues desinències (*0* i *-es*); la desinència *0* té el valor de 'singular', i la desinència *–es*, el de 'plural'. Per a conèixer aquesta informació, el programa consulta el camp DERIV; tanmateix, la informació relativa a aquest valor de les desinències no es dóna de forma directa sinó que en aquest camp hi ha un únic caràcter que constitueix un codi alfanumèric que remet a un tercer fitxer on s'especifica tant el valor de la desinència com l'etiqueta que, segons l' etiquetari castellà de l'*IULA*, correspon a aquest valor.

8.3.2.3 Fitxer d'etiquetes

En una darrera base de dades s'explica el valor dels codis alfanumèrics del camp DERIV de DESI_MAN.DBF. Aquest fitxer, anomenat CODIS_D.DBF, és comú pel català i el castellà, i està al directori J:\PUBLIC\MAT_AUX\DBF_ASC. D'aquest fitxer són pertinents per al procés d'expansió els tres camps següents:

a) DERIV: codis utilitzats al camp *DERIV* de *DESI_MAN.DBF*.

Hi ha una entrada per a cada codi o combinació possible de codis utilitzats a *DESI_MAN*. Per exemple, la forma *-ón* del model de flexió 10, que abans hem fet servir com a il·lustradora del funcionament del camp MODEL, és una forma de singular. A DESI_MAN, el registre corresponent a aquesta forma tenia un codi alfanumèric al camp DERIV; en el fitxer CODIS_D s'especifica que el valor d'aquest codi és, efectivament, el de singular.

b) EXPL: explicació del valor del codi de DERIV.

Aquesta explicació és una paràfrasi en llengua natural del valor del codi, i només serà utilitzada per l'usuari humà que codifiqui el diccionari.

c) CODITEI: etiqueta que en el procés d'expansió s'assignarà a les formes flexives dels mots generades amb la desinència que remet a aquest codi.[5]

8.3.3 Les tasques de manteniment del diccionari

8.3.3.1 El DALE

La versió electrònica del *DALE* ha de respectar la informació que hi ha a la versió en paper.[6] Tanmateix, de vegades hi ha errors en la codificació de la base de dades, i cal corregir-ne la informació. A tall d'exemple, explicarem sota aquest epígraf quina informació hauria d'introduir-se a la base de dades perquè el programa d'anàlisi reconegués el mot *antecedente*, i com es correlaciona la informació dels diferents mòduls del diccionari perquè es pugui generar el fitxer de formes flexionades de manera satisfactòria.

[5]Tant el repertori d'etiquetes que contemplem com el significat d'aquestes estan recollits en el Working Paper de Morel, J. *et al*. (1998).

[6] Llevat dels aspectes relatius al gènere dels noms esmentats més amunt.

Segons llegim a *DALE*, aquesta paraula pot ser tant un nom masculí com un adjectiu; per tant, les etiquetes categorials que ha de dur són A i M. D'altra banda, és un mot que no presenta variació deguda al gènere i que afegeix *-s* per a formar el plural; segons els paradigmes de flexió que s'han tingut en compte per a confeccionar la versió electrònica del *DALE*, això significa que és una paraula que segueix el model de flexió 3.

Coneixent aquestes característiques de la paraula, qui volgués introduir-la al diccionari hauria d'incloure la següent informació:

MOT	antecedente
CODI_M	3
CATG	AM

Amb la codificació d'aquesta informació categorial i flexiva associada a la paraula, la tasca pròpia de qui volgués donar d'alta la paraula al diccionari ja s'hauria acabat. Tanmateix, perquè el programa informàtic que fa l'etiquetatge pugui utilitzar aquesta informació cal tornar a expandir el diccionari, ja que aquest programa se serveix de la informació del *DALE* expandit (més endavant, s'explica amb detall com es duu a terme l'expansió del diccionari).

L'expansió del diccionari procedirà, en primer lloc, a comprovar quines desinències corresponen al CODI_M que li hem donat a la paraula.[7] Per fer això, el programa informàtic encarregat de generar el diccionari expandit consultarà el fitxer DESI_MAN. En aquest fitxer s'especifica com flexionen les paraules de cadascun dels models, és a dir, quines desinències corresponen als diversos codis utilitzats a CODI_M; per al codi que hem assignat al mot *antecedente* (3), el programa trobaria que fa referència a dues desinències diferents: *0* i *-s*. Amb aquesta informació, el programa generaria les dues formes flexives de la paraula: *antecedente, antecedentes*.

En aquest punt del procés, cal un mecanisme que especifiqui quin valor té cadascuna de les formes de la paraula flexionada. Aquest està explicitat al fitxer CODIS_D, i la informació de DESI_MAN està vinculada amb la d'aquest fitxer gràcies a la coincidència dels codis del camp DERIV: cadascuna de les desinències que s'han fet servir per a generar les variants flexives d'un mot està aparellada amb un codi al camp DERIV que remet al camp d'idèntic nom del fitxer CODIS_D. En l'exemple del mot *antecedente*, la desinència *0* que ha generat la primera de les formes flexives està aparellada amb un codi que remet a un registre de CODIS_D que proporciona la informació de singular (amb gènere pendent d'assignar), i per al codi aparellat a la desinència *-s*, la informació de plural (amb gènere pendent d'assignar); aquestes informacions es tradueixen, a la pràctica, en l'assignació, respectivament, de les etiquetes *6S* i *6P*.

D'aquesta manera, a cada forma flexiva li queda assignada una etiqueta que n'especifica el gènere i el nombre, però no la categoria. Per tal de poder obtenir l'etiqueta final que s'assignarà a les diverses formes flexives d'una paraula, la informació de CODIS_D s'ha de combinar amb la informació categorial que

[7] A la figura 1 d'aquest annex hem il·lustrat la relació que s'estableix entre els diferents camps que intervenen en el procés d'expansió del diccionari.

prèviament s'ha inclòs al camp CATG de *DALE*. Com que *antecedente* pot ser, segons hem especificat a *DALE*, tant un nom com un adjectiu, la informació sobre el gènere i el nombre s'ha d'associar, consecutivament, amb la categoria d'*adjectiu* i la de *nom masculí*. De la combinació de la informació categorial A amb aquesta altra informació sobre la flexió se n'obtenen dues formes diferents:

antecedente adjectiu, gènere pendent, singular

antecedentes adjectiu, gènere pendent, plural

De la combinació de la informació categorial M amb la informació sobre la flexió se n'obtenen unes altres dues formes. Tal i com estan codificades a CODIS_D, les desinències *0* i *-s* deixen l'assignació del gènere pendent; tanmateix, en el cas dels noms, aquest es desprèn de la categoria; en el nostre cas, per exemple, *M* significa nom masculí. Per tant, les dues formes que es generarien serien:

antecedente nom, masculí, singular

antecedentes nom, masculí, plural

Amb això, el diccionari de formes flexionades per al castellà tindria quatre registres (dos com a adjectiu i dos com a substantiu) per al mot *antecedente*, que constitueixen una única entrada al diccionari de formes no flexionades.

8.3.4 L'expansió del diccionari

L'expansió del diccionari de formes flexionades del castellà es fa automàticament a partir del diccionari de formes no flexionades. El procés d'expansió del diccionari no es pot fer amb la còpia del diccionari que hi ha a la xarxa (la còpia del disc *J:*), sinó que es fa en local. Per a dur a terme l'expansió, cal seguir les passes següents:

1) Copiar al directori C:\DBF_ES les versions dels fitxers DALE_BO.DBF o DNLE_BO.DBF que s'han d'actualitzar (les del directori J:\PRIVAT\DBF_ES), i canviar-los el nom a DALE.DBF i DNLE.DBF respectivament. També cal copiar-hi el fitxer DESI_MAN.DBF si s'hi han introduït canvis.

2) Des del directori C:\DBF_ES\EXPAN fer córrer el programa FLEXION.EXE. L'ordre per fer córrer aquest programa és:

- Per expandir *DALE*: flexion dale
- Per expandir *DNLE*: flexion dnle

3) Un cop expandit el diccionari, cal copiar-lo de nou al directori J\PRIVAT\DBF_ES. Els fitxers que cal moure en cada cas són:

a) DALE:
- Fitxer de lemes: DALE.DBF
- Fitxers de formes flexionades:
  DALE06.DBF
  DALE08.DBF
  DALE10.DBF
  DALE12.DBF

DALE14.DBF
DALE30.DBF

b) Diccionari de neologia:

- Fitxer de lemes: DNLE.DBF
- Fitxer de formes flexionades: DNLE_F.DBF

4) Un cop copiats els fitxers al disc *J:* cal fer una còpia del diccionari de lemes per si cal fer-hi modificacions en el futur (amb el nom DALE_BO.DBF o DNLE_BO.DBF segons correspongui).

8.4 Annex 4 : Incorporació de la neologia del castellà als diccionaris de l'IULA

8.4.1 Anàlisi del document del CT del qual es vol introduir la neologia

Es tracta de processar lingüísticament el document que hem escollit, de tal manera que s'assignin tots els lemes i les etiquetes gramaticals als mots inclosos en el diccionari. Aquest és un pas previ per detectar les paraules desconegudes.

8.4.1.1 Comprovació de l'estructura del document

En tots els casos, abans d'iniciar el preprocés del document, cal comprovar que no s'hagi afegit algun nou error de marcatge. Per aquest motiu, a l'entorn Unix, al servidor *morgana* passarem un nou parser al document sencer.

S'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el password de l'usuari.

Des de *morgana*, escriurem la següent ordre a la línia de comandes:

```
nsgmls –s nom_del_document.sgm
```

Si hi ha algun problema de marcatge, ens indicarà la mostra i la línia en què s'ha trobat el problema, que evidentment haurem de corregir en el mateix entorn Unix editant el fitxer (si es domina aquest entorn) o mitjançant un editor de textos (Crimsom, EditPad).

8.4.1.2 Preprocessament del document

Per tal de preprocessar el document, s'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el password de l'usuari.

Des de *morgana*, escriurem la següent ordre a la línia de comandes:

```
pretag1.pl -i nom_del_document.sgm -bd -q
```

Si volem analitzar el document *g00159.sgm* escriurem:

```
pretag1.pl -i g00159.sgm -bd
```

Un cop preprocessat el document hem d'obrir una finestra MS-DOS i començar l'anàlisi pròpiament lingüística.

8.4.1.3 Anàlisi lingüística del document escollit

Obrim una finestra MS_DOS.

A la línia de comandes escrivim:

```
ctget nom_del_document.sgm
```

Si volem analitzar el document 105 de medicina escriurem:

```
ctget m00105.sgm
```

En donar aquesta ordre, se'ns demanarà el nom d'usuari i el password que utilitzem a l'entorn Novell.

Un cop finalitzada l'anàlisi hem de donar l'ordre següent:

ctput nom_del_document.sgm

que en el nostre cas d'exemple seria

ctput m00105.sgm

En donar aquesta ordre, se'ns demanarà el nom d'usuari i el password que utilitzem quan treballem a l'entorn Unix.

Quan es dóna aquesta ordre, pot ser que a la pantalla aparegui algun missatge que ens indicarà que alguna/es mostra/es d'aquell document no s'han analitzat correctament. Si succeeix això, convindria revisar el fitxer c:\tmp\lemacial.dat i comprovar amb l'ajuda de la persona encarregada de corpus què pot ser el que està passant.

8.4.2 Consulta de la llista de mots no reconegut pel diccionari (errors i neologismes)

El següent pas és la consulta de la llista de paraules no reconegudes pel diccionari.

En l'entorn UNIX i des del servidor morgana cal donar l'ordre següent:

apilado1xl.pl -i nom_del_document.sgm –bd -q

Ex:

Per crear un fitxer amb la neologia del document m00105.sgm de medicina s'ha d'executar la comanda següent:

apilado1xl.pl -i m00105.sgm –bd -q

El resultat d'aquesta comanda és una llista on trobem les paraules de cada mostra del document que no han rebut cap anàlisi. El programa posa dues etiquetes de categoria gramatical per defecte i un signe d'interrogació a cada paraula desconeguda.

Per tal de comprovar si les paraules que no estan reconegudes són errors o neologismes ha de consultar-se amb exactitud el context d'aparició d'aquestes paraules. Per fer-ho, cal obrir les mostres corresponents des del Crimson Editor o un altre editor de textos.

Un cop obert el editor, s'ha d'anar a File > FTP > Open Remote

Aquesta selecció obre una finestra que dóna accés al servidor Morgana. Per accedir-hi, s'ha de posar la contrasenya que fem servir a l'entorn Unix. A continuació, s'han de buscar les mostres analitzades al directori:

usr3/iula/corpus/ÀREA/mostres 4

Ex:

Per consultar els context d'aparició dels mots desconegudes del document de medicina m00105 cal obrir totes les mostres d'aquest document. Aquestes mostres es troben al directori usr3/iula/corpus/medicina/mostres 4.

Un cop tenim les mostres obertes, es tracta de veure les paraules que no han rebut anàlisi mitjançant la funció Search > Find in files. A la finestra de cerca, escriurem ? (el signe d'interrogació).

La consulta del context d'aparició de cada paraula ens permetrà saber si es tracta de un error en el text o d'un neologisme.

8.4.3 Correcció d'errors i registre escrit de neologismes

8.4.3.1 Errors

Quan la paraula no ha estat reconeguda pot ser que es tracti d'un error d'ortografia o de marcatge. En aquest cas, cal corregir l'error a la mostra original que es troba al directori usr3/iula/corpus/ÀREA/mostres.

Les mostres a corregir s'obren des de Crimson Editor: File > FTP > Open Remote.

Un cop fetes les correccions, s'han de guardar els canvis amb Save.

Ex:

Si es vol corregir un error dins la mostra 00599mma.ey0 del document m00105 de medicina, cal obrir-la des del directori usr3/iula/corpus/medicina/mostres.

Cal comprovar que en modificar els originals no s'hagi afegit algun nou error de marcatge. Per aquest motiu, a l'entorn Unix, al servidor morgana passarem un nou *parser* al document sencer.

Des de UNIX, en la línia de comandes cal escriure l'ordre següent:

nsgmls -s nom del document.sgm

Ex: nsgmls –s m00105.sgm

Si no hi ha cap nou error de marcatge ens sortirà de nou la línia de comandes. Si hi ha algun problema, ens indicarà la mostra i la línia en què s'ha trobat el problema, que evidentment haurem de corregir, (tornar a obrir de nou la mostra amb Crimson, corregir els errors, tornar a guardar el document i tornar a passar el *parser*).

8.4.3.2 Neologisme

Finalment, un cop s'ha descartat que la paraula desconeguda no sigui cap error, s'està realment davant d'un neologisme a incorporar a la base de dades del CT.

Per entrar la neologia en castellà del CT de l'IULA utilitzem la base de dades DBASE on es recull la neologia detectada abans de ser incorporada definitivament al diccionari electrònic utilitzat per a l'anàlisi del castellà.

Per accedir a DBASE cal obrir una finestra MSDOS i escriure DBASE. S'obrirà el programa DBASE. Ara hem d'accedir a la base de dades. Des de la línia de comandes hem d'escriure:

use j:\public\mat_aux\dbf_asc\neo_es excl

brow

Quan no hi ha cap registre introduït prèviament escriurem a la línia de comandes:
insert

A la base de dades haurem d'entrar el lema del neologisme i la categoria gramatical del lema.

La base de dades té tres columnes.

- A la primera cal entrar el lema del neologisme que volem incorporar al diccionari neològic del castellà. (Vegeu l'annex 7 on s'especifiquen els criteris de lematització).
- A la segona columna hem d'entrar la categoria gramatical del lema. La codificació de la categoria gramatical del mot a entrar és la següent:

  adjectius qualificatius ..... A (ex. verde)

  noms masculins ..... M (ex. barbero)

  noms femenins ..... F (ex. cabra)

  noms invariables de gènere ..... C (ex. electricista)

  verbs ..... V (ex. tomar)

- A la tercera columna introduirem informació referent a les possibles irregularitats de gènere o nombre que presenti una paraula.

Ex:

- De *locus* fa el plural en *loci*

Un cop haguem entrat la neologia:

Esc (escape)

i a la línia de comandes escriurem quit

Un cop fets tots aquests passos, la neologia introduïda s'incorporarà al diccionari castellà de neologia del CT.

8.4.4 Incorporació dels neologismes al diccionari de neologia castellana del CT

Aquesta tasca es realitza un cop a la setmana.

8.4.5 Incorporació dels documents a la base de dades definitiva

D'aquesta manera, la següent setmana, el fitxer en què havíem detectat la neologia haurà de tornar-se a processar per tal que pugui incorporar-se definitivament a la base de dades del CT de l'IULA. Aquestes són les instruccions a seguir:

pretag1.pl -i nom_del_document.sgm -bd –q (entorn Unix)

ctget nom_del_document.sgm (entorn Novell)

ctput nom_del_document.sgm del (entorn Novell)

apilado1xl.pl -i nom_del_document.sgm -bd –q (entorn Unix)

TagCorpus.pl -i nom_del_document.sgm –q (entorn Unix)

El pròxim procés és la inclusió del document a la BD textual (bwanaNet). El protocol on s'explica com dur a terme aquest procés es troba a l'annex 12.

8.5 Annex 5: Etiquetari anglès per al tractament del textos del CT

		Interjeccion
Category		I

		Noun
Category		N
Case	Nominative Genitive Pending	N G 6
Number	Singular Plural Pending	S P 6
Type (in 1st position)	Used adverbially Pending	A 6

		Verb
Category		V
Mood	Indicativa Subjunctive Imperative Infinitive Auxmod Pending	D J R I A 6
Tense	Present Past Pending	R A 6
Number	Singular Plural Pending	S P 6
Person		1,2,3,6

		Preposition
Category		P

		Adverb
Category		D
Degree	Absolute Comparative Superlative Wh- Pending	A C S W 6

		Conjunction
Category		C
Type	Coordinating Subordinating	C S

		Participle
Category		H
Type	-ing -ed	1 2

		Determiner
Category		A
Type	Wh- Demonstrative Absolute Comparative Superlative Pending	W D A C S 6
Case	Nominative Genitive Pending	N G 6
Number	Singular Plural Pending	S P 6

		Adjective
Category		J
Degree	Absolute Comparative Superlative	A C S

		Pronoun
Category		R
Type	Demonstrative Wh- Personal Reflexive Reciprocal Pending	D W P X R 6
Subtype (1st position)	Interrogative Relative Pending	I L 6
Degree	Absolute Comparative Superlative	A C S
Gender	Masculine Femenine Pending	M F 6
Case	Nominative Genitive Accusative Pending	N G A 6
Subcase	Independent Pending	I 6
Number	Singular Plural Pending	S P 6

		Numeral
Category		M
Type	Cardinal Ordinal Fraction	C O F
Number	Singular Plural Pending	S P 6

8.6 Annex 6: *Catmorf*: Analitzador morfològic de català

8.6.1 Introducció

Aquest document és el manual d'usuari de l'analitzador morfològic per al català de l'IULA: *CATMORF*; un analitzador de dos nivells de cobertura àmplia per al català.

En aquest document presentem l'organització interna de l'analitzador, els passos que cal seguir per instal·lar-lo i el seu funcionament.

Cadascuna de les tasques a realitzar requereix uns coneixements de partida i la lectura de les seccions corresponents d'aquest document. Tot seguit especifiquem alguns d'aquests requeriments:

- Per instal·lar el software, és convenient tenir coneixements en instal·lació de software sota Unix i coneixements de Prolog.
- Per usar el sistema, cal estar familiaritzat amb el sistema operatiu Unix i és recomanable tenir experiència com a usuari de Prolog, encara que no imprescindible.
- Per afegir entrades lèxiques al sistema, calen uns coneixements mínims de morfologia catalana, una certa familiaritat amb el sistema operatiu Unix, el coneixement d'algun editor en aquest entorn i són recomanables alguns coneixements de Prolog. Prèviament és necessari familiaritzar-se amb el formalisme de les entrades.
- Per augmentar la cobertura del sistema, caldrà modificar la gramàtica de la paraula i les regles de dos nivells. En aquest cas, cal tenir coneixements de morfologia de dos nivells i una certa experiència en el camp de la lingüística computacional.
- Per fer que l'analitzador generi unes etiquetes morfosintàctiques diferents de les actuals caldrà tenir coneixements de Prolog. (Les etiquetes que es generen actualment segueixen el patró descrit a Morel *et al*., 1997).

Finalment, per augmentar l'eficiència del sistema o per modificar-ne substancialment el funcionament intern, calen també coneixements de Prolog.

8.6.2 Breu explicació de CATMORF

CATMORF és un analitzador morfològic de dos nivells de cobertura àmplia per al català. [1] Les seves característiques principals són:

- Implementació en Sicstus Prolog.
- Cobertura completa de la flexió nominal, adjectival i verbal i d'alguns processos derivatius (alguns casos de prefixació, alguns casos de formació de diminutius i la formació d'adverbis acabats en *-ment*).[2]

Des del punt de vista d'estratègia de processament, el sistema rep com a entrada una paraula i retorna la paraula analitzada morfològicament, però no realitza cap mena de preprocés amb els textos ni cap mena de desambiguació.

8.6.3 Organització dels fitxers de CATMORF

En instal·lar l'analitzador morfològic CATMORF es generen una sèrie de directoris i fitxers amb el següent contingut:

```
%% ls -F
      Afegir_Lexic/           ind_pref_startup_dlc.pl
      pref_startup.pl
      Documentacio/           ind_pref_startup_dlc.ql
      pref_startup.ql
      Driver/                 ind_pref_startup_diec.pl
      LEXIC/                  ind_pref_startup_diec.ql
      Regles_Tlm/
      WG+Etiq/
```

[1] Vegeu Sproat (1992) per a una bona introducció a la morfologia de dos nivells, i en general, a la morfologia computacional.

[2] CATMORF s'ha provat a l'IULA amb la utilització de dos diccionaris:

- El diccionari DIEC (1995) de 71000 entrades, de les quals 68000 són lemes.

- El diccionari DLC83 + DLC93 +DIEC que conté 85000 entrades, de les quals 82000 són lemes. Per a una descripció més detallada del contingut d'aquest lexicó pot consultar-se Bach *et al*. (1997).

- Afegir_Lexic: Conté el programa *AFEGIRLEX* per a crear noves entrades lèxiques en un format apte per a *CATMORF*.
- Documentacio: Conté la documentació relacionada amb *CATMORF*, entre la qual trobem aquest manual.
- Driver: Conté l'analitzador pròpiament; és a dir, els programes que apliquen les regles de dos nivells i fan cerques al lèxic (tant per a noms i adjectius com per a verbs, i tant si el lèxic està en memòria com en disc).
- LEXIC: Conté el(s) diccionari(s) que formen el lèxic del sistema. El lèxic pot presentar-se en dues versions: en disc (format Dbm) i/o amb els fitxers prolog. Consulteu els fitxers *README* d'aquests directoris.
- Regles_Tlm: Conté les regles de dos nivells, tant les que s'apliquen a noms i adjectius com les que s'apliquen als verbs.
- WG+Etiq: Conté la gramàtica de la paraula de *CATMORF* i el generador d'etiquetes (per a cada paraula analitzada es genera una etiqueta segons l'etiquetari definit a Morel *et al*., (1997)).
- ind_pref_startup_dlc.pl: Fitxer que carrega *CATMORF* amb el lèxic del diccionari Dlc(83) + ampliacions. Està indexat en disc.
- ind_pref_startup_dlc.ql: La versió compilada, que es pot cridar des d'altres aplicacions.
- ind_pref_startup_diec.pl: Fitxer que carrega *CATMORF* amb el lèxic del diccionari Diec. Està indexat en disc.
- ind_pref_startup_diec.ql: La versió compilada, que es pot cridar des d'altres aplicacions.
- pref_startup.pl: Fitxer que carrega *CATMORF* amb lèxic del diccionari en memòria. [3]
- pref_startup.ql: La versió compilada, que es pot cridar des d'altres aplicacions.

A continuació expliquem amb una mica més de detall el contingut d'algun d'aquests directoris. Aquesta informació és especialment útil en cas que es vulguin afegir i/o canviar entrades lèxiques o modificar/substituir el generador d'etiquetes.

8.6.3.1 El subdirectori LEXIC

El subdirectori *Diec* conté tot el lèxic extret semiautomàticament del diccionari normatiu de DIEC (1995), mentre que el directori *Dlc+Lx93+Lx95* conté el lèxic extret del diccionari DLC(83) més les diferències que hi ha en el DLC(93) més les del DIEC.

A continuació s'explica l'estructura d'aquests directoris, que és simètrica.

[3] La versió en memòria comporta una càrrega excessiva del sistema. Per aquest motiu només funciona amb el diccionari del DIEC. Si es vol utilitzar una versió en memòria que utilitzi el Dlc per a fer les anàlisis cal modificar un *script* del programa, per la qual cosa es pot contactar amb l'IULA.

```
%% ls -F

Diec/   Dlc+Lx93+Lx95/  README
```

```
%% ls -F

Dbm/  Pl+Ql/
```

El directori *Dbm* conté la versió indexada en disc que ha estat construïda amb la llibreria *dbm* de Sicstus Prolog. El directori *Pl+Ql* conté els fitxers del lèxic amb extensions *.pl* i *.ql*. S'ha de tenir en compte que la versió indexada es construeix a partir dels fitxers *.pl* i *.ql*.

El contingut del directori *Dbm* és el següent:

```
%% ls -F

nmenys/                 tancades/                       verbs/
creacio_index_noms.pl   creacio_index_tancades.pl
creacio_index_verbs.pl
```

El directori *nmenys* conté els noms i adjectius dels diccionaris indexats. El directori *tancades* conté les paraules que pertanyen a categories tancades (preposicions, determinants ...). El directori *verbs* conté els verbs del diccionari.

Finalment, el fitxers *creacio_index_noms.pl*, *creacio_index_tancades.pl* i *creacio_index_verbs.pl* són els que permeten afegir entrades lèxiques al diccionari.

El contingut del directori *Pl+Ql* és el següent:

```
%% ls -F

adverbis.pl             verbs_1a.pl
adverbis.ql             verbs_1a.ql
nmenys.pl               verbs_2a.pl
nmenys.ql               verbs_2a.ql
sufixos_nominals.pl     verbs_3a.pl
sufixos_nominals.ql     verbs_3a.ql
tancades.pl             formes.pl
tancades.ql             formes.ql
```

El lèxic s'ha distribuït en una sèrie de fitxers; a la capçalera dels quals es pot trobar una explicació del seu contingut. El gruix de noms i adjectius es troba a *nmenys.pl*.

El contingut del subdirectori WG+Etiq és aquest:

```
%% ls -F

catalan_word_grammar2.pl      gen_etiquetes2.pl
catalan_word_grammar2.ql      gen_etiquetes2.ql
ind_catalan_word_grammar2.pl
ind_catalan_word_grammar2.ql
```

El fitxer *catalan_word_grammar2.pl* és la gramàtica de la paraula que tracta la flexió nominal i verbal, mentre que el fitxer *deriv_nominal_WG.pl* és la gramàtica de la paraula que tracta alguns casos de prefixació i la formació d'adverbis que acaben en *-ment*.

El fitxer *gen_etiquetes2.pl* és el programa que transforma la sortida de *CATMORF* en una etiqueta morfosintàctica, segons l'etiquetari descrit a Morel *et al*., (1997).

8.6.4 Requeriments i instal·lació

Aquesta secció està pensada com a guia d'instal·lació de *CATMORF*. Tot i que la instal·lació és molt senzilla és convenient que qui se n'encarregui tingui coneixements de Prolog i experiència en instal·lació de software en UNIX.[4]

8.6.4.1 Requeriments per a la instal·lació

La configuració ideal que permet la instal·lació i el funcionament correcte de *CATMORF* és la següent:

Arquitectura: Sparc 20 o Sparc 10
Sistema Operatiu: Solaris 2.3 (UNIX per a màquines Sparc de SUN)
Prolog: Sicstus Prolog, versió 3.#2
Espai de disc: 203 Mb per a la versió completa

[4]Cal també consultar la secció 3 per a una òptima instal·lació del sistema.

Ara bé, si no es compleix algun d'aquests requeriments, CATMORF pot funcionar també adequadament sempre que es tinguin en compte els següents aspectes:

- *CATMORF* corre sobre *Sicstus Prolog*; si no es té aquest Prolog se'n pot fer servir un altre, però caldrà revisar les parts *potencialment conflictives* dels fitxers que estan en Prolog. Caldrà consultar:[5]

  - Les declaracions *multifile*

  - Les declaracions *use_module*

  - Les declaracions *dynamic*

  - Les crides a llibreries de Sicstus (llistes i dbm)

- No cal que el sistema operatiu sigui *Solaris*, però es necessita un sistema operatiu on es pugui instal·lar correctament el Prolog i que permeti treballar amb un gran volum de dades; probablement serà alguna variant de UNIX.

- Evidentment, no cal que la màquina utilitzada sigui una Sparc 20; simplement cal que a la màquina es pugui instal·lar el Prolog, amb el sistema operatiu adient i amb l'espai en disc i en memòria suficient.

- La versió *COMPLETA* ocupa 203 Mb d'espai en disc, però aquest espai es pot reduir sensiblement:

  - Eliminant completament algun dels diccionaris.

  - Eliminant parcialment algun dels diccionaris; és a dir, eliminant la versió amb el lèxic indexat o eliminant la versió en memòria.

8.6.4.2 Instal·lació

CATMORF es distribueix com un fitxer *tar* comprimit (*catmorf.tar.gz*). Per descomprimir-lo i obtenir-ne tots els fitxers des de la línia de comandes cal executar l'ordre següent:[6]

```
zcat catmorf.tar.gz | tar xf -
```

[5] Cal tenir present que és possible que aquesta llista no sigui completa.
[6] S'ha de tenir en compte que la comanda *zcat* ha de ser la de *GNU*, i no la que ve com a estàndard en la majoria de les màquines UNIX.

Un cop descomprimida tota la distribució, i si tot ha anat correctament, es poden eliminar els fitxers que no siguin necessaris, si es creu convenient, i ja es pot utilitzar l'analitzador *CATMORF* sempre que es disposi del lèxic en algun dels formats adequats.

8.6.5 Com fer servir CATMORF

CATMORF es pot utilitzar de dues maneres diferents: a) des de la línia de comandes (de forma interactiva) i b) des d'una altra aplicació.

Sigui quina sigui l'opció escollida, per poder utilitzar *CATMORF*, en primer lloc s'ha de carregar el programa i l'índex del diccionari a partir del qual es vol fer l'anàlisi. En segon lloc, cal interrogar el sistema.

Per interrogar l'analitzador morfològic es poden fer servir dos predicats diferents:
(1) interrog(+ Paraula, - Etiqueta)
tant *Paraula* (variable instanciada) com *Etiqueta* (variable sense instanciar) han de ser *àtoms*.
(2) morf2n(+ Paraula, - Etiqueta)
tant *Paraula* (variable instanciada) com *Etiqueta* (variable sense instanciar) han de ser *strings*.

Com es pot observar, la única diferència és que en un cas cridem l'analitzador morfològic amb un *àtom* i en l'altre cas amb un *string*.

Presentem tot seguit un esquema dels passos a seguir per usar *CATMORF* :

1. Situar-se al directori principal de *CATMORF*:

```
%% cd directori_de_catmorf
```

2. Carregar el Prolog:

```
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?-
```

3. Carregar tot el sistema. Hi ha dues opcions:

a) carregar la versió del diccionari en memòria (les anàlisis seran més ràpides)

```
?- [ppref_startup].
{consulting /usr3/iula/soft/catmorf/ppref_startup.pl...}
{loading /usr3/iula/soft/CATMORF/Driver/catmorf.ql...}
...
...  molts missatges
Carregats els especificadors, preposicions, ...
{/usr3/iula/soft/CATMORF/ppref_startup.pl consulted, ...}
yes
| ?-
```

b) carregar la versió indexada (es carregarà amb més rapidesa però les anàlisis seran més lentes).

Per carregar la versió indexada cal substituir en la comanda *pref_startup* per *ind_pref_startup_dlc.pl*, si es vol utilitzar el diccionari modular per a les anàlisis, o *ind_pref_startup_diec.pl* en cas de voler analitzar el text a partir del lèxic del DIEC.

4. Interrogar el sistema (o dit d'una altra manera, fer una anàlisi).

```
| ?- interrog('casa',Etiqueta).
Etiqueta = '#casa\\casa:N5-FS\\casar:V8R6S-#' ?
yes

| ?- interrog('cases',Etiqueta).
Etiqueta = '#cases\\casa:N5-FP\\casar:VDR2S-#' ?
yes
| ?-
```

Cada etiqueta consta d'una primera part on apareix la forma que es vol analitzar, i a continuació totes les seves possibilitats morfològiques:

5. Finalitzar:

```
| ?- halt.
```

8.6.6 Modificació del lèxic de CATMORF

L'ampliació del lèxic de *CATMORF* es pot fer o bé utilitzant el programa *AFEGIRLEX*[7] o bé de forma manual.

A continuació expliquem el procés d'incorporació d'entrades al diccionari utilitzat per CATMORF de forma manual.

Una de les tasques més habituals en un analitzador morfològic és l'ampliació i/o modificació del seu lèxic; en el cas de *CATMORF* distingirem els següents casos:

- Ampliació i/o modificació d'entrades verbals.
- Ampliació i/o modificació de noms i adjectius.
- Ampliació i/o modificació de paraules que pertanyen a categories tancades.

En tots els casos, els passos que s'han de seguir seran aquests:

1. Modificació o creació d'entrada o entrades ja existents en els fitxers del lèxic.
2. Compilació del(s) fitxer(s) que contenen el lèxic.
3. Indexació del(s) fitxer(s) que contenen el lèxic.

Seria convenient que en aquestes tasques es tingués la supervisió d'algú que tingui coneixements de Prolog, almenys al principi.

Les entrades lèxiques de *CATMORF* són termes prolog que tenen el següent format:

```
Functor(+Entrada,+Bloqueig,+Info_Morf,+Vocals,+Vocals_acc).
```

[7]Aquest programa s'ha dissenyat en el marc del projecte de recerca de *Llenguatges Especialitzats. Corpus Multilingüe* (CIRIT CS93-4.009).

on la interpretació de cadascun dels arguments és la següent:

- `Functor.` És el nom del predicat de les entrades del diccionari, que segons la seva naturalesa pot ser:

  ```
  ln (per a les entrades nominals)
  lv (per a les entrades verbals)
  lex_tanc (per a les entrades de forma tancada)
  ```

  - `+Entrada`. Ha de ser un àtom instanciat; és la clau per accedir a l'entrada, i habitualment serà el seu lema.
  - `+Bloqueig`. És una llista de les TLR que l'entrada bloqueja. El bloqueig de regles s'especifica mitjançant parells: *nom_regla:n*. Si no hi ha cap regla a bloquejar la llista és buida.
  - `+Info_Morf`. Ha de ser un terme prolog, i representa la informació morfològica de l'entrada.
  - `+Vocals`. Ha de ser un número; especifica el nombre de vocals que apareixen a la grafia de l'entrada.
  - `+Vocals_acc`. Ha de ser un número; especifica el nombre de vocals accentuades que apareixen a la grafia de l'entrada.

Les entrades lèxiques corresponents a classes diferents es diferenciaran en el *functor* i en el camp *Info_Morf*. Les pròximes subseccions aclariran aquests punts.

8.6.6.1 Ampliació i/o modificació d'entrades verbals

En el cas que es vulguin afegir entrades verbals, aquestes hauran de ser necessàriament de la primera conjugació (la segona i la tercera són conjugacions tancades i per això no s'ha previst que puguin ampliar-se).

a) Per afegir una nova entrada verbal, s'ha d'editar el fitxer d'entrades verbals de la primera conjugació i duplicar l'entrada de *cantar*:

```
lv('cant',[],
    morf(
         cat(arrel_verbal),
         v(v-intr-tr),
         conj(1),model(1),buit(no),
         lema('cantar'),temps(... moltes coses)
        )
  ).

lv('cant',[],
    morf(
         cat(arrel_verbal),
         v(v-intr-tr),
         conj(1),model(1),buit(no),
         lema('cantar'),temps(... moltes coses)
        )
  ).
```

b) A continuació cal substituir les dades que apareixen de *cantar* per les de la nova entrada. Per exemple, suposem que es vulgui afegir al lèxic el verb (impossible) *pklftar*:

```
lv('cant',[],
    morf(
         cat(arrel_verbal),
         v(v-intr-tr),
         conj(1),model(1),buit(no),
         lema('cantar'),temps(... moltes coses)
        )
  ).

lv('pklf',[],
    morf(
         cat(arrel_verbal),
         v(v-intr-tr),
         conj(1),model(1),buit(no),
         lema('pklfar'),temps(... moltes coses, que es deixen
igual)
        )
  ).
```

Per afegir una nova entrada verbal d'aquesta manera, només cal substituir els camps corresponents a l'entrada i al lema. Per afegir entrades lèxiques verbals, no cal saber interpretar el significat dels altres camps.

c) A continuació cal compilar el fitxer que conté el lèxic dels verbs:

```
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- fcompile('verbs_1a').
{fcompiling ...verbs_1a.pl}
  ...
yes
| ?-
```

Si la compilació ha anat correctament, cal indexar el lèxic que hem creat. Per això, cal situar-se al directori on es troba el lèxic que es vol indexar, esborrar el directori que correspon als verbs i crear el nou lèxic:

```
%% cd LEXIC/Diccionari_escollit/Dbm
%% rm -r verbs
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- [creacio_index_verbs].
{loading ...}


{Warning: ... - goal failed}

yes
| ?-
```

Ara ja es pot utilitzar *CATMORF* amb el nou lèxic creat.

8.6.6.2 Ampliació i/o modificació de noms i adjectius

Afegir noms i adjectius al lèxic és una tasca més complicada que la d'afegir verbs. Com a consell pràctic recomanem que abans de començar a afegir entrades

s'estudiïn detingudament algunes de les entrades que es troben al lèxic i es comparin amb les entrades corresponents dels diccionaris en paper.

A mode d'exemple, suposem que es vulgui afegir l'entrada del nom *nòvio* i de l'adjectiu *aconseguible*. Les entrades corresponents seran aquestes (compareu amb *casa*):

```
ln('casa',[],
   morf(
        lema('casa'),
        cat(nmenys),
        class(nom),
        gen(f),nom(sing),
        flex(gen(no),nom(si))
       ),
   2,0).

ln('nòvio',[],
   morf(
        lema('nòvio'),
        cat(nmenys),
        class(nom),
        gen(m),nom(sing),
        flex(gen(si),nom(si))
       ),
   2,1).

ln('aconseguible',[],
   morf(
        lema('aconseguible'),
        cat(nmenys),
        class(adj),
        gen(mf),nom(sing),
        flex(gen(no),nom(si))
       ),
   2,0).
```

En el cas de les entrades nominals és important tenir en compte el significat de cadascun dels camps de les entrades:

- el primer camp és l'entrada (*casa* o *nòvio*),
- el segon camp és una llista de regles de dos nivells a bloquejar (en aquests casos la llista és buida, però després veurem exemples on no és així),
- el tercer camp és un camp compost que conté informació morfosintàctica de l'entrada (lema, classe, gènere i nombre, si flexiona en gènere i nombre).

En l'exemple mostrat, *aconseguible* no flexiona en gènere (s'aplica tant a masculins com a femenins) i només pot ser adjectiu. *nòvio*, en canvi, sí que flexiona pel gènere (considerem que el lema de *nòvia* és *nòvio*).

Algunes entrades tenen el mateix paradigma tant si funcionen com a noms com si funcionen com a adjectius; en aquests casos les entrades tindran aquest aspecte:

```
ln('espanyol',[],
   morf(
        lema('espanyol'),
        cat(nmenys),
        class(nom-adj),
        gen(m),nom(sing),
```

```
        flex(gen(si),nom(si))
       ),
   2,0).
```

Hi ha casos en què tenir tota aquesta informació en una única entrada no és possible; per exemple, *percussor* només flexiona pel gènere si l'utilitzem com a adjectiu, però com a nom només és masculí. En aquests casos cal duplicar l'entrada, modificant els camps morfosintàctics necessaris:

```
ln('percussor',[],
   morf(
        lema('percussor'),
        cat(nmenys),
        class(adj),
        gen(m),nom(sing),
        flex(gen(si),nom(si))
       ),
   2,0).

ln('percussor',[],
   morf(
        lema('percussor'),
        cat(nmenys),
        class(nom),
        gen(m),nom(sing),
        flex(gen(no),nom(si))
       ),
   2,0).
```

Com hem indicat anteriorment, el segon camp en les entrades nominals marca quines són les regles de dos nivells que cal bloquejar en cada cas. Si anteriorment hem vist entrades en què aquesta informació no era pertinent (perquè no s'havia de bloquejar cap regla), a continuació, i a mode d'exemple, presentem entrades on aquesta informació és necessària:

```
ln('absolut',[perduda:n],
   morf(
        lema('absolut'),
        cat(nmenys),
        class(nom),
        gen(m),nom(sing),
        flex(gen(si),nom(si))
       ),
   2,0).
```

L'entrada corresponent a *absolut* bloqueja la regla *perduda*. El motiu pel qual cal afegir aquesta informació és que el femení de *perdut* (*perduda* es considera *obligatori* a *CATMORF*, mentre que el femení d'*absolut*, *absoluta*, és *irregular* per a l'analitzador morfològic. Per poder donar compte de tots dos fenòmens ha calgut bloquejar l'aplicació d'una de les regles.

8.6.6.2.1 Selecció de patrons de bloqueig

Per poder saber quan cal aplicar un patró de bloqueig, presentem un algoritme que permet decidir la selecció de patrons de bloqueig per a una entrada determinada:

(1) Comprovar si l'entrada que es vol afegir segueix algun dels patrons de cerca de la taula de la figura 1. En cas contrari, l'entrada no necessita cap bloqueig.

(2) Si les formes resultants de la flexió de l'entrada en gènere o nombre segueixen el patró definit a la columna *flexió*, s'ha d'assignar el patró definit a la columna *Patró de bloqueig*.

La taula de la *figura 2* explicita per a cada bloqueig la llista de regles a bloquejar.

A continuació mostrarem alguns exemples de selecció de patrons per a algunes entrades:

(1) Estudiem la paraula *abús*. Aquesta segueix el patró de cerca 1 definit a la figura 1 perquè a les seves formes flexionades no hi ha duplicació de "s" (*abusos*). Per aquesta raó li assignem el patró de bloqueig 6.

(2) Estudiem la paraula *espès*. Aquesta paraula també segueix el patró de cerca 1 (figura 1) però com que en les seves formes flexionades es duplica la "ss" (a partir d'*espès* obtenim *espessa* i *espessos*, per exemple), no cal assignar-li cap patró de bloqueig.

Patró de cerca	Expressió regular	Flexió	Patró de bloqueig	Exemple
patro_1	[Consonant][à|é|è|ís|ò|ú]s$	!(.ss.)	bloc_6	abús
patro_2	.[^q][Vocal][í|ú]s$	!(.ss.)	bloc_1	país
		(.ss.)	bloc_2	suís
patro_3	[Vocal]ig$	!(.tj.)	bloc_5	boig
		(.tj.)	bloc_4	lleig
patro_4	[Vocal|ï]t$	!(.da$)	bloc_15	absolut
patro_5	[Consonant][à|é|è|í|ó|ò|ú]$	!(.ns$)	bloc_11	tabú
patro_6	[Vocal|l|n|r|s]c$	(.gues$)	bloc_17	amic
		(.*qües$)	bloc 16	oblic
		per defecte	bloc_3	amnèsic
patro_7	[Consonant][Vocal]l$	!(.l·la$)	bloc_14	amonal
patro_8	[Consonant][Vocal]s$	!(.ss.)	bloc_7	cas
patro_9	quí$	!(.ns.)	bloc_12	esquí
patro_10	què s$	per defecte	bloc_8	iroquès
patro_11	[Consonant][Vocal]x$	!(.os$)	bloc_13	clímax
patro_12	[Consonant]{1-3}[vocal] $	!(.ns$)	bloc_19	te
patro_13	[Consonant]{1-3}a$	(.ns$)	bloc_10	gra
patro_14	[Consonant]{1-3}[e|o]$	(.*ns$)	bloc_9	fre
patro_15	[a|i|o]u$	!(.va$)	bloc_18	nadiu

Figura 1. Patrons de cerca i bloqueigs de regles

Estudiem les paraules *sec, amic, oblic*. Totes tres segueixen el patró 6. Observem que hi ha tres casos possibles:

- En les formes flexionades apareix el patró "gues" (per exemple, *amigues*). En aquest cas, el bloqueig a assignar és el 17.

- En les formes flexionades apareix el patró "qües" (per exemple, *obliqües*). En aquest cas, el bloqueig a assignar és el 16.

- En cas que no es segueixi cap dels patrons anteriors (per exemple, *seques*), el patró de bloqueig a assignar és el 3.

Patró de bloqueig	Llista de regles
bloc_1	[abús_0:n,duplic_s_3:n]
bloc_2	[abús_0:n]
bloc_3	[amic_1:n,oblic_11:n,oblic_12:n,aparicio_u_12:n,aparicio_u_13:n]
bloc_4	[boig_1:n,boig_2:n,boig_3:n,boig_4:n]
bloc_5	[lleig_1:n,lleig_2:n,lleig_3:n,lleig_4:n]
bloc_6	[duplic_s_1:n]
bloc_7	[duplic_s_2:n]
bloc_8	[duplic_s_3:n]
bloc_9	[e_absorcio_2:n]
bloc_10	[ea_2:n,e_absorcio_3:n]
bloc_11	[elim_n_1:n,pagana:n]
bloc_12	[elim_n_3:n,marroquí:n]
bloc_13	[elim_o_2:n]
bloc_14	[gal:n]
bloc_15	[perduda:n]
bloc_16	[seques:n,amic_1:n,aparicio_u_11:n]
bloc_17	[seques:n,oblic_11:n,oblic_12:n,aparicio_u_12:n,aparicio_u_13:n]
bloc_18	[blava:n]
bloc_19	[elim_n_4:n]

Figura 2. Patrons de bloqueig i llistes de regles associades

Un cop ja s'han afegit totes les entrades lèxiques, cal compilar el fitxer que conté el lèxic dels noms i adjectius:

```
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- fcompile('nmenys').
{fcompiling ...nmenys.pl}

yes
| ?-
```

Si la compilació s'ha fet correctament, cal indexar el lèxic que hem creat. En primer lloc, s'ha d'anar al directori on es troba el lèxic que es vol indexar, esborrar el directori que correspon als noms i adjectius i crear el nou lèxic:

```
%% cd LEXIC/Diccionari_escollit/Dbm
%% rm -r nmenys
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- [creacio_index_noms].
{loading ...}


{Warning: ... - goal failed}

yes
| ?-
```

Ara ja es pot utilitzar *CATMORF* amb el nou lèxic creat.

8.6.6.3 Ampliació i/o modificació de paraules que pertanyen a categories tancades

El procediment és similar als anteriors.

Suposem que es decideix ampliar el lèxic afegint l'adverbi *mateixament*. Primer generem l'entrada corresponent a *mateixament*:

```
lex_tanc(zelosament,[],morf(cat(adv),lema(zelosament),"D4"),4,0).
...
lex_tanc(mateixament,[],morf(cat(adv),lema(mateixament),"D4"),4,0)
.
```

A continuació hem de compilar el fitxer que conté el lèxic dels adverbis:

```
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- fcompile('adverbis').
{fcompiling ...adverbis.pl}

yes
| ?-
```

Si la compilació ha anat correctament, cal indexar el lèxic creat. Primer cal anar al directori on es troba el lèxic que es vol indexar, esborrar el directori que correspon als adverbis i crear el nou lèxic:

```
%% cd LEXIC/Diccionari_escollit/Dbm
%% rm -r tancades
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- [creacio_index_tancades].
{loading ...}


{Warning: ... - goal failed}

yes
| ?-
```

Ara ja es pot utilitzar *CATMORF* amb el nou lèxic creat.

8.6.6.4 Criteris per a la lexicalització de formes

És possible que tot i seguir els criteris de selecció de patrons de bloqueig, algunes formes d'un lema no es reconeguin. Això pot passar per determinats lemes i formes que segueixen un comportament flexiu totalment irregular. Per exemple, existeix un lema *fa* que fa el seu plural *fas*. En aquests casos la solució és la següent:

- Lexicalitzar la forma *fas*, indicant que no flexiona en nombre, que és plural i que el seu lema és *fa*.
- Lexicalitzar *fa* i especificar a l'entrada que no flexiona pel plural (això evitarà que es reconegui *fes* com a forma flexionada de *fa*.

L'aspecte d'aquestes entrades seria aquest:

```
ln('fas',[],
   morf(
        lema('fa'),
        cat(nmenys),
        class(nom),
        gen(f),nom(p),
        flex(gen(no),nom(no))
       ),
   2,0).

ln('fa',[],
   morf(
        lema('fa'),
        cat(nmenys),
        class(nom),
        gen(f),nom(sing),
        flex(gen(no),nom(no))
       ),
   2,0).
```

Tot i que hem tractat el tema de les lexicalitzacions de formes amb un exemple particular, considerem que aquest exemple il·lustra els casos en els quals *caldrà fer lexicalitzacions* d'aquest tipus.

8.6.6.5 Ampliació i/o modificació de formes lexicalitzades

Per augmentar la rapidesa del sistema, s'ha incorporat un fitxer en què hem incorporat aquelles paraules d'ús més freqüent en el corpus de l'IULA. Aquest fitxer conté parells de paraula analitzada i la seva etiqueta associada.

El fitxer es troba a cadascun dels directoris *Pl+Ql* dels diccionaris i es diu *formes.pl*. La versió indexada en disc es troba sota el directori *Dbm* de cadascun dels diccionaris i es diu *formes*.

El fitxer *formes.pl* conté termes prolog que tenen la següent estructura:

```
l(+Paraula,+Etiqueta).
```

Tant *Paraula* com *Etiqueta* han de ser àtoms. Per exemple, les entrades corresponents a *tot* i *poden* són les següents:[8]

```
l('tot','#tot\\tot:EN--MS\\tot:N5-MS\\tot:D4#').
l('poden','#poden\\poder:VDR3P-\\podar:VDR3P-#').
```

Aquest fitxer es pot ampliar amb noves formes, i només s'ha de respectar el format de les entrades. Com la resta de fitxers de lèxic, un cop s'han afegit les entrades, s'ha de compilar el fitxer:

```
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- fcompile('formes').
{fcompiling ...formes.pl}

yes
| ?-
```

Posteriorment cal indexar aquest lèxic:

```
%% cd LEXIC/Diccionari_escollit/Dbm
%% rm -r formes
%% sicstus
SICSTUS 3  #2: Mon Feb 12 17:08:17 WET 1996
| ?- [creacio_formes].
{loading ...}
...
...
{Warning: ... - goal failed}

yes
| ?-
```

Ara ja es pot utilitzar *CATMORF* amb el nou lèxic creat.

8.6.7 Llista de BUGS

A continuació s'exposa la llista de Bugs que hem trobat:

1. *Sicstus Prolog* no permet que hi hagi un número superior a 65535 clàusules d'un *mateix predicat* compilat en memòria (és un bug de Sicstus); per tant, caldrà fer servir la versió indexada en disc si s'utilitzen tots els noms i adjectius del diccionari *DLC + Ampliacions* i/o s'amplia el lèxic i es sobrepassa aquesta xifra amb algun dels predicats.
2. Si es vol cridar l'analitzador morfològic des de *Perl* es recomana utilitzar la versió de l'analitzador amb el lèxic indexat en disc; la versió amb el lèxic en memòria ha causat problemes quan es carregava amb altres mòduls Prolog si tots es cridaven des de *Perl*.
3. El sistema és bastant lent. Si es necessita que sigui molt més ràpid, es recomana tenir indexades en disc les X (on X podria ser igual a 200.000) paraules més freqüents (analitzades per *CATMORF*), i que s'utilitzi aquest índex conjuntament amb l'analitzador morfològic per al processament de

[8] Consulteu Morel *et al*., (1997) per interpretar les etiquetes morfosintàctiques.

textos. D'aquesta manera, la part lenta del sistema (l'analitzador) només hauria d'analitzar paraules poc freqüents.

8.7 Annex 7: Criteris de lematització del CT

En aquest document es presenten els criteris que se segueixen en CT a l'hora d'assignar lemes a les peces lèxiques analitzades.

1. MOTS GRAMATICALS

- Formes gramaticals que presenten flexió: (especificadors, determinants, adjectius quantificadors, indefinits i possessius)

Els assignem el lema que correspon a la forma de masculí singular, la forma menys marcada:

unes ..... *un*

ambdues ..... *ambdós*

- Formes gramaticals invariables[1]

Els assignem com a lema la forma que presenten:

abans ..... *abans*

però ..... *però*

millor (adv.) ..... *millor*

laberínticament ..... *laberínticament*

en lloc de ..... *en lloc de*

de debò ..... *de debò*

- Pronoms personals

Els assignem el lema *pr*:

jo ..... *pr*

me ..... *pr*

si ..... *pr*

2. FORMES NOMINALS: (noms i adjectius, excepte els noms propis marcats en el preprocés <NAME>)

- Noms i adjectius que presenten formes diferents per al masculí i per al femení.

Els assignem com a lema el masculí singular; la forma menys marcada:

fills ..... *fill*

filles ..... *fill*[2]

blaves ..... *blau*

[1]Atenció: Incloem sota aquest grup les locucions, que es tracten en l'etapa del preprocés, anterior a la de la lematització.

[2]Tot i que la lematització dels substantius podria fer-se respectant el fet que el gènere és un tret inherent al nom i per tant lematitzant independentment els noms masculins i el femenins coincidents per la forma, en aquest projecte hem optat per lematitzar-los junts sota la forma de masculí singular.

pobra ..... *pobre*

- Noms i adjectius que presenten una sola forma per al masculí i el femení

Els assignem com a lema la forma que correspon al singular:

acompanyant ..... *acompanyant*

supranacional ..... *supranacional*

simples ..... *simple*

- Noms només masculins o només femenins

Els assignem com a lema la forma que correspon al singular:

casa ..... *casa*

mares ..... *mare*

camins ..... *camí*

excapellans ..... *excapellà*

- Formes nominals invariables

Els assignem com a lema la mateixa forma:

llapis ..... *llapis*

cactus ..... *cactus*

millor (adj.) ..... *millor*

molls ..... *molls*

- Formes nominals a les quals s'ha incorporat un sufix aspectiu

Es lematitzaran a partir del nom o adjectiu del qual provinguin, seguint els criteris indicats en aquest mateix punt:

fillets ..... *fill*

caseta ..... *casa*

pobretes ..... *pobre*

pobríssim ..... *pobre*

caminet ..... *camí*

milloret ..... *millor*

llapisset ..... *llapis*

Es dóna el cas que algunes d'aquestes formes nominals amb variació aspectuals s'han lexicalitzat (*caseta*). Les formes lexicalitzades tindran com a lema la mateixa forma:

caseta (entrada lexicalitzada) ..... *caseta*

3. NOMS PROPIS MARCATS EN EL PREPROCÉS COM A <NAME>, N4

- Noms propis formats per un sol mot.

Els assignem com a lema la mateixa forma que presenten:

Codi ..... *Codi*

Code ..... *Code*

- Grups polilexemàtics marcats en el preprocés com a noms propis (N4)

Els assignem com a lema la forma que presenten:

Boletín oficial del estado ..... *Boletín oficial del estado*

Filles de la Caritat ..... *Filles de la Caritat*

Manchester United ..... *Manchester United*

Futbol Club Barcelona ..... *Futbol Club Barcelona*

Registre Mercantil ..... *Registre Mercantil*

4. FORMES VERBALS

- Els assignem com a lema el verb en infinitiu:

canta ..... *cantar*

menjat ..... *menjar*

patiria ..... *patir*

patint ..... *partir*

En el corpus de l'IULA, per tal de facilitar la desambiguació dels documents, hem creat l'etiqueta H que representa les formes que són considerades tant adjectiu com participi. El lema que assignem a aquestes peces lèxiques un cop col·lisionades és la forma del verb en infinitiu:

tret ..... *treure*

- Verbs que presenten variacions formals: (aquest criteri s'aplica també als verbs que deriven dels que llistem en aquest apartat).

- De les dues formes d'infinitiu possibles, escollim com a lema la que ha estat prioritzada pel DIEC:

cabre/caber ..... *cabre*

caldre/caler ..... *caldre*

doldre/doler ..... *doldre*

jaure/jeure ..... *jeure*

nàixer/néixer ..... *néixer*

tindre/tenir ..... *tenir*

trac ..... *treure*

traiem ..... *treure*

traure/treure ..... *treure*

trec ..... *treure*

vindre/venir ..... *venir*

- Quan el DIEC no prioritza cap de les dues formes, aleshores escollim com a lema la més estesa:

ésser/ser.....*ser*

5. ABREVIACIONS: (Abreviatures, sigles i símbols)

Hem observat que en diverses ocasions una mateixa abreviació apareix amb més d'una forma. En aquests casos entrarem totes les formes en un fitxer en què s'assignarà un mateix lema a totes les variants.

- Abreviatures

Se'ls assigna com a lema aquell que correspon a la forma que abreugen (en funció dels criteris aplicats a 1, 2 o 3):[3]

veg......*veure*

v. .....*veure*

etc. .....*etcètera*

- Sigles

Els assignem com a lema la forma en què apareixen. Si trobem més d'una variant per a cada sigla se'ls assignarà com a lema la forma sense punts ni espais:

BOE .....*BOE*

EGB .....*EGB*

E.G.B. .....*EGB*

- Símbols

Els donem com a lema *símbol*, ja que en certes ocasions pot ser molt difícil saber a què remeten:

N .....*símbol*

^{0}C .....*símbol*

6. XIFRES

Els assignem com a lema *num* que és la forma amb què queden marcades després del preprocés:

11 .....*num*

7. DATES

Els donem com a lema *date* que és la forma que els assigna el preprocés:

11 de setembre.....*date*

8. IDENTIFICADORS

En el corpus de l'IULA, per tal de facilitar el tractament lingüísitic de les dades, hem creat l'etiqueta *identificador, B,* com a categoria major per tal de donar compte d'estructures del tipus *38), a)*...

A aquestes formes els donarem el lema *label* que és la marca que es genera en el preprocés:

38).....*label*

[3]Per poder lematitzar aquestes formes caldrà mantenir un fitxer d'abreviatures.

a) ..... *label*

9. SEQÜÈNCIES NO ANALITZABLES

En el marc del projecte s'han marcat algunes seqüències no lingüístiques com a no analitzables per tal de facilitar el tractament lingüístic dels documents: fórmules, símbols químics complexos, ...

- A aquestes formes els assignem el lema *noanat*:

V=4+3+2 ..... *noanat*

10. PARAULES DESCONEGUDES

És necessari assignar un lema a les paraules que no han estat reconegudes per l'analitzador morfològic per tal que es pugui passar el desambiguador estadístic.

El lema que es proposa de moment és *?*. Posteriorment en l'entrada de la neologia als diccionaris de l'IULA ja se'ls assignarà el lema que els correspongui:

referendament ..... *?*

conformemement ..... *?*

8.8 Annex 8: Procediment per a l'adquisició de textos amb l'escàner i posterior etiquetatge estructural

L'objectiu d'aquest document és especificar el procediment a seguir per adquirir textos a través de l'escàner i per a la incorporació de les marques que defineixen l'estructura del document.

El procediment que es detalla a continuació és aplicable al cas més complex, és a dir, a l'adquisició d'un document mitjançant l'escàner. Per als textos obtinguts en format electrònic s'han d'obviar aquelles fases que no tinguin sentit en aquesta situació (apartats 1 i 2). En qualsevol cas caldrà completar l'imprès de l'Annex III.

8.8.1 Selecció de les mostres que s'introduiran al Corpus Textual (CT)

8.8.1.1 Determinació de l'estructura del document

El text escrit pot adquirir diferents organitzacions estructurals, algunes molt simples (seqüències d'un o més paràgrafs separats per títols) i altres poden ser jerarquies complexes de capítols, seccions, subseccions, etc. Aquestes organitzacions complexes poden variar en el nom que es dóna a cada una de les parts, segons si es tracta de llibres, textos legals, diaris, etc. Per evitar aquesta complexitat, la normativa EAGLES assigna noms genèrics a les divisions de text en blocs, des de *div1* (nivell més alt) a *div8* (nivell més baix), independentment dels noms que l'editor pot haver assignat a aquestes divisions.

Les mostres que es prenguin d'un document seran sempre porcions de la divisió més alta (*div1*). També s'haurà de decidir fins a quina profunditat s'incorporen les marques estructurals. De totes maneres, en la majoria de textos i per tal de simplificar el procés, només es marcarà una divisió.

8.8.1.2 Selecció de les porcions del document a reconèixer

Quan es tracta de textos obtinguts a través de l'escàner, el criteri de base és recollir 10 mostres (no consecutives) de 3000 a 5000 paraules cadascuna. El més comú serà escollir una mostra de cada capítol del document, començant des del principi de capítol.

El criteri per determinar aproximadament quantes pàgines s'han d'explorar és el següent:

explorar el contingut d'una pàgina representativa,

reconèixer-la utilitzant l'OCR (omnipage),

obrir-la en el processador de textos Word,

comptar el nombre de paraules de la pàgina escollida amb l'opció Herram > Contar palabras,

el resultat de dividir 3000 pel nombre de paraules obtingut en el pas anterior serà el nombre de pàgines a explorar (arrodonir tenint en compte el número sencer superior).

Els documents disponibles en més d'una llengua tindran prioritat sobre els estrictament monolingües.

Si es disposa del document en diversos idiomes s'exploraran per separat les parts en cada llengua i es guardaran en fitxers separats, assegurant que hi hagi correspondència exacta en el text de cada un d'ells. Així, a un mateix document que estigui en dos idiomes li correspondran dos fitxers, un per a cada un dels dos idiomes.

Les mostres començaran sempre al principi d'un capítol i acabaran sempre amb un paràgraf complet (punt i a part), i si és possible coincidiran amb el final d'una secció.

8.8.1.3 Còpia escrita del material que s'incorpora al Corpus Textual (CT)

Cal fotocopiar les mostres escollides del document juntament amb les pàgines que continguin les dades d'identificació del document i l'índex, si n'hi ha. Cal alinear el document perquè les fotocòpies no surtin tortes, ja que acostumen a ser necessàries per al procés d'OCR.

8.8.2 Operacions preliminars

Abans de començar a escanejar, corregir i marcar un document, s'han de completar els següents documents amb les dades que es demanen:

8.8.2.1 Completar el "document resum"

Aquest document és el que es mostra a l'annex III. És molt important fer-hi constar la data d'inici i final del marcatge estructural, per tal de poder calcular el ritme en què avança la constitució del corpus. Aquest document s'ha d'arxivar a l'arxiu que correspongui, juntament amb les fotocòpies del document original (en cas d'haver-n'hi).

8.8.2.2 Completar el full de la base de dades Access

Cal informar a la coordinació del corpus del número de document que es treballarà.

8.8.3 Procés des del programa Omnipage (OCR)

8.8.3.1 Operacions preliminars

Abans d'escanejar un document, s'han de comprovar les dades de les mostres seleccionades a Excel (en el cas de dret, economia, informàtica, medicina) o a Access (en el cas de medi ambient): número de document, nom, llengua i si existeixen documents paral·lels. Si n'hi ha, cal assegurar-se que les mostres siguin les mateixes per a totes les llengües.

8.8.4 Procés des del programa Caere Omnipage Pro 10.0

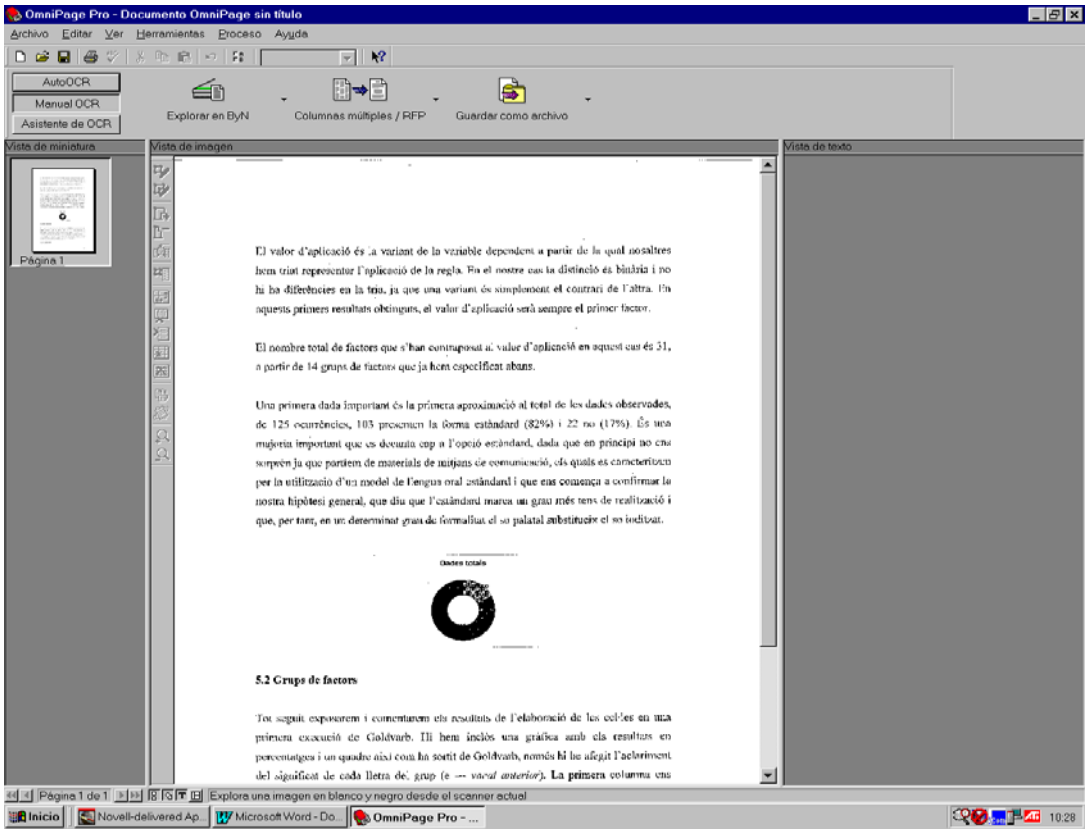
8.8.4.1 Operacions preliminars

Abans d'escanejar un document, se n'han de comprovar les dades amb la coordinació del corpus: número de document, nom, llengua i si existeixen documents paral·lels. Si n'hi ha, cal assegurar-se que les mostres siguin les mateixes per a totes les llengües.

Aquest procés es realitza a l'ordinador personal que té connectat l'escàner (RC018). Per fer-ho, després de connectar l'equip (ordinador i escàner), s'ha d'arrencar el programa *Caere Omnipage Pro 10.0* fent un doble clic a la icona del programa o des de inicio > aplicaciones > Caere OmniPage Pro 10.0. Des d'aquest programa es realitza l'escaneig dels fragments del document escollits anteriorment. Per dur a terme aquest procés, serà necessari:

En l'opció menú "Ver > Caja de herramientas de OmniPage", seleccionar l'opció Manual OCR si encara no ho està. A la barra d'eines apareixeran tres icones que ens serviran per escanejar els documents. Per escanejar un document cal seguir 3 passos que s'especifiquen a continuació:

Explorar en ByN: reconeix el text des de l'escàner i el porta a la pantalla. A l'esquerra tenim la Vista de Miniatura on apareix el full escanejat en petit, que seleccionarem per a què aparegui a la finestra central anomenada Vista de



imagen, on definirem les àrees de text que ens interessen. Si la pàgina no sembla ben escanejada cal fer edicion > eliminar la página actual.

El procés de selecció del fragments de text es farà en el mateix ordre en què estan sobre el paper. Per seleccionar-los s'utilitza el botonet de dalt de tot de la barra d'eines de Vista de imagen que diu dibujar zonas rectangulares. A la mateixa barra hi ha altres botonets d'ajut com els quatre de sota que serveixen respectivament per omplir o reduir la vista, per redreçar i per girar el full. S'evitarà incloure informació no rellevant: capçaleres, números de pàgina, notes a peu de pàgina amb referències bibliogràfiques, epígrafs i taules numèriques (encara que s'han de conservar els títols que porten aquestes figures o taules). Quan hi ha algun problema amb les caixes de selecció de text cal fer edición > borrar zonas.

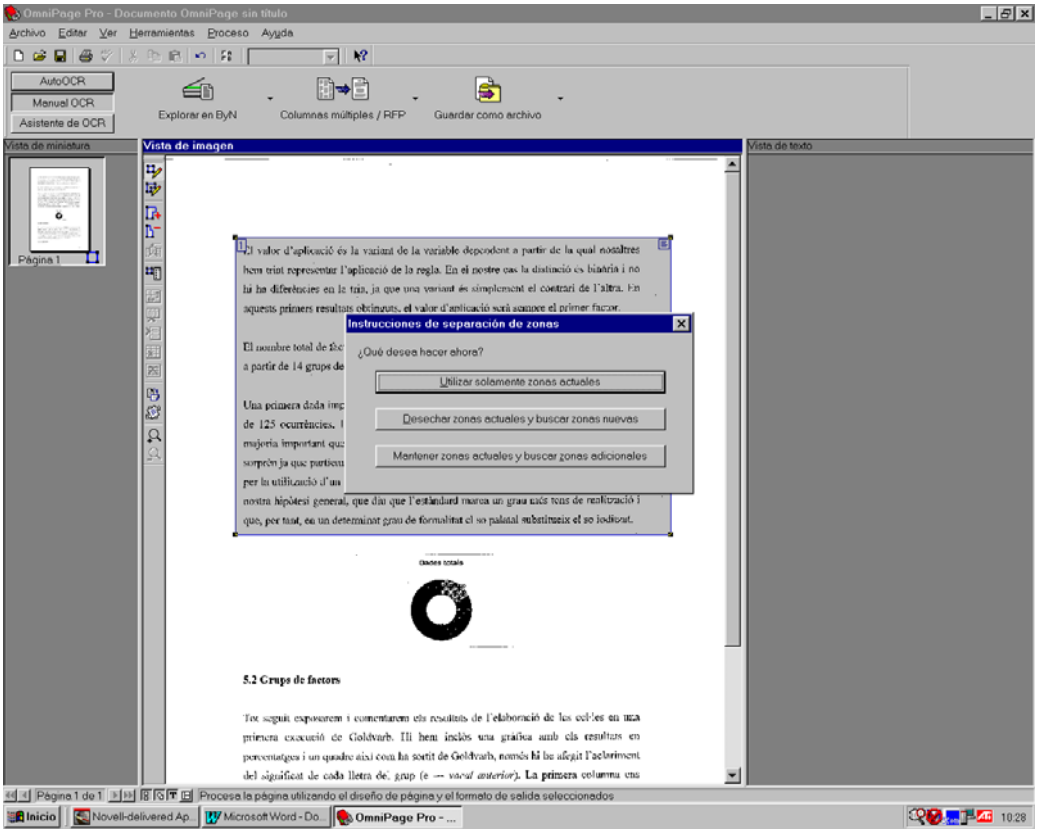
I. Columnas múltiples / RF: un cop seleccionat el text que ens interessa fem clic en aquest botonet que ens permet reconèixer les zones seleccionades. Un quadre de diàleg ens permet seleccionar:

   i. Utilizar solamente zonas actuales,

   ii. Desechar zonas actuales y buscar zonas nuevas o

   iii. Mantener zonas actuales y buscar zonas adicionales

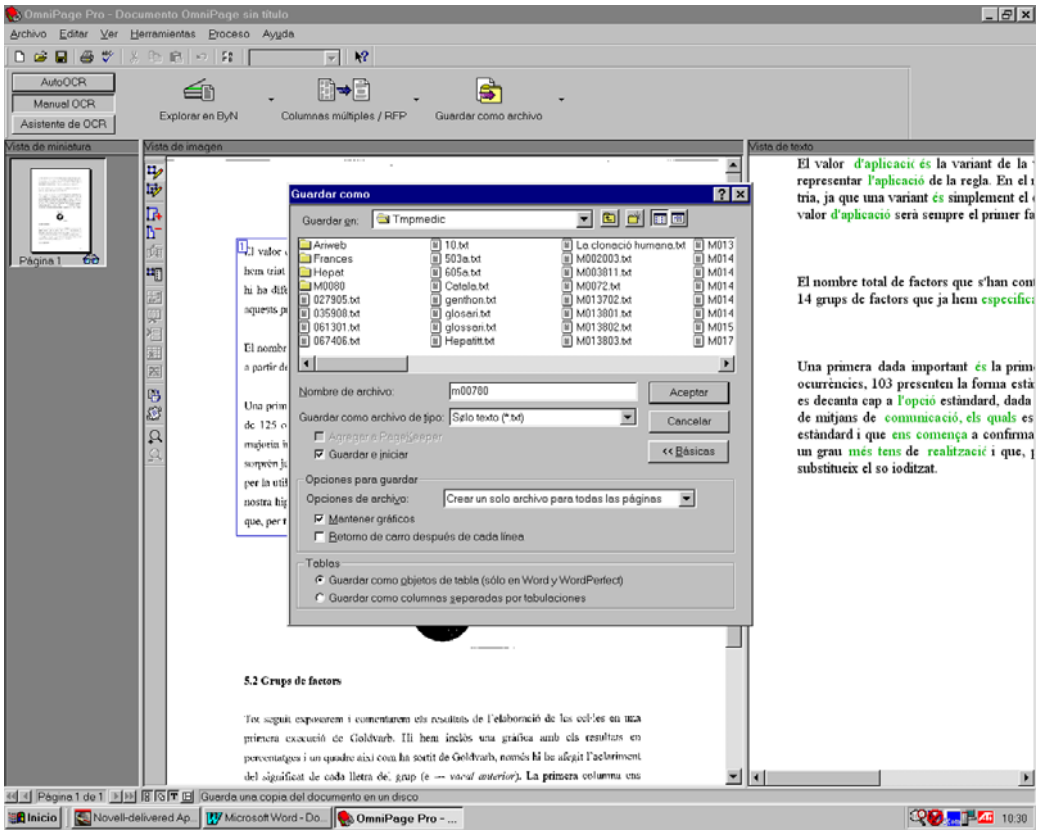
Normalment, si la nostra selecció ha estat correcta escollirem la opció **i.**



Finalitzat el procés de reconeixement ens surt el text a mà dreta en Vista de texto. Per visualitzar bé el resultat es pot moure la pantalla i ampliar l'estat de la

mostra. Els errors que presenti el text com a conseqüència d'aquesta fase de reconeixement, no s'han de corregir des del programa d'OCR. Quan l'estat del text sigui molt dolent es plantejarà de repetir les fotocòpies, reescanejar des de l'original o bé deixar de banda el document per deficient.

II. Guardar como archivo: Quan el resultat ja ens sembla bé el guardarem a Corpus en 'Elsinore\Vol3' (k:)



Cada mostra s'ha de guardar en un fitxer diferent, en el directori que correspongui:

K:\tmpdret si és de dret,

K:\tmpecon si és d'economia,

K:\tmpinfo si és d'informàtica,

K:\tmpma si és de medi ambient,

K:\tmpmedic si és de medicina.

El nom que cal assignar a cada mostra ha de ser amb la clau de l'àrea (d, e, i, ma, m), el nombre del document i el nombre de la mostra; per exemple:

m000101.txt

a b c

a) àrea *medicina*

b) document *0001*

c) mostra *01*

Cal especificar que volem guardar el documento com Sólo texto i en Opciones para guardar, Opciones de archivo seleccionarem crear un sólo archivo para todas las páginas. Ens sortirà una finestra al Bloc de Notas amb el text, que tancarem cada vegada.

Per a cada pàgina comprovarem que les opcions anteriors continuen activades. Cada vegada que guardem ens sortirà una pantalla dient "¿Desea reemplazar el archivo existente?", posarem que sí. Quan acabem d'escanejar la mostra: archivo>cerrar documento.

8.8.5 Procés en Word

8.8.5.1 Operacions preliminars

Quan ja disposem dels fitxers ANSI produïts pel programa d'OCR, o bé de fitxers obtinguts directament en versió electrònica, cal seguir les següents passes per a cada una de les mostres:

Obrir el fitxer de la mostra, escollint l'opció: Solo texto

Reconstruir el format del text que s'ha perdut en el procés d'OCR (línies en blanc, paraules tallades, etc.).

Algunes indicacions que faran més àgil aquest procés són les següents:

anar a l'inici del doc.	=	control	+	inicio
anar al final del doc.	=	control	+	final
buscar i canviar	=	control	+	L
buscar	=	control	+	B
copiar	=	control	+	C
desfer	=	control	+	Z
enganxar	=	control	+	V
refer	=	control	+	Y
seleccionar un fragment	=	shift	+	fletxes
seleccionar-ho tot	=	control	+	E
tallar	=	control	+	X

8.8.5.2 Correccions ortogràfiques derivades del procés d'escaneig i reconeixement

Per seleccionar l'idioma del corrector, cal seguir les següents passes des de la barra d'eines de Word:

Edición > Seleccionar todo (el document quedarà marcat com un únic bloc)

Herramientas > Idioma > seleccionar la llengua del document

Herramientas > Ortografía (comanda que activa el corrector ortogràfic de Word)

Els errors ortogràfics del document original s'han de corregir en aquesta fase.

Durant aquest procés, caldrà marcar les paraules escrites en un idioma diferent a les fotocòpies (el corrector les detectarà), per tal de recuperar-les en la fase posterior de marcatge i assignar-los la marca de mot estranger (veg. apartat 6.9).

Precaució! Les unitats manllevades a altres llengües, com ara els castellanismes que apareixen en documents catalans, no es corregiran.

8.8.5.3 Tractament de les referències bibliogràfiques

Les referències bibliogràfiques que apareguin en el text, a les notes a peu de pàgina, etc., s'han d'eliminar sempre que no formin part de l'estructura lingüística on apareixen (exemple A). En cas que constitueixin un element estructural de la construcció lingüística, s'han de mantenir (exemple B).

A) El text següent:

> Això ha fet que la racionalització de la despesa energètica a l'agricultura passi, cada vegada més, per l'estudi dels consums energètics al llarg dels diversos cultius. Així, adquireix sentit la proliferació de treballs d'aquesta mena publicats en els darrers anys (Pimentel, D. and M. Pimentel, 1979; Gámir, A., 1980; Pimentel, D., 1980).

quedarà tal com segueix:

> Això ha fet que la racionalització de la despesa energètica a l'agricultura passi, cada vegada més, per l'estudi dels consums energètics al llarg dels diversos cultius. Així, adquireix sentit la proliferació de treballs d'aquesta mena publicats en els darrers anys.

B) En canvi, en el següent fragment es mantindrà la referència bibliogràfica:

> És realment sorprenent comprovar que la possibilitat que ofereix el Decret 2822/1974 quan estableix el Pla Comptable per a les Petites i Mitjanes Empreses per facilitar que s'acullin als beneficis de la Regularització de Balanços no ha estat aprofitada per la majoria dels nostres empresaris.

Un cop efectuada la correcció del text, cal guardar el document novament en format Solo texto con saltos de línea.

8.8.6 Marcatge

L'objectiu d'aquesta fase és afegir la informació bàsica d'estructura del document, així com la relativa a característiques tipogràfiques o a fragments en un idioma diferent al del document.

Si és necessari afegir comentaris o dades específiques a una mostra, cal seguir la normativa SGML que inclou els comentaris començant per <!-- i acabant amb -->.

Ex:

```
<!--    Fórmula de promulgació de la Unió Europea
        origen: Tractats contitutius de les Comunitats Europeas    -->
```

8.8.6.1 Operacions preliminars

Des de Word, seleccionar la plantilla de corpus.dot per a les macros de tractament de corpus des d' Herramientas > Plantillas y complementos > Corpus.dot. En cas que aquesta plantilla no aparegui seleccionada, cal fer agregar i seleccionar-la a k:\ i dir acceptar.

8.8.6.2 Recompte de les paraules de la mostra

Per comptar el nombre de paraules de la mostra s'ha d'escollir l'opció "Herram > Contar palabras". Un cop obtingut el resultat s'anotarà en la pàgina on es recullen les dades per a cada document del CT (Document resum).

8.8.6.3 Aplicació de la macro Postcorrecció

No cal seleccionar gaire text.

La macro Postcorrecció converteix els punts volats de les eles geminades (l·l), per al català a codis SGML (l·l). Per aplicar-la no cal seleccionar gaire text.

8.8.6.4 Tractament de les divisions

<div1 n= >, </div1>; cal seleccionar tot el text.

Tot text escrit es composa bàsicament d'una seqüència de frases, les quals s'agrupen en paràgrafs, els quals, al seu torn, s'agrupen en nivells superiors. Aquests nivells poden adquirir graus molt variats de complexitat: capítols, seccions, subseccions, etc. Aquests són els nivells que es volen indicar en aquest apartat. La macro que realitza aquest marcatge (divisió) necessita una sèrie d'informacions addicionals. A continuació s'indiquen quines són aquestes dades i el seu significat.

- *Nivell de Bloc:*

Nivell dins de la jerarquia; va des del nivell més alt div1, que correspon habitualment als capítols d'un llibre o document, fins al nivell més baix: div 7, per les (sub)(...)seccions.

Tipus de bloc:

Nom formal que s'assigna a la divisió que s'està marcant. Està pensat únicament per als textos legals (lleis i reglaments).

Número de bloc:

Número d'ordre corresponent al tipus de bloc (textos legals) o bé al nivell de bloc (la resta de textos). Aquest número s'indicarà en el format que apareix en el text original (xifres aràbigues, números romans, lletres, etc.)

Bloc:

Indicació sobre si la divisió que s'està marcant ha estat recollida en la seva totalitat o no (cas dels textos recuperats a través de l'escàner). La consideració de divisió completa/incompleta s'aplica al document de manera global i no en una mostra en concret.

Important: Tots els documents de les àrees d' economia, informàtica, medi ambient i medicina, es redueixen en un únic nivell, que s'aplica als blocs principals del document; és a dir, si un document està dividit en capítols i aquests en seccions, només es marcaran els capítols com a div1 (les seccions no portaran cap marca específica). En tots els casos sempre cal indicar si la divisió està o no completa. El marcatge manual de les divisions fins a un nivell jeràrquic 8 es manté només per als documents de dret.

En els documents que no són de dret tampoc no es consignarà el tipus de divisió; per defecte la gramàtica SGML interpretarà que es tracta de capítols. En canvi, sí s'ha d'indicar.

8.8.6.4.1 Exemples només per als documents de dret:

A) El text següent:

.....
8.12 Compra-venda d'una finca rústica arrendada

que correspon al Capítol 8-Sección 12, quedarà marcat de la següent manera:

.....
<div2 type=secc n=12>
<head type=main>Compra-venda d'una finca rústica arrendada </head>
.....
</div2>

B) El text següent:

Article 5

Tot el que no preveu aquest Conveni es regirà per les normes de l'Ordenança de treball en el comerç, aprovada per l'Ordre de 24 de juliol de 1971 (BOE de 14 d'agost) i rectificada per l'Ordre de 4 de juny de 1975 (BOE de 14 del mateix mes), com també per les lleis, pels reglaments i per altres disposicions de compliment obligatori.

quedarà marcat de la manera següent:

<div2 type=art n=5>
<s>Tot el que no preveu aquest Conveni es regirà per les normes de l'Ordenança de treball en el comerç, aprovada per l'Ordre de 24 de juliol de 1971 (BOE de 14 d'agost) i rectificada per l'Ordre de 4 de juny de 1975 (BOE de 14 del mateix mes), com també per les lleis, pels reglaments i per altres disposicions de compliment obligatori.</s>

C) El següent fragment de la Constitució:

CAPÍTULO PRIMERO

De los españoles y los extranjeros

12. Los españoles son mayores de edad a los dieciocho años.

quedarà marcat de la manera següent:

<div5 n=PRIMERO type=cap>
<head type=main> CAPÍTULO PRIMERO </head>
<head type=sub r=it> De los españoles y los extranjeros</head>
<div7 n=12 type=art> Los españoles son mayores de edad a los dieciocho años.
</div7>

Convé remarcar el fet que, a diferència de l'exemple B, en aquest darrer cas s'ha marcat el número del capítol com a títol principal. S'ha fet així considerant que l'autor del document volia posar èmfasi en la indicació del capítol. Fer-ho d'una manera o d'una altra queda a criteri del codificador en el cas dels documents recuperats a través de l'escàner. Habitualment, les divisions es codificaran seguint els esquemes d'A i B.

Precaució! Quan el capítol d'un llibre té associat un o més noms d'autor, tots aquests noms s'han d'eliminar. El següent fragment del capítol 9 d'un llibre (recuperat a través de l'escàner)

Cap IX

Gestió documental amb macroordinadors: característiques, estructura i tecnologia dels Sistemes de Gestió Documental. Introducciò.

Lluís Codina Bonilla

Universitat Pompeu Fabra

Ernest Abadal Falgueras

Escola Universitària "Jordi Rubió i Balaguer" de Biblioteconomia i Documentació

ha de quedar marcat com s'indica a continuació:

<div1>
<head type=main rend=bo>Gestió documental amb microordinadors: característiques, estructura i tecnologia dels Sistemes de Gestió Documental</head><head type=sub rend=bo>Introducció</head>
<p><s> ...
</div1>

8.8.6.5 Tractament dels títols

<head>, </head>

Una divisió o bloc de text pot contenir un títol o encapçalament previ al text. Aquest fragment de text s'ha de marcar seleccionant com a bloc el títol i activant la macro Títol.

La informació que cal incorporar és la següent:

a) tipus de títol:

La norma EAGLES preveu les següents possibilitats:

- *principal:* títol principal de la divisió <head type=main>, </head>
- *secundari*: títol/s secundari/s de la divisió <head type=sub>, </head>
- *no especificat:* tipus de títol no especificat o desconegut

<head type=unspec>, </head>

Anteriorment s'indicava també en els títols amb la informació relativa a l'autor i la procedència d'un article. Actualment aquesta informació ha deixat de marcar-se; en conseqüència, caldrà eliminar aquests títols dels documents en què apareguin.

b) característica tipogràfica del títol:

L'especificació d'aquesta informació acaba el procés.

Precaució! Aquesta informació només es tindrà en compte quan es marquen els títols manualment, és a dir per als documents recuperats mitjançant l'escàner. En cas que aquesta informació ja vingui consignada des del document en format electrònic, no s'ha de retocar.

El tractament que es dóna a les divisions no impedeix que els títols que puguin aparèixer en divisions diferents de les de primer nivell (div1) no s'hagin de marcar. La gramàtica que es fa servir en aquestes àrees ja preveu aquest tipus de situacions.

Sobre l'aplicació de l'atribut de títol principal i títol secundari, cal tenir en compte les següents consideracions:

1. Títol format per dues parts separades únicament per un punt

Es considerarà cadascuna de les parts com a títols independents.

L'exemple següent:

Recursos administrativos. Principios generales
Introducción

quedarà marcat d'aquesta manera:

<head type=main>Recursos Administrativos</head>
<head>Principios generales </head>
<head type=sub>Introducción</head>

Cal suprimir els signes de puntuació que apareixen en el títol. Això ha de ser així per a qualsevol títol, per tal d'evitar problemes posteriors en la fase d'introducció de les marques de frase i paràgraf (veg. secció 8.8.8).

2. Títol format per dues parts separades per un canvi de línia

Es considerarà la línia superior com a títol principal i la línia inferior com a títol secundari. L'exemple següent:

Recursos administrativos
Principios generales

quedarà marcat d'aquesta manera:

<head type=main>Recursos Administrativos.</head>
<head type=sub>Principios generales </head>

3. Títol que indica explícitament el començament d'una divisió

El text següent:

CAPÍTOL DOS
Formes d'adjudicació del contracte d'obres

quedarà marcat tal com es mostra:

<head type=main rend=bo>CAPÍTOL DOS</head>
<head type=sub>Formes d'adjudicació del contracte d'obres </head>

4. Títol que incorpora una nota de qualsevol tipus

La nota ha de formar part del títol. L'exemple següent:

Disposicions generals[2]
.....
.....
2. Vid els articles 1 a 18 del Reglament.

quedarà marcat com es mostra (veg. tractament de les notes, al punt 8.8.6.11):

<head type=sub r=il>Disposicions generals<ptr target='d23n2'></head>
.....
.....
<note place=foot n=2 resp=ed id='d23n2'> <s>Vid els articles 1 a 18 del Reglament. </note>

5. Títol que incorpora paraules en un altre idioma

L'exemple:

Procedimiento de *habeas corpus*

quedarà marcat de la manera següent:

<head type=main r=bo>Procedimiento de<foreign lang=LA rend=il>habeas corpus</foreign></head>

6. Títol principal compost per dues parts en una mateixa línia.

L'exemple:

Recursos administrativos. Principios generales

quedarà marcat:

<head type=main>Recursos administrativos</head>
<head> Principios generales </head>

8.8.6.6 Tractament de la informació tipogràfica

<hi rend=" ">, </hi>

Si hi ha una o més paraules del text amb característiques tipogràfiques especials (negreta, cursiva o subratllat) s'inseriran la/les marca/ques corresponents. Per fer-ho, se seleccionen la/les paraula/es com a bloc i s'activa la macro tipografia que demana la/les característica/ques tipogràfiques que cal marcar. De tal manera que la negreta quedarà marcada com a <hi rend="bo">, la cursiva com a <hi red="il"> i el subratllat com a <hi rend="ul">.

Precaució! Si les característiques tipogràfiques es corresponen amb un títol o subtítol (veg. 8.8.6.5), paraules en un altre idioma (veg. 8.8.6.9) o, en general, a blocs sencers (veg. 8.8.7.2) no s'han de marcar seguint aquest procediment. En aquests casos, les marques s'incorporaran seguint el procés corresponent a cada un d'aquests casos.

Cal tenir en compte que un atribut tipogràfic es pot estendre més enllà de la jerarquia que el conté; per exemple, una frase. La macro preveu inserir l'etiqueta en tres situacions diferents: una seqüència de paraules (dins d'una mateixa frase), un paràgraf sencer o bé una frase.

8.8.6.7 Tractament de les taules

<table n=' '><head type=main>, </head></table>

Les taules estan formades normalment per un títol i un conjunt de cel·les. Aquestes cel·les contenen informació que es pot dividir en:

- text lingüísticament rellevant,
- text format per números, noms propis, sigles, paraules en una llengua diferent a la del document, etc.

Només es codificaran el títol[1] i les cel·les que contenen text lingüísticament rellevant. Per tant, les cel·les que només contenen números, noms propis o sigles, entre altres, s'han d'eliminar.

Per exemple, quan la taula conté informació del següent tipus:

Taula 24. Pressupost de despeses de la Generalitat de Catalunya

Secció	Departaments	MPTA	%
2	Presidència, la	38.283	2,6
4	Governació	33.569	2,3
5	Economia i Finances	. . .	. . .
6	Ensenyament	. . .	. . .
7	Cultura	. . .	. . .
. . .	. . .	. . .	. . .
Total		1.462.178	100.0

la codificació es limitarà al títol; o sigui:

> <table n=24><head type=main>Pressupost de despeses de la Generalitat de Catalunya</head></table>

En canvi, la informació de la taula següent s'ha de mantenir tal com es mostra, posant els marques <row>, </row>, <cell> i </cell> manualment[2]:

[1] Si el títol té característiques tipogràfiques, cal introduir aquestes marques manualment perquè la macro no dóna aquesta opció. D'aquesta manera, si el títol està en negretes cal escriure el títol entre les marques <head rend="bo">Títol</head>, en cursiva <head rend="il">Títol</head> o subratllat <head rend="ul">Títol</head>.

[2] En casos excepcionals, per tal de facilitar el marcatge, es pot contemplar la possibilitat d'extreure els elements textuals del cos de la taula. És a dir, es podria marcar el títol com a text de la taula i el cos com a text normal.

Tabla 5-0. Estadística de gestión de mensajes

Estadística	Descripción
Hora de la última reinicialización	Fecha y hora de la última reinicialización de los datos estadísticos actualmente en pantalla.
Total	Número total de mensajes que entregó el servidor de gestión de mensajes del MHS.
Tamaño total	Tamaño total de los mensajes que entregó el servidor.

<tablen=5-0><head type=main> Estadística de gestión de mensajes </head>
<row><cell> Estadística </cell><cell> Descripción</cell></row>
<row><cell> Hora de la última reinicialización </cell><cell> Fecha y hora de la última reinicialización de los datos estadísticos actualmente en pantalla.</cell></row>
<row><cell></cell><cell> Número total de mensajes que entregó el servidor de gestión de mensajes del MHS</cell></row>
<row><cell></cell><cell> Tamaño total de los mensajes que entregó el servidor </cell></row>
</table>

Un altre cas, es presenta quan trobem una taula amb informació mixta (fórmules i text):

tabla 6.1 Nomenclatura estándar de los cariotipos cromosómicos

Cariotipo	Descripción
46,XY	Constitución cromosómica de hombre normal
47,XX, + 21	Mujer con trisomía 21, síndrome de Down
47,XY, + 21 46, XY	Mosaico de células con trisomía 21 y de células normales en hombe
46,XY, del(4)(p14)	Hombre con delación distal del brazo corto de la banda 14 del cromosoma 4
46,XX,dup(5p)	Mujer con duplicación del brazo corto del cromosoma 5
45,XY, -13, -14, t(13Q;14q)	Hombre con una translocación robertsoniana equilibrada de cromosomas 13 y 14
46,XY, t(11;22)(q23;q22)	El cariotipo muestra la pérdida de un 13 normal y un 14 normal Hombre con una translocación recíproca equilibrada entre los cromosomas 11 y 22 Los puntos de rotura se localizan en 11q23 y 22q22

S'ha de marcar de la següent manera:

```
<table n='6-1'><head type=main> Nomenclatura estándar de los cariotipos cromosómicos
</head>
<row><cell>Cariotipo</cell><cell>Descripción</cell></row>
<row><cell>Constitución cromosómica de hombre normal</cell></row>
<row><cell>Mujer con trisomía 21, síndrome de Down</cell></row>
<row><cell>Mosaico de células con trisomía 21 y de células normales en hombre</cell></row>
<row><cell>Hombre con delación distal del brazo corto de la banda 14 del cromosoma 4</cell></row>
<row><cell>Mujer con duplicación del brazo corto del cromosoma 5</cell></row>
<row><cell>Hombre con una translocación robertsoniana equilibrada de cromosomas 13 y 14</cell></row>
<row><cell>El cariotipo muestra la pérdida de un 13 normal y un 14 normal</cell></row>
<row><cell>Hombre con una translocación recíproca equilibrada entre los cromosomas 11 y 22</cell></row>
<row><cell>Los puntos de rotura se localizan en 11q23 y 22q22</cell></row>
</table>
```

8.8.6.8 Tractament de les figures

<figure n=''><head type=main>, </head></figure>

Existeixen alguns documents en els quals s'inclouen figures o gràfics per il·lustrar alguns aspectes del text. Tot i que el contingut textual d'aquests tipus d'elements, en principi, no és interessant, sembla oportú de reflectir-ne l'existència i aprofitar el text que forma el títol.[3]

La macro per marcar els títols de taules i figures és Taula / Figura, on se'ns demana el número, si és que en té.

Exemple:

> figure 1: Normas de resultados NOx (mg/m3) alcanzables mediante modificaciones de la combustión (Ver Repertorio Cronológico Legislación 1991, TOMO I, pg. 1718)

quedarà marcat com

> <figure n=1><head type=main>Normas de resultados NOx (mg/m3) alcanzables mediante modificaciones de la combustión (Ver Repertorio Cronológico Legislación 1991, TOMO I, pg. 1718) </head> </figure>

8.8.6.9 Seqüències en un altre idioma

<foreign lang= >, </foreign>

Mots en un altre idioma

[3] Si el títol té característiques tipogràfiques, cal introduir aquestes marques manualment perquè la macro no dóna aquesta opció. D'aquesta manera, si el títol està en negretes cal escriure el títol entre les marques <head rend="bo">Títol</head>, en cursiva <head rend="il">Títol</head> o subratllat <head rend="ul">Títol</head>.

S'entén per paraules en un altre idioma aquelles que estan en un idioma diferent al del document. No tractarem com a *foreings* els mots isolats, que es consideren manlleus. Així, si en realitzar la correcció ortogràfica es detecten paraules amb aquesta característica, caldrà afegir per a cada grup de paraules l'etiqueta corresponent utilitzant la macro foreign.

Per fer-ho, cal marcar com a bloc les paraules que s'han de marcar i activar la macro foreign, la qual obrirà una finestra com la que es mostra a continuació. En aquesta, caldrà seleccionar l'idioma, ja sigui algun dels que s'han previst, ja sigui mitjançant el codi ISO 639 corresponent:

Alemany	de
Anglès	en
Català	ca
Espanyol	es
Francès	fr
Italià	it
Llatí	la
Grec	gr
Holandès	nl
Japonès	jp

També cal anotar si la seqüència de paraules en un altre idioma té alguna característica tipogràfica especial: negreta, cursiva o subratllat. Seleccionant la característica que correspongui, es tanca la finestra i s'insereixen les marques corresponents en el text. Una expressió anglesa (com per exemple *key word*) en un text castellà haurà de quedar marcada de la manera següent:

.... <foreign lang=EN> key word </foreign> ....

Precaucions**!**

a) Les paraules que podrien ser tractades com a simples a nivell gràfic –és a dir, paraules morfològicament simples, derivades i compostes amb guió no s'han de considerar paraules estrangeres.

b) Tampoc s'han de marcar com a mots estrangers els noms propis, ja que això impediria que en la fase del preprocessament lingüístic se'ls reconegués com a tals. Concretament, el preprocessament reconeix com a noms propis les seqüències formades per un mot iniciat en majúscula el qual no estigui precedit per cap punt.

8.8.6.9.1 Frases en un altre idioma

La macro extr_tei preveu el cas, més habitual, que hi hagi una frase o un paràgraf sencer en un altre idioma. Per això només cal fer la indicació corresponent a la finestra. En cas que la porció de text en un altre idioma constitueixi un fragment que va més enllà d'on s'acaba la frase, aquesta informació ha de consignar-se en el bloc de text (frase, paràgraf, ...) corresponent.

8.8.6.10 Tractament de les llistes

<list>, </list>

Existeixen dos tipus de llistes, l'enumeració dins del text o bé la que queda fora d'aquest.

Dins d'una llista es defineixen les parts següents (encara que no totes hagin d'estar presents sempre):

cos: el conjunt de la llista,

títol: títol de la llista,

ítem: element de la llista, que pot contenir un identificador de cada element de la llista o etiqueta amb la indicació "n=", més el codi alfanumèric pertinent.

Per marcar una llista s'ha de:

seleccionar cadascun dels ítems i aplicar la macro **Unificar Paràgrafs** per evitar errors d'aplicació de la macro **Llista**;

seleccionar la llista sencera i activar la macro **Llista**, indicant l'opció que correspongui;[4]

vigilar amb les llistes incrustades i estar al cas de l'últim element de cada subllista.

Exemples:

A) El text següent:

Expedients de contractació
1. de tramitació ordinària
2. de tramitació urgent
3. de règim excepcional

s'ha d'etiquetar tal com es mostra:

```
<list><head>Expedients de contractació</head>
<item n=1>de tramitació ordinària</item>
<item n=2>de tramitació urgent</item>
<item n=3>de règim excepcional</item>
</list>
```

B) El text següent:

... els expedients de contractació podran ser: a) de tramitació ordinària b) de tramitació urgent c) de règim excepcional segons correspongui en cada cas.

s'ha d'etiquetar tal com es mostra:

```
<p><s>.... els expedients de contractació podran ser: <list><item n=a>de tramitació ordinària</item><item n=b>de tramitació urgent </item><item n=c>de règim excepcional </item> </list> segons correspongui en cada cas.</s></p>
```

[4] Quan es tracti d'una llista numerada, s'han de treure els punts, parèntesis i altres marques que acompanyin el número.

8.8.6.10.1 Tractament de les subllistes

Criterios de inclusión
1. Consentimiento del paciente a participar en el estudio
2. Pacientes de ambos sexos con edad superior a 18 años
3 Diagnóstico de infección ósea establecido por presencia de dos o más de los siguientes signos:
Fiebre
Signos inflamatorios locales
Fístula con o sin supuración y/o herida accidental o quirúrgica con exposición de hueso con o sin supuración
4. Presencia de alguno de los siguientes signos radiológicos:
Osteoporosis y lisis del hueso reticulada
Lucencia cortical y lisis Periostitis e involucro
Tumefacción de tejidos blandos
Radiolucencia única o múltiple
Secuestro
Migración de fragmentos corticales

Per marcar una subllista s'ha de seleccionar cadascun dels ítems

```
<list><head>Criterios de inclusión</head>
<item n='1'><s>Consentimiento del paciente a participar en el estudio </s></item>
<item n='2'><s>Pacientes de ambos sexos con edad superior a 18 años </s></item>
<item n='3'><s>Diagnóstico de infección ósea establecido por presencia de dos o más de los siguientes signos: </s>
<list>
<item> Fiebre </item>
<item> Signos inflamatorios locales </item>
<item> Fístula con o sin supuración y/o herida accidental o quirúrgica con exposición de hueso con o sin supuración </item>
</list></item>
<item n='4'><s>Presencia de alguno de los siguientes signos radiológicos: </s>
<list>
<item> Osteoporosis y lisis del hueso reticulada </item>
 <item> Lucencia cortical y lisis Periostitis e involucro </item>
 <item> Tumefacción de tejidos blandos </item>
 <item> Radiolucencia única o múltiple </item>
<item> Secuestro</item>
<item> Migración de fragmentos corticales </item>
</list></item>
</list>
```

8.8.6.11 Tractament de les notes

<note place=foot n= id=' '> <s>, </s> </note>

Si el text inclou notes d'algun tipus (peu de pàgina, final de document o intercalades en el text), aquestes s'han de codificar de la manera següent:

• en la posició del text on es fa referència a la nota s'afegeix una marca que indica l'existència de la nota 1 (<ptr target='d1n1'>).

• el cos de la nota (element note) es col·loca al final de la divisió que s'està marcant, però abans de l'inici de qualsevol altra divisió de nivell inferior.

Esquemàticament un text genèric com el següent:

<div3 n=24 type art>
<s> .....frase.....[2] </s>
..... continuació del text .....
2. ... text de la nota ...

es marca tal com s'indica a continuació:

<div3 n=24 type art>
<s>... frase ... <ptr target='**nota2**'></s>
..... continuació del text .....
<note n=2 place=foot id='**nota2**'> <s>... text de la nota ...</s>
</div3>

o bé :

<div3 n=24 type art>
<s>... frase ... <ptr target='**nota2**'></s>
..... continuació del text .....
<note n=2 place=foot id='**nota2**'> <s>... text de la nota ...</s>
<div4 n=1 type art>
....
</div4>
</div3>

Considerem que té una importància especial el valor de l'atribut target de l'element ptr i el valor de l'atribut id de l'element note. Els valors assignats a aquests atributs han de coincidir ja que és la manera que té l'estàndard SGML per establir una relació entre la referència a la nota (element ptr) i la seva aparició física en el document (element note). En el cas de l'exemple anterior, el valor compartit per tots dos elements és 'nota2'.

L'estàndard SGML preveu el valor d'aquests atributs en un codi que ha d'identificar unívocament la nota dins del document i que està format per una combinació de lletres (en minúscules) i números. En els textos del CT s'ha adoptat la convenció següent per a la identificació de les notes:

àrea - número_de_ document - n - número_de_nota

Així, la nota 24 del document número 0012 de l'àrea de dret s'identificarà com d12n24.

Per exemple, el text real següent del document d0005:

Las normas relativas a los derechos fundamentales y a las libertades que la Constitución reconoce se interpretarán de conformidad con la Declaración Universal de Derechos Humanos y los tratados y acuerdos internacionales sobre las mismas materias ratificados por España.[2]

. . . . . .

2. La libertad religiosa es un derecho fundamental puesto que queda recogida en el artículo 16 de la de la Constitución entre los acuerdos internacionales que han de tenerse en cuenta para su interpretación en primer lugar están los convenios Internacionales de Derechos Humanos (ver §§ 2 al 6), y también los convenios con confesiones que disfrutan de la consideración de Tratados Internacionales, como es el caso de los Acuerdos con la Santa Sede.

s'hauria d'etiquetar de la manera següent:

<s>Las normas relativas a los derechos fundamentales y a las libertades que la Constitución reconoce se interpretarán de conformidad con la Declaración Universal de Derechos Humanos y los tratados y acuerdos internacionales sobre las mismas materias ratificados por España.<ptr target='d5n2'></s>

. . .

<note place=foot n=2 id='d5n2' > <s> La libertad religiosa es un derecho fundamental puesto que queda recogida en el artículo 16 de la Constitución entre los acuerdos internacionales que han de tenerse en cuenta para su interpretación en primer lugar están los convenios Internacionales de Derechos Humanos (ver §§ 2 al 6), y también los convenios con confesiones que disfrutan de la consideración de Tratados Internacionales, como es el caso de los Acuerdos con la Santa Sede. </note>

Per realitzar aquesta codificació es farà el següent:

Seleccionar com a bloc el text de la nota.

Activar la macro Nota i completar la informació que correspon al número i tipus de nota.

Seleccionar l'origen de la nota (autor, traductor o editor).

El resultat del pas anterior serà la creació de dos elements (ptr i note) que hauran de moure's manualment a la posició que els correspongui: ptr, al punt del text on es fa referència a la nota i l'element note, a la posició definitiva dins del document SGML: final de la secció.

En l'aplicació de la macro Nota, el codi assignat als atributs target i id no inclou el número de document. Per exemple, la nota anterior tindria aquesta forma: id='dn2'. El número de document s'ha d'afegir manualment o bé utilitzant l'opció de buscar i substituir, quan ja s'han modificat totes les notes de la mostra.

Cal esmentar de manera especial les notes marcades amb caràcters no alfanumèrics, per ex. el "*". Aquest caràcter no està permès com a valor d'un atribut per l'estàndard SGML. Per aquesta raó, el fragment del codi que identifica el número de nota en els atributs target i id s'ha de substituir per un o més caràcters permesos, però cal tenir en compte que el codi resultant ha de continuar essent una identificació única per a la nota dins del document (p. ex. d123n0, d123na1

Una altra situació peculiar es presenta quan dins d'un mateix document apareixen més d'una nota amb el mateix número, quan per exemple, la numeració de les notes es canvia a 1 amb l'inici de cada capítol). En aquest cas, les notes s'identificaran de la manera següent:

àrea - número_de_ document - m - número_de_muestra - n - número_de_nota

Així, la nota 24 de la mostra 7 pertanyent al document número 0012 de l'àrea de dret s'identificarà com d12m7n24.

Precaució! En els acords i textos legals s'ha de considerar com autor de la nota l'editor.

En cas que una nota a peu de pàgina (o d'un altre tipus) no contingui text lingüísticament rellevant s'ha d'eliminar: Això significa que no s'hauran d'incloure els elements <note> i <ptr>.

Per exemple, les següents notes es poden eliminar:

[1] Sigerist, H.E. Introduction to historical approach to medicine. In: *A history of medicine*. New York: Oxford University Press, 1951 (reprint 1977); pàgs 3-4.

[2] López Piñero, J.M. Los estudios histórico-sociales sobre la medicina. Introducción. En: Lesky, E. (ed) *Medicina social. Estudios y testimonios históricos.* Madrid: Ministerio de Sanidad y Consumo, 1984; pàg 4.

[3] López Piñero, J.M. *op. cit.*, 1984; pàg. 22.

[4] Homer, La Ilíada, Himne XII, v. 147-152.

[5] V.arts. 94.1 i 106.1 LCX.

[6] Definit a la sec. 4.3.

[7] Com suposa Hart (1982).

I les següents, no, ja que contenen text d'interès per al CT:

[1] Prego als meus col·legues, els arxivers, que disculpin la simplificació amb la qual presento a continuació l'essència de la seva delicada, complexa i alhora difícil feina.

[2] Los protocolos del cliente NetWare soportan desde la Capa o nivel 3 (la capa o nivel de red) a la Capa o nivel 4 (la capa o nivel de transporte) del modelo de referencia de red de Interconexión de sistemas abiertos (OSI) de la Organización internacional para la estandarización (ISO).

[3] El desenvolupament del concepte de probabilitat es troba magníficament explicat al llibre de Ian Hacking, The Emergence of Probability, Cambridge University Press, 1976.

[4] Es posible que esta asignación ya esté configurada como una unidad de búsqueda en el guión de entrada para la estación cliente. Utilice la utilidad MAP a fin de visualizar una lista con las asignaciones de unidad existentes. Si una asignación de unidad ya existe en el directorio SYS:PUBLIC proceda al paso 2.

8.8.6.12 Fórmules matemàtiques

<na>, </na>

En certs documents de les àrees d'economia i d'informàtica apareixen fórmules matemàtiques de difícil marcatge i difícil processament lingüístic. Per tal de resoldre aquesta qüestió i considerant que la codificació de les fórmules és molt costosa, s'afegirà una marca **<na>** (no analitzable) a l'inici de la fórmula (o fragment de text assimilable a una fórmula) i **</na>** al final de la fórmula, mitjançant la macro No Analitzable. Aquest procés cal fer-lo també per al símbols "<" (més petit que) i ">" (més gran que).

Donat aquest fragment:

> D'altra banda, si recordem la definició de la taxa d'explotació o plus-vàlua:
> $e = s/v = [1 - (b_1 \lambda\delta_1+b_2\lambda_2)]/(b_1 \lambda\delta_1+b_2\lambda_2)$
> i, per tant,
>
> $(b_1 \lambda\delta_1+b_2\lambda_2)(1+\Xi) = 1$,

és a dir; $b_1 \lambda\delta_1+b_2\lambda_2$ és el temps necessari i $e(b_1 \lambda\delta_1+b_2\lambda_2)$ és el *temps d'excedent* per hora-home. ...

el marcatge que cal aplicar és el següent:

> <s>D'altra banda, si recordem la definició de la taxa d'explotació o plus-vàlua: <na>xxx</na> i, per tant, <na>xxxx</na>, és a dir; <na>xxx</na> és el temps necessari i <na>xxxx</na> és el <hi rend=il>temps d'excedent</hi> per hora-home.</s><s> ...

El text que s'inclou en l'element **<na>** serà el que es recuperi de l'escàner o del fitxer original no editat.

8.8.6.13 Detecció dels finals de frases que no estan indicats pel signe de puntuació "." (*punt*)

** , ; <s>, </s>;**

Un altre tipus d'informació que cal consignar durant la fase de marcatge és la delimitació de frases i paràgrafs. Aquesta informació s'insereix al document per mitjà d'un procés semiautomatitzat des de la plataforma MS-DOS.). Es tracta d'un procés que marca els límits de cada frase del text a partir dels punts que troba (signes de puntuació "."). Tanmateix, hi ha una sèrie de contextos propis de frontera entre frases que aquest procés no és capaç de detectar. Es tracta dels següents signes de puntuació:

":" dos punts
"!" signe d'exclamació
"?" signe d'interrogació

En conseqüència, un cop s'han realitzat totes les fases de marcatge anteriors, convé aplicar la macro Revisar Puntuació, dissenyada per detectar i marcar pertinentment aquest tipus de fronteres entre frases:

</s> <s> = final i inici de frase
** <s>** = inici de paràgraf i frase
**</s> ** = final de frase i paràgraf

<s>	=	inici de frase
</s>	=	final de frase

8.8.6.14 Inserció de noms propis

<name>, </name>

Tot i que els noms propis els reconeix automàticament el preprocés, cal tenir en compte que les seqüències inicials en majúscula no es reconeixeran com a noms propis. En aquest cas tenim dues opcions:

a) esperar que surtin a la neologia com a candidats i aleshores inserir-los les marques de <name></name>

b) marcar-los si els veiem en aquesta fase. Per això posarem les marques següents:

després del punt </s><s><name>NOM TROBAT</name>. Si el nom propi apareix a inici de paràgraf cal posar també les marques d'inici i tancament de paràgraf </s> <s><name> NOM TROBAT</name>

8.8.6.15 Com cal guardar el document

A continuació cal guardar el text en format Solo texto con saltos de línea.

8.8.7 Precaucions generals

És necessari tenir algunes precaucions a l'hora d'inserir etiquetes en un text. En general, hem de recordar que l'estàndard SGML organitza jeràrquicament el text i tots els seus components a mode de nines russes. Presentem seguidament algunes qüestions que cal tenir en compte.

8.8.7.1 Organització de les divisions d'un text de l'àrea de dret

És habitual que els textos s'organitzin en capítols, seccions, etc. Podem considerar cada una d'aquestes parts com a peces que han d'encaixar una dins de l'altra. Així no és possible tenir la marca de final de secció d'un capítol més enllà de la marca de final del capítol.

Exemples:

```
Correcte                        Incorrecte
<div1>                          <div1>
        <div2>                                  <div2>
                <div3>                  <div2>
                </div3>                 </div2>
        </div2>                                 </div2>
        <div2>                                  <div2>
        </div2>                                 </div2>
<div1>                          <div1>
```

També són incorrectes les situacions següents:

```
<div1>                          <figure>
        <div2>                                  <head>
                <div3>                          </figure>
                </div3>         </head>
                <div3>
        </div2>
                </div3>
        <div2>
        </div2>
<div1>
```

8.8.7.2 Etiquetes el valor de les quals s'estén més enllà de la seva jerarquia

Les etiquetes d'inici i final de qualsevol element SGML no poden estendre més enllà de la seva jerarquia. Però és possible trobar-se amb elements la vigència dels quals ha d'estendre's més enllà de la seva jerarquia. Per exemple, els casos en què les marques de *hi* (característica tipogràfica especial) o *foreign* (seqüència en un altre idioma) es prolonguen més enllà d'una frase o paràgraf. El procediment en aquests casos consisteix a tancar l'element en qüestió dins de la frase o paràgraf i tornar a obrir-lo en la frase o paràgraf següent.

A continuació es presenta un exemple del procediment a seguir per a un paràgraf que inclou un text en un altre idioma:

> *Tal i com afirma J. ALONSO DAVILA "El derecho publicitario en España y la Directiva de la CEE de 10 de septiembre de 1984 sobre Publicidad Engañosa" La Ley 1985-4, p. 1046 i ss, que "el ritmo de jurisdicción ordinaria no es suficientemente ágil para responder a los conflictos publicitarios. En la actualidad las campañas duran semanas o meses, mientras que los juicios ordinarios, desafortunadamente, duran años. Por eso las partes no recurren a la via judicial: porque la medicina llegaría cuando el enfermo estuviese muerto" (p. 1047).*

Pel que fa a la marca de paraules en un altre idioma, el text ha de quedar marcat de la manera següent:

> <s>*Tal i com afirma J. ALONSO DAVILA* ***<foreign lang=ES>*** *"El derecho publicitario en España y la Directiva de la CEE de 10 de septiembre de 1984 sobre Publicidad Engañosa" La Ley* ***</foreign>*** *, 1985-4, p. 1046 i ss, que* ***<foreign lang=ES>*** *"el ritmo de jurisdicción ordinaria no es suficientemente ágil para responder a los conflictos publicitarios* ***</foreign>****.</s>****<s lang=ES>*** *En la actualidad las campañas duran semanas o meses, mientras que los juicios ordinarios, desafortunadamente, duran años.</s>****<s lang=ES>*** *Por eso las partes no recurren a la via judicial: porque la medicina llegaría cuando el enfermo estuviese muerto" (p. 1047).*</s>

8.8.8 Inserció de les marques de paràgraf i frase

El procés d'inserció de les marques de paràgraf i frase es realitza a través d'un programa que funciona a l'entorn DOS. Per aplicar-lo, cal fer un doble clic a la icona MS-DOS del menú principal. Un cop realitzada aquesta operació, caldrà executar la següent seqüència d'ordres:

a. Canviar a k: <Return> i donar l'ordre:

K:\ cd tmpÁREA <Return>

b. Obrir cada una de les mostres del document i canviar l'extensió *.txt* per el seu número de mostra [5]

K:\TMPÀREA> move d000101.txt d0001.1 <Return>

c. Donar l'ordre

K:\TMPÀREA> Addsyp nom de la mostra que cal processar <Return>

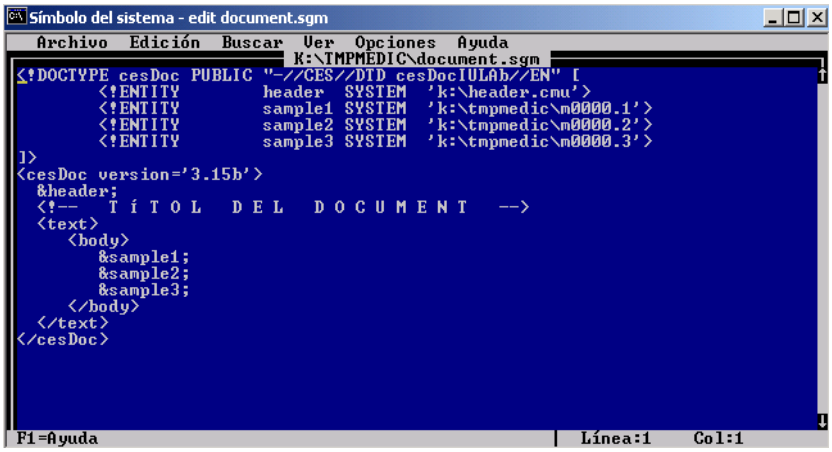
Aquesta última ordre és la que realment comporta el procés d'inserir les marques i per aquesta raó s'ha de repetir per a cada una de les mostres.

d. Construir un document SGML fictici des de DOS:

K:\TMPÁREA> copy document.sgm *nom_usuari*.sgm <Return>

e. Editar el nou document

K:\TMPÁREA> edit *nom_usuari*.sgm <Return>



El document que editem s'ha d'adaptar a l'estructura del document que volem analitzar. Així, cal augmentar o suprimir les línies d'<!ENTITY samplen> que siguin necessàries; també s'han de canviar les seqüències 0000 pel número que identifica el document,per exemple:

<!ENTITY sample 0 SYSTEM k:\TMPDRET\d0000.0

[5] Això es pot fer també des de l'*Explorador de Windows*.

s'ha de canviar per:

<!ENTITY sample 1 SYSTEM k:\TMPDRET\d0001.1

i substituir l'expressió[6]

<!-- TÍTOL DEL DOCUMENT-->

pel títol del document, i revisar que la informació que hi ha sota d'aquest coincideixi amb el nombre de les mostres.

Finalment, cal donar les ordres:

Archivo > Guardar
Archivo > Salir

8.8.9 Anàlisi del document (parser)

És molt habitual que en el procés de marcatge del document es cometin errors que un analitzador s'encarrega de detectar. És el que anomenem "passar el parser".

Per analitzar el document, cal fer:

k:\TMPDRET> parser *nom_usuari*

Per a totes aquelles situacions detectades com a incorrectes, l'analitzador dóna informació que permet localitzar el problema, per exemple:

k:\tmpdret\d0032.1:9:3:E:...

L'exemple anterior assenyala que està verificant la mostra d0032.1 i que en la línia 9, columna 3, hi ha un problema. El procediment per corregir aquest problema és editar la mostra que es vulgui corregir. Un cop es té la mostra editada, cal anar a la línia corresponent per verificar el problema i solucionar-lo. La correcció es pot fer des de qualsevol programa d'edició de textos com ara l' Edit Pad[7] (així no cal tancar la finestra del parser cada cop, encara que s'han de salvar els canvis abans de verificar-lo).

Amb l'Edit pad es treballa amb el document d'errors i l'usuari.sgm, es corregeixen els errors, es tanca el document errors i es torna a passar el "parser" desde MS-DOS perquè s'actualitzi el document d'errors.

8.8.10 Protecció de les mostres marcades

Després de corregir-los i quan el parser ja no dóna cap error, cal canviar el nom del document SGML fictici i protegir les mostres contra escriptura:

[6] En el cas del català, cal posar els accents i els punts volats de les eles geminades amb codis SGML:
´ (& acute;), ` (& grave;), etc. (l'espai correspon a la vocal, per exemple: á).
l·l (l·l)

[7] Cal verificar que la orde *Opción > Romper líneas* no hi està activada.

K:\TMPDRET> move *nom_usuari*.sgm k:\panorama\ÀREA*nom_document*.sgm

K:\TMPDRET> attrib +r k:\panorama\ÀREA*nom_document*.sgm

K:\TMPDRET> attrib +r *nom_document*.*

on ÀREA és: dret / medicina / mediamb / info / econ.

I ja podem passar a fer la capçalera del document.

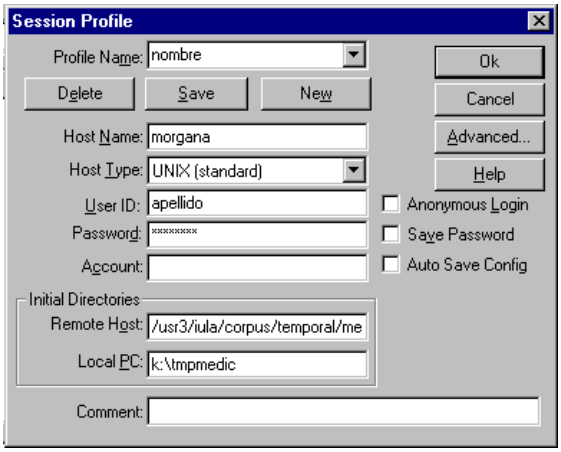
8.8.11 Confecció de la capçalera

Un problema freqüent a l'hora de passar un text d'un entorn de treball a un altre (ex. DOS a Windows, DOS a UNIX, etc.) són els caràcters especials (vocals accentuades, *ñ*, etc.). De fet, cada entorn de treball representa aquests caràcters internament d'una manera diferent i no és compatible amb els altres. Per solucionar aquest inconvenient, l'estàndard SGML defineix el concepte d'*entitat*, que consisteix molt bàsicament a substituir aquests caràcters especials per una seqüència de caràcters no especials. Per exemple ***á*** se substitueix per á o ***ç*** per ç.

Aquests canvis es fan automàticament en el moment de fer la capçalera, que té dues fases:

8.8.11.1 des de l'FTP

L'ftp permet copiar les mostres i el document.sgm al servidor. La finestra esquerra (client) correspon a NOVEL, i la dreta (servidor) correspon a UNIX. Així:



a) hem de copiar les mostres

des de local system> k:\tmpÀREA\

a remote system> cd /usr3/iula/corpus/temporal/ÀREA/mostres

b) i el document.sgm

des de local system> k:\panorama\ÀREA

a remote system> cd /usr3/iula/corpus/temporal/ÀREA

8.8.11.2 des de Morgana

Des de Morgana (icona Conexión Servidor) és des d'on efectivament s'elabora la capçalera.

a) s'ha de seleccionar la terminal:

vt100

b) només cal escriure:

setheader

c) respondre les preguntes que fa, entre elles el número dels documents paral·lels:

IMPORTANT:

En cas d'equivocació, s'ha de donar l'ordre "Control + c" i tornar a fer la capçalera des de l'inici, ja que el programa no permet esborrar les dades.

- àrea del document (dret, economia, medi ambient, medicina, informàtica, general)
- idioma en què està escrit el document (català, castellà, anglès, francès, alemany)
- número de document (4 xifres)
- número de mostres del document
- text original o traduït
- altres idiomes del document (no n'hi ha, hi serà o ja hi és)
    * número del document [paral·lel] en [idioma] (5 xifres)
    * el títol correspon al document en [idioma] (s/n)
- subàrea a què pertany el document [en el cas de medicina, hi ha encara una altra classificació per aquells que pertanyen al projecte genoma]
- publicació independent [sempre s'ha posar l'opció "sí"]
- títol de l'obra
- autor de l'obra (cognoms, nom)
- Responsable intel·lectual de l'obra [traductor, editor, etc.]
- Número de l'edició
- Data de publicació (dd/mm/aa)
- Lloc de publicació
- Qui ha publicat l'obra (organització, persona)
- Nom de l'organització o de la persona que ha publicat l'obra
- Classe de codi d'identificació de l'obra (isbn,)
- Si s'ha agafat el document sencer (sí= revistes, no= capítols de llibres) [en cas de *no*, es demana el número inicial i final de cada una de les mostres]
- Informació addicional que es vulgui afegir
- Mètode d'obtenció del document
- Tipus de document (propi de l'àrea, normatiu, instrumental)
- Rectificació de dades
- Comprovació del document SGML

8.8.12 Epíleg

Un cop marcat i arxivat el document cal tornar a completar les dades del document a la base de dades del CT. Concretament, s'han d'omplir les caselles relatives al nombre total de paraules del document i el número sgm del document.

CONTROL DE LES ETAPES DE TREBALL EN LA CADENA DE PROCESSAMENT DEL CORPUS TÈCNIC
DOCUMENT : _____
DATA D'INICI: _____
DATA D'INCORPORACIÓAL CORPUS TÈCNIC: _____

ETAPA / DOCUMENT		OBSERVACIONS
Selecció de documents i base de dades (1)		
Escaneig		
Revisió ortogràfica		
Marcatge		
Parser (MS-DOS)		
Capçalera		
Base de Dades Access (1+ .sgm)		
Preprocés (pretag)[8]		
Palic-Ambilic (ctget)		
Ctput.		

Neologia:

Correcció d'errors a les mostres		
Incorporació de neologia		
Base de Dades Access (B)		
Parser (UNIX)		
Espera d'incorporació dBASE		
Preprocés (pretag)		
Palic-Ambilic (ctget)		
Verificació de la neologia introduïda		

Ctput		
Apilado		
TagCorpus		

Codificació dels documents (CWB)		
FullBalanceig		
Preindexació		
Comprovació concordances bwanaNet		
Base de dades (6)		

[8] En el cas de l'anglès l'ordre és `en-cg2ct.pl -d .....sgm -bd` i després d'aquest pas hem de passar directament a la codificació dels documents (CWB)

Document resum

Títol de l'obra:

Autor:

Identificador de document: Identificador sgm:
Classificació dins de l'arbre d'àrea:
Tipus de document:

Mostra 01

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 02

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 03

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 04

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 05

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 06

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 07

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 08

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 09

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 10

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 11

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Mostra 12

Nombre de paraules:	
Pàg. inicial:	Pàg. final:

Data inici: ____/____/200_

Data fi:____/____/200_

8.9 Annex 9: Processament lingüístic de documents en català

En aquest document s'explica com s'ha de processar lingüísticament un document català, de tal manera que s'assignin tots els lemes i les etiquetes gramaticals als mots inclosos en el diccionari.

Per tal de processar el document, s'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el password de l'usuari.

Des de *morgana*, escriurem les següents ordres a la línia de comandes:

nsgmls –s nom_del_document.sgm

(Amb aquesta ordre comprovem que no hi ha cap error nou de marcatge que s'hagi afegit en algun moment).

Si hi ha algun problema de marcatge, ens indicarà la mostra i la línia en què s'ha trobat el problema, que evidentment haurem de corregir en el mateix editant el fitxer (si es domina l'entorn UNIX) o mitjançant un editor de textos (Crimsom, EditPad).

pretag5.pl -i nom_del_document.sgm -bd –q (anàlisi pròpiament)

apilado1xl.pl -i nom_del_document.sgm -des -bd –q (pas al format verticalitzat)

TagCorpus.pl -i nom_del_document.sgm –q (desambiguació estadística)

Per exemple, si volem analitzar el document m00173.sgm de medicina escriurem:

pretag5.pl -i m00073.sgm -bd -q

apilado1xl.pl -i m00073.sgm -des -bd -q

TagCorpus.pl -i m00073.sgm -q

Quan fem l'ordre d'apilado, cal posar atenció als resultats per tal de comprovar que no surti cap dels mots neològics que havíem introduït en la fase de detecció de neologia (vegeu annex 2). Si no sorgeix cap problema, un cop fet el TagCorpus passarem a seguir les instruccions per incorporar els documents del corpus a *bwananet*. El protocol on s'explica com dur a terme aquest procés es troba a l'annex 12.

8.10 Annex 10: Processament lingüístic de documents en castellà

En aquest document s'explica com s'ha de processar lingüísticament un document castellà, de tal manera que s'assignin tots els lemes i les etiquetes gramaticals als mots inclosos en el diccionari.

8.10.1 Comprovació de l'estructura del document

En tots els casos, abans d'iniciar el preprocés del document, cal comprovar que no s'hagi afegit algun nou error de marcatge. Per aquest motiu, a l'entorn Unix, al servidor *morgana* passarem un nou parser al document sencer.

S'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el password de l'usuari.

Des de *morgana*, escriurem la següent ordre a la línia de comandes:

```
nsgmls –s nom_del_document.sgm
```

Si hi ha algun problema de marcatge, ens indicarà la mostra i la línia en què s'ha trobat el problema, que evidentment haurem de corregir en el mateix entorn Unix editant el fitxer (si es domina aquest entorn) o mitjançant un editor de textos (Crimsom, EditPad).

8.10.2 Preprocessament del document

Per tal de preprocessar el document, s'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el password de l'usuari.

Des de *morgana*, escriurem la següent ordre a la línia de comandes:

```
pretag1.pl -i nom_del_document.sgm -bd -q
```

Si volem analitzar el document *g00159.sgm* escriurem:

```
pretag1.pl -i g00159.sgm -bd
```

Un cop preprocessat el document hem d'obrir una finestra MS-DOS i començar l'anàlisi pròpiament lingüística.

8.10.3 Anàlisi lingüística del document escollit

Obrim una finestra MS_DOS.

A la línia de comandes escrivim:

```
ctget nom_del_document.sgm
```

Si volem analitzar el document 105 de medicina escriurem:

```
ctget m00105.sgm
```

En donar aquesta ordre, se'ns demanarà el nom d'usuari i el password que utilitzem a l'entorn Novell.

Un cop finalitzada l'anàlisi hem de donar l'ordre següent:

```
ctput nom_del_document.sgm
```

que en el nostre cas d'exemple seria:

ctput m00105.sgm

En donar aquesta ordre, se'ns demanarà el nom d'usuari i el password que utilitzem quan treballem a l'entorn Unix.

Quan es dóna aquesta ordre, pot ser que a la pantalla aparegui algun missatge que ens indicarà que alguna/es mostra/es d'aquell document no s'han analitzat correctament. Si succeeix això, convindria revisar el fitxer c:\tmp\lemacial.dat i comprovar amb l'ajuda de la persona encarregada de corpus què pot ser el que està passant.

El següent pas és la consulta de la llista de paraules no reconegudes pel diccionari i la introducció de la neologia als diccionaris de l'IULA. (Vegeu annex 4).

El fitxer en què havíem detectat la neologia haurà de tornar-se a processar per tal que pugui incorporar-se definitivament a la base de dades del CT de l'IULA. Aquestes són les instruccions a seguir:

pretag1.pl -i nom_del_document.sgm -bd –q (entorn Unix)

ctget nom_del_document.sgm (entorn Novell)

ctput nom_del_document.sgm del (entorn Novell)

apilado1xl.pl -i nom_del_document.sgm -bd –q (entorn Unix)

TagCorpus.pl -i nom_del_document.sgm –q (entorn Unix)

El pròxim procés es la inclusió del document a la BD textual (bwanaNet). El protocol on s'explica com dur a terme aquest procés es troba a l'annex 12.

8.11 Annex 11: Processament lingüístic de documents en anglès

En aquest document s'explica com s'ha de processar lingüísticament un document en anglès, de tal manera que s'assignin tots els lemes i les etiquetes gramaticals als mots inclosos en el diccionari.

8.11.1 Anàlisi lingüística del document del CT i desambiguació estadística

Es tracta de processar lingüísticament el document que hem escollit, de tal manera que s'assignin tots els lemes i les etiquetes gramaticals als mots inclosos en el diccionari.

Per tal de processar el document, s'ha d'obrir una finestra en entorn UNIX, seleccionar el servidor morgana i entrar el password de l'usuari.

Des de *morgana*, escriurem les següents ordres a la línia de comandes:

nsgmls –s nom_del_document.sgm (amb aquesta ordre comprovem que no hi ha cap error nou de marcatge que s'hagi afegit en algun moment)

Si hi ha algun problema de marcatge, ens indicarà la mostra i la línia en què s'ha trobat el problema, que evidentment haurem de corregir en el mateix editant el fitxer (si es domina l'entorn UNIX) o mitjançant un editor de textos (Crimsom, EditPad).

Quan no hi ha cap error cal seguir les instruccions següents:

en-cg2ct.pl –d .....sgm –bd

Amb aquesta ordre s'afegeixen els lemes i etiquetes que corresponen a cada mot i a més a més es fa directament la desambiguació estadística.

Per exemple si volem processar el document m00732.sgm l'ordre serà:

en-cg2ct.pl –d m00732.sgm –bd

Si no sorgeix cap problema, passarem a seguir les instruccions per incorporar els documents del corpus a *bwananet*. El protocol on s'explica com dur a terme aquest procés es troba a l'annex 12.

8.12 Annex 12: Incorporació dels documents del CT a *bwanaNet*

Aquest document explica el conjunt de passos a seguir per fer que un document ja processat lingüísticament pugui ésser accessible des de *bwanaNet*.

8.12.1 Visió general del procés

El corpus de l'IULA utilitza el CWB[1] per fer que els documents siguin accessibles des d'Internet. El procés d'incorporació d'un document a la base de dades textual té bàsicament dues parts:

La primera és manual, és a dir, un operador humà ha d'efectuar certes operacions per tal de comprovar (i corregir, si és necessari) que no hi hagi errors en el document.

La segona part del procés, en canvi, és automàtica i no requereix cap intervenció humana.

Tot el procés que descrivim en aquest protocol es realitza des de morgana.upf.es. A la Figura 1, trobareu una representació visual d'aquests processos i com s'insereixen en la cadena de processament del CT. Tots els processos que s'expliquen en aquest document s'han d'executar des del directori arrel de cada usuari.

[1] CWB (*Corpus Work Bench*) és un conjunt d'eines desenvolupades al IMS (*Institut fur Maschinelle Sprachverarbeitung*) de la Universitat de Stuttgart per a la recuperació de text complet a partir de corpora voluminosos. Trobareu més informació a: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/

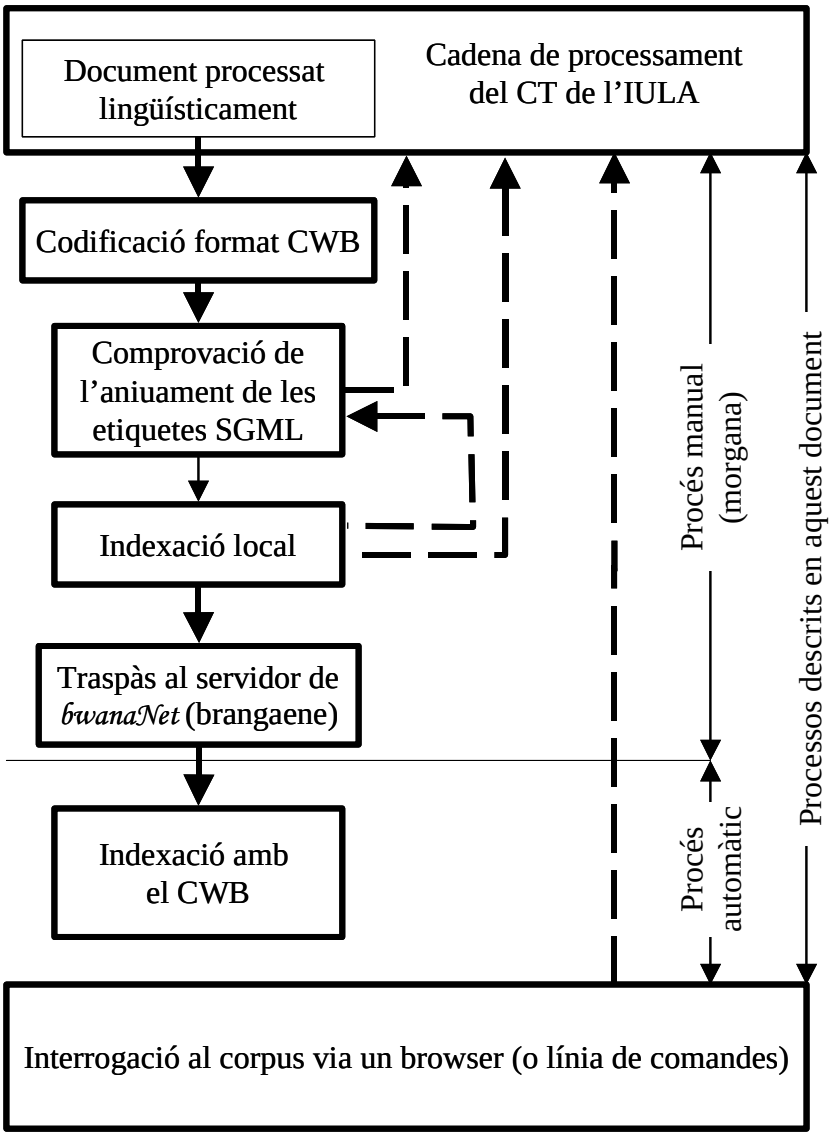


Figura 1. Visió global del procés

8.12.2 Codificació dels documents en format CWB

El CWB permet incorporar textos a la base de dades textual sempre que estiguin en un format determinat. Per generar el text en el format adequat cal executar la comanda creaCQPfiles.pl. El format bàsic d'aquesta comanda és el següent:

```
creaCQPfiles.pl –d document_sgml
```

Per a convertir el document m00206 la comanda és:

```
creaCQPfiles.pl –d m00206.sgm
```

Com a resultat de l'execució d'aquesta comanda es creen dos fitxers:

`m00206.cqp`	Document únic que conté totes les mostres del document m00206.sgm codificades en format CWB. Sobre aquest document s'han d'efectuar totes les comprovacions.
`provacqp_m00206`	*Script* a executar en l'etapa de preindexació (veure secció 8.12.3.2)

El format del nom del primer fitxer és el nom base del document (m00206) amb l'extensió 'cqp' mentre que el nom del segon fitxer es construeix amb la seqüència 'provacqp_' seguida del nom base del document.

El format del fitxer creat preveu dos tipus d'informació:

a) text:

forma tabulació *lema* tabulació *etiqueta* tabulació *tipus_de_token*

b) marques SGML:

etiqueta

Algunes etiquetes i/o atributs s'han eliminat degut a limitacions del CWB.

8.12.3 Correcció d'errors

El procés de detecció d'errors es divideix en almenys dues etapes, la primera per trobar errors en el marcatge estructural i la segona per trobar errors que es podrien presentar en la fase d'indexació del CWB. En las seccions següents expliquem cadascuna d'aquestes dues etapes.

8.12.3.1 Verificació del marcatge estructural

El marcatge estructural que s'ha introduït en la cadena de processament del corpus podria tenir errors que fins ara no es podien detectar o bé que el programa d'explotació utilitzat fins ara no ens en informava. Per exemple, fins ara es podien barrejar majúscules i minúscules en el nom de la etiqueta com per exemple <Figure> o <Item>. Aquesta codificació no es permesa pel CWB perquè tracta el text com si fos XML i per tant s'han de substituir per <figure> o <item> respectivament.

Per fer aquesta comprovació s'ha d'executar el programa fullBalanceig.pl Aquest programa té dues modalitats: a) la primera per comprovar el balanceig de totes les etiquetes incloses en un document i b) la segona per analitzar específicament el comportament d'una determinada etiqueta de la qual sospitem que hi ha problemes.

A continuació, veurem el resultat obtingut en la primera modalitat amb un exemple en concret (més endavant veurem la segona modalitat).

La següent és la línia de comanda bàsica:

morgana_26: fullBalanceig.pl –d *nom_del_document*

Així, si el que volem es comprovar el document m00206, la comanda a executar és:

morgana_26: fullBalanceig.pl -d m00206.cqp

que donarà com a resultat la taula següent:

```
Comprovació del balanceig de la etiquetes "SGML" en el
document "m00206.cqp".

 ----------------------------------------
      Tag       Num     bal    maxnest
 ----------------------------------------
     abbr       133     0       1
     cell       104     0       1
     date         2     0       1
     div1        11     0       1
      doc         1     0       1
   figure        48     0       1
  foreign         6     0       1
      gap        15     -       -
     head       226     0       1
       hi       148     0       1
     item       285     0       3
     list        78     0       3
      loc       135     0       1
       na       329     0       1
     name       388     0       1
     note         2     0       1
      num       940     0       1
        p       826     0       1
      ptr         2     -       -
      row        37     0       1
        s      1588     0       1
    table        42     0       1
     text      2203     0       3
 ----------------------------------------
```

Aquesta taula conté quatre columnes amb la següent informació:

Columna	Significat
`Tag`	Etiqueta SGML
`Num`	Nombre d'aparicions de `Tag`
`bal`	Balanceig de `Tag`
`maxnest`	Aniuament màxim de `Tag`

Las condicions d'error poden ser produïdes per tres circumstàncies:

1. A la columna `Tag` apareix una etiqueta SGML amb algunes de las lletres en majúscula.
2. La columna `bal` controla que totes les etiquetes que s'obren (ex. `<s>`) també es tanquin (ex. `</s>`). Qualsevol valor diferent de zero (0) que aparegui en aquesta columna indica un error. Si és més gran que zero indica que manquen etiquetes de tancament i si és més petit que zero indica que manquen etiquetes d'obertura.
3. La columna `maxnest` controla quantes vegades s'incrusta una etiqueta dins d'una altra del mateix tipus. Per exemple, una seqüència d'aquest tipus:

```
<s> ...
     <list>
             <item><s> ...</s></item>
             <item><s> ...</s></item>
     </list>
</s>
```

implica que columna maxnest per l'etiqueta s tindria un valor de 2. El CWB imposa uns límits al nivell d'aniuament que permet sobre les etiquetes SGML. Si aquests valors se superen l'etapa del preindexador indicarà el problema (amb un missatge del tipus: 13 <s> regions dropped because of deep nesting.). El valor establert pel nivell màxim d'aniuament és zero (0), excepte per a les etiquetes SGML que s'indiquen a continuació:

etiqueta	Màxim aniuament permès
<s>	3
<p>	2
<list>	4
<text>	6

Qualsevol de les condicions d'error ja mencionades s'ha de corregir (normalment a l'original) abans de passar a la següent fase. A continuació analitzem en detall una situació d'error i el seu procés de detecció.

8.12.3.1.1 Error en el tancament d'una etiqueta SGML

Un exemple de situació d'error és la manca de l'etiqueta de tancament d'un títol. En l'exemple anterior la taula obtinguda tindria aquest aspecte:

```
morgana_27: fullBalanceig.pl -d m00206.cqp
Comprovació del balanceig de la etiquetes "SGML" en el
document "m00206.cqp".

 ----------------------------------------
       Tag       Num     bal    maxnest
 ----------------------------------------
       .....
       gap        15     -       -
      head       226     1       2      <-- ?
        hi       148     0       1
       .....
----------------------------------------
```

La presència d'1 en la columna bal ens hauria de fer pensar en una situació d'error. Aquesta situació també és la responsable que la columna maxnest mostri el valor 2. Per tant, el que s'ha de fer és analitzar en detall les etiquetes <head> del document. Entrem així en la segona modalitat d'execució del programa fullBalanceig.pl. La comanda a executar i el resultat obtingut són els següents:

```
morgana_28: fullBalanceig.pl -d m00206.cqp -t head
Comprovació del balanceig de la etiqueta "head" en el
document "m00206.cqp".

$count= 1
Nivell màxim d'aniuament: 2 (222 vegades, la primera a
1033)
Consulteu el resultat detallat en m00206.cqp.head
```

El resultat obtingut confirma les dades de la execució anterior (veure: $count= 1 i Nivell màxim d'aniuament: 2) i crea un fitxer (m00206.cqp.head) amb més informació.

Aquesta és la segona modalitat de funcionament del programa fullBalanceig.pl. En aquesta modalitat es genera un fitxer auxiliar que indica cada línia en la qual es tanca una etiqueta oberta prèviament. També indica la línia on troba una etiqueta de tancament sense la corresponent d'obertura. En aquest últim cas s'atura l'execució.

En l'exemple que estem analitzant el contingut del fitxer auxiliar és el següent:

Comprovació del balanceig de la etiqueta "head" en el document "m00206.cqp".

```
conteo a cero en linea 6
conteo a cero en linea 13
conteo a cero en linea 787
$count= 1
Nivell màxim d'aniuament: 2 (222 vegades, la primera a
1033)
```

Aquesta informació indica que en la línies 6, 13 i 787 (de m00206.cqp) hi ha etiquetes de tancament d'altres prèviament obertes. A partir d'aquesta línia ja no hi ha etiquetes de tancament que balancegin les d'obertura. Per tant l'error està més endavant de l'última línia de balanceig i segurament en la propera etiqueta d'obertura de títol (<head>). Per tant, el següent pas és obrir el document (en el format CWB, és a dir m00206.cqp) per identificar la zona on es troba l'error. El procediment a seguir és anar a l'última línia de balanceig i partir d'aquí buscar la primera aparició d'una etiqueta d'obertura de títol i comprovar si hi ha o no l'etiqueta de tancament. El contingut del document que presentem a continuació confirma la manca d'una etiqueta de tancament de títol:

```
de      de      P       TOK
los     el      AMP     TOK
siguientes      siguiente       JQ--6P  TOK
eslabones       eslabón N5-MP   TOK
fundamentales   fundamental     JQ--6P  TOK
(       =       Z       DLS
<abbr>
fig.    fig.    N5-66   TOK
</abbr>
<num pos='X'>
1-1     num     X       TOK
</num>
).      =       Z       DLS
</s>
</text>
</p>
<text>
<head type=sub rend="bo" tree="1.1.12">
Observación     observación     N5-FS   TOK
</text>
<p>
<text>
<s tree="1.1.13.1">
La      el      AFS     TOK
capacidad       capacidad       N5-FS   TOK
de      de      P       TOK
observación     observación     N5-FS   TOK
es      ser     VDR3S-  TOK
absolutamente   absolutamente   D6      TOK
esencial        esencial        JQ--6S  TOK
para    para    P       TOK
todo    todo    EN--MS  TOK
buen    bueno   JQ--MS  TOK
científico      científico      JQ--MS  TOK
,       =       Z       DLD
```

Manca l'etiqueta de tancament (</head>)

El proper pas és verificar en quin moment de la cadena de processament del corpus (i per quina raó) s'ha perdut l'etiqueta. Qualsevol canvi s'ha d'efectuar a l'original, s'ha de repetir el processament i comprovar que el problema s'ha resolt.

La tercera situació d'error és molt poc probable però si es presentés fa necessari simplificar el marcatge estructural per disminuir la complexitat de les parts del document involucrades.

8.12.3.2 Preindexació

L'objectiu d'aquesta etapa és detectar qualsevol situació que pugui generar una condició d'error al indexar el document que estem processant juntament amb la resta de documents del corpus. Un error típic que s'ha de detectar és la presencia de caràcters que el CWB no pot indexar com ara els signes '<' i '>'.

Per efectuar aquesta comprovació només cal executar l'*script* que hem generat en l'etapa de creació del document en format CWB (secció 8.12.2). Un exemple típic d'execució d'aquest script amb una condició d'error és la següent:

```
morgana_39: provacqp_m00206
rm CWBdata/*
...
cwb-encode m00206.cqp ...
```

Malformed tag < < W TOK, inserted literally (file m00206.cqp, line #3644).

El missatge d'error que proporciona el procés d'indexació normalment es prou explícit com per localitzar l'error dintre del fitxer en format CWB. En aquest cas, ens indica que en la línia 3644 hi ha un error i a més a més ens mostra el contingut de la línia en qüestió. El procediment a seguir és localitzar aquesta línia en una de les mostres del document que estem processant, fer la/les correccions pertinents i tornar a processar el document.

Un altre exemple d'execució d'aquest script amb una condició d'error més complexa és quan hi ha un aniuament correcte des de el punt de vista del balanceig de les etiquetes però no previst pel CWB. Suposem un document que incorpora un fragment com el següent:[2]

... de forma que sólo aparecían como fachadas o como galerías cubiertas dentro de una manzana de viviendas (el más famoso es la galería <name>Vittorio Emmanuele II de <name>Milán</name></name>).</s></p>

El procés d'indexació donaria un missatge d'aquest tipus:

```
1 <name> regions dropped because of deep nesting
```

Aquest error no es detecta abans perquè les etiquetes `<name>` estan ben balancejades però aquest aniuament no està previst per la indexació (veure Secció 8.12.3.1, condició d'error 3). Per detectar aquest error s'ha de tornar a aplicar el programa fullBalanceig.pl amb l'opció `-t` sobre l'etiqueta SGML que causa el problema. En aquest cas:

```
morgana% fullBalanceig.pl -d e00143.cqp -t name
```

que dóna com a resultat:

```
Comprovació del balanceig de la etiqueta "name" en el
document "e00143.cqp".
$count= 0
Nivell màxim d'aniuament: 2 (1 vegades, la primera a
18112)
Conselteu el resultat detallat en e00143.cqp.name
morgana%
```

La resposta del programa indica que hi ha etiquetes aniuades fet que ja fa pensar en un error.[3] Com en els casos anteriors la solució és trobar l'error en el document que estem processant, fer la/les correccions pertinents i tornar a processar el document.

[2] Aquest és un cas fictici. Normalment no s'han de posar etiquetes de nom propi a l'original.

[3] Aquesta mateixa informació es dóna també quan s'executa el programa fullBalanceig.pl sense l'opció `-t`.

8.12.4 Traspàs del document al servidor de *bwanaNet*

Una vegada que ens hem assegurat que el document no conté cap error podem traspassar-lo al servidor de *bwanaNet*. Per efectuar aquesta operació cal executar la comanda cwbput.pl.

El format bàsic d'aquesta comanda és la següent:

cwbput.pl document_cqp

Per a convertir el document m00206 la comanda és:

cwbput.pl m00206.cqp

Com a resultat de l'execució d'aquesta comanda, en primer lloc, el programa ens demana el nom d'usuari i la contrasenya a brangaene.upf.es i a continuació transferirà (per ftp) el fitxer en format CWB i esborrarà els fitxers temporals. Una sessió típica és la següent:

```
morgana_41: ../bin/cwbput.pl m00206.cqp
Nom d'usuari: vivaldi
Contrasenya:
Transferint document m00206.cqp ...ok
Establint els permissos sobre aquest document ...ok
Esborrant directoris i fitxers temporals ... ok ;-)
morgana_42:
```

8.12.5 Compleció del procés

Està previst que el document s'incorpori definitivament a *bwanaNet* a partir del primer dia de la setmana següent. Per tant, l'últim pas a seguir és comprovar, mitjançant la interfície de *bwanaNet* (http://brangaene.upf.es/bwananet0/bwananet1a.es.htm), que el document s'ha incorporat efectivament i que es possible extreure'n concordances.

També s'haurà de comprovar que no quedin errors de marcatge SGML. Per això cal comprovar els següents patrons per a totes les llengües (concordança estàndard):

1r)

Forma		
Lema		
POS	nom	n. propi

D'aquesta manera detectarem seqüències errònies com aquestes:

Adecuado ni para diversificar sus ***exportaciones. Los*** países que han logrado
estudiados en los últimos diez ***años. En*** otros 17 países se observó
formas , no admite conclusiones ***simples. Sin*** embargo , los acontecimientos

Però també seqüències correctes com ara:

noviembre de 1991 , el ***presidente Hosni Mubarak*** anunció la creación del

las costas y del ***mar Mediterráneo*** , de la flora ,
soleado ático situado en la ***calle Montesquinza*** , inmueble muy señorial ,

De nou, si es detecten seqüències incorrectes s'ha de fer la correcció de les mostres originals i tornar a processar el document.

2n)

Forma		
Lema		
POS	n. propi	nom

D'aquesta manera detectarem seqüències errònies com aquestes:

ni para diversificar sus exportaciones. ***Los países*** que han logrado buenos
para corregir estas distorsiones ? ***Algunas opciones*** son la agrupación de
en función de l riesgo. ***Los fondos*** que comercializa la Caja de

Però també seqüències correctes com ara:

Que consideremos que existen en ***España regiones*** muy pobres en relación
cobrando en la delegación de ***Madrid servicios*** realizados en diversas
la tecnología innovadora .</item></s> <s>El ***Estado miembro*** participa

Si es detecten seqüències incorrectes s'ha de fer la correcció de les mostres originals i tornar a processar el document.

3r) També necessitarem fer una concordança complexa per buscar un altre tipus d'error. Cal escriure:

- a la primera casella :

```
[word=".+\." & (pos="N5.*"|pos="JQ.*")][word=")" & pos="Z"]
```

- a la segona casella:

```
cat Last;
```

D'aquesta manera detectarem seqüències errònies com aquestes:

querría más espacio del ***disponible.)*** En cierto sentido se tra
sociología y la ciencia ***política.)*** Todas estas materias tra
rada sin importaciones ni ***exportaciones.)*** El Estado obtiene sus in

Però també seqüències correctes com ara:

violeta , de menos de 300 ***nm.)*** perturban moléculas que
(obras públicas , agua , ***etc.)*** con los que existen obli
empresas , sindicación , ***etc.)</s>*** <s>Desde esta pers

Si es detecten seqüències incorrectes s'ha de fer la correcció de les mostres originals i tornar a processar el document.

4r) També cal comprovar els següents patrons (concordança estàndard) que depenen de la llengua:

a) Català i Castellà:

Forma	.*\.	
Lema		
POS	puntuació	n. propi

b) Anglès:

Forma	.*\.	
Lema		
POS	punctuation	common

D'aquesta manera detectarem seqüències errònies com aquestes:

pes molecular deduït de 54.023 kDa ***. La*** proteïna va ser purificada a
tècnica de SSO (Sequence Specific Oligonucleotide ***). Tipificació*** de l locus HLA-DPB1 amb
individus heterozigots per a aquesta ***.</item> Les*** delecions fan palesa una
màxim de valor « anticipat ***». És*** evident que en aquestes
Rhodopes and Vrondous mountains ***. Protection*** and management of brown

Però també seqüències correctes com ara:

dos tipus de transmissions lucratives ***.</item> <item>Espanya*** aplica un tipus marginal màxim
la tuberculosi i sense contagi ***.</s> <s>Mantoux*** Negatiu
pero no el riego arterial ***.</s> <s>Snyder*** ha demostrado que el torniquete
sistema operativo central NetWare ***.</s> <s>Novell*** recomienda que se mantenga
vehicles , railways and aircraft ***.</item> <item>research*** in the telematics applications

c) Cal fer una concordança estàndard per als documents en anglès, per comprovar que l'apòstrof no se separa de la paraula que acompanya:

Forma	\'	s
Lema		
POS	punctuation	

D'aquesta manera detectarem seqüències errònies com aquestes:

C. elegans finished , Drosophila ***' s*** near completion , the full
the gp120 protein on HIV ***' s*** surface can be integrated with
referred to as the individual ***' s*** genotype .</s> <s>Some loci (
females .</s> <s>Arabic number % ***' s*** are the risk of mental
removed two-thirds of each rat ***' s*** liver to stimulate liver cell

d) Cal fer una concordança estàndard per als documents en català, per comprovar que l'apòstrof no se separa de la paraula que acompanya:

Forma		‘
Lema		
POS	n. comú	puntuació

D’aquesta manera detectarem seqüències errònies com aquestes:

es dediquen a l‘ ***agricultura ‘*** altres a la ramaderia
fama : La teoria de ***l '*** economia política (1871)
és evident : retards, ***formalitats ‘*** , despeses, que augmenten els

Però també seqüències correctes com ara:

' el millor del ***mercat '*** immediatament , sense despeses desproporcionades
o nous ' de les ***arts '*** les ciències i de tota
' orientació general de la ***contrada '*** el drenatge ha generat per

e) Cal fer una concordança complexa únicament per als documents en català, per comprovar que no quedin apòstrofs al mig d’una paraula:

- a la primera casella:

```
[word=".+'.*"&pos!="N4.*"&pos!="W"&pos!="P"&pos!="D4"&pos!="T"]
```

- a la segona casella:

```
cat Last;
```

D’aquesta manera detectarem seqüències errònies com aquestes:

d ' ells són superiors a ***1'1*** % .</s> <s>
són , entre d ' altres , ***('*** asma i la població del
que ocórrega una cosa o ***1'altra*** dependrà de mecanisme
el primer dels quals mapa ***2'5*** kb. distal a

I com sempre, si es detecten seqüències incorrectes s’ha de fer la correcció de les mostres originals i tornar a processar el document.

8.12.6 Resum

La figura 2 resumeix el conjunt de passos a seguir per completar el processament d'un document del CT.

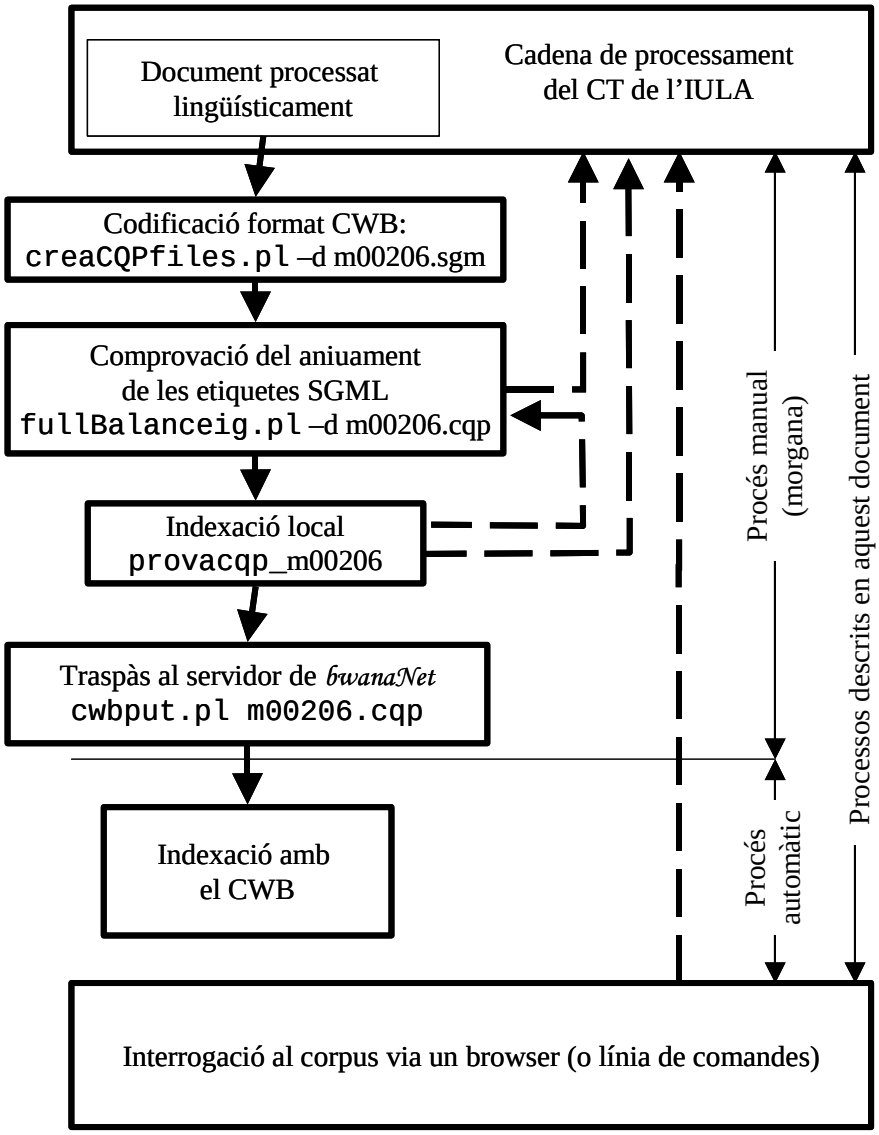


Figura 2. Processament del document m00206.sgm