Marcatge Estructural i morfosintàctic del Corpus Tècnic amb l'estàndard SGML

Vivaldi J., Ll. DeYzaguirre, X. Solé, M.T. Cabré

Sèrie Informes, 1

Barcelona
Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
1996

Direcció de les Publicacions de l'IULA: M. Teresa Cabré

Primera edició: 1996

La Rambla, 30-32
08002 Barcelona

Dipòsit legal: B-34.225-2002

Aquest article s'ha realitzat en el marc del projecte de recerca *Llenguatges especialitzats. Corpus Multilingüe* (CIRIT CS93-4.009).
En l'elaboració d'aquest article, també hi han co_laborat, en l'establiment dels materials base, els investigadors del projecte *Llenguatges especialitzats*: M.T. Cabré (i.p.), J. Vivaldi (coord.), C. Bach, R. Bayá, E. Bernal, L. Borrás, Ll. de Yzaguirre, M. Gonzalez, J. Morel, M. Pujol, R. Saurí, X. Solé, D. Soler, S. Torner, T. Vallés, M. Ribera (documentalista).

Índex

1.- L'estàndard SGML

1.1.- Generalitats.

Tradicionalment en el procés de confecció de documents, llibres i, en general, de tot tipus de material imprès s'han utilitzat marques per identificar els títols i les seccions, per indicar atributs de cert fragments de text (negreta, itàlica, ...), etc. Fins fa ben poc aquestes marques eren introduïdes manualment en l'original per part de l'especialista. Amb l'arribada de la informàtica i la seva massiva difusió durant els anys 80, aquesta forma de treballar ha anat canviant. Ja no és només l'especialista qui posa les marques, sinó fins i tot el mateix autor del text, el qual es veu capaç de fer-ho en la mesura que el seu processador de textos li permet afegir tota mena d'indicacions sobre l'aspecte final del document.

En aquest marc cada processador de textos té el seu sistema i els seus codis per representar internament la informació. Això fa que un document no es pugui llegir amb un altre processador i que s'hagin de fer servir programes de conversió, els quals no sempre són eficaços. A més a més, els ordinadors necessiten d'un mecanisme de gestió interna (sistema operatiu) que imposa certes restriccions al processador de textos que el fan incompatible amb el d'una altra màquina.

Aquesta situació ha generat tot un seguit de problemes: manca de coherència, dificultat per compartir i reutilitzar la informació, etc. Qualsevol que hagi tingut ocasió (o necessitat) de fer servir un text escrit, per exemple, en WordPerfect en un altre processador similar (Word, AmiPro, etc.) o bé de passar aquest text escrit, per exemple, de l'entorn MS-DOS a entorns MacIntosh o UNIX podrà comprovar els problemes que apareixen en la recuperació del text en el lloc de destinació.

Per intentar resoldre aquesta problemàtica es va crear l'estàndard *SGML* (*Standard Generalized Markup Language*) el qual a partir de 1986 ha esdevingut un estàndard internacional (ISO 8879:1986).

A continuació, farem un resum de les característiques més bàsiques d'aquest estàndard. Per aprofundir més en el tema consulteu Bryan (1988), Herwijnen (1993) o bé la descripció detallada de Golfarb (1990).

Un llenguatge de marques no és més que un conjunt de convencions que han de permetre a l'usuari codificar un text tot fent explícita una de les seves interpretacions. Per obtenir aquest resultat un llenguatge de marques ha de permetre les operacions següents:

- definir quines marques es necessiten i la seva posició dintre del text,
- distingir entre marques i text,
- donar significat a cada marca.

L'estàndard SGML té els mecanismes necessaris per satisfer els dos primers requisits, tot permetent, així, la creació de múltiples llenguatges[1] de marques independentment de l'objectiu de l'usuari. La semàntica de cada una de les marques es deixa a la iniciativa de cada usuari en funció de les seves necessitats particulars. Aquesta llibertat, imprescindible per atendre les múltiples necessitats del món real, ha donat origen a diverses iniciatives que han aplicat aquest estàndard a àrees diferents. Les més conegudes són : CALS (*Computer-aided Acquisition and Logistic Support*), TEI (*Text Encoding Initiative*) i HTML (*HyperText Markup Language*).

[1] Donada aquesta capacitat de definir llenguatges de marques, se sol dir que SGML és un metallenguatge.

Cadascuna d'aquestes iniciatives intenta resoldre un o més problemes concrets aportant la definició de les marques necessàries i la seva semàntica associada.

Aquest tipus d'organització permet també una altra característica fonamental: un text marcat en SGML pot ésser tractat per qualsevol aplicació informàtica que tingui en compte aquest estàndard sense conèixer la semàntica de les marques[2]. L'exemple més clar i estès és la xarxa mundial Internet i, en particular, l'accés amb el WWW (*World Wide Web*). En accedir a aquesta xarxa es pot comprovar la facilitat per obtenir qualsevol text sense conèixer el tipus de plataforma que el conté.

Els conceptes bàsics que fa servir l'estàndard per aconseguir els objectius fixats són els següents:

- entitats,
- elements i atributs associats,
- tipus de documents.

1.2. Entitats.

Una entitat SGML és simplement una cadena de caràcters amb un nom associat. Aquesta seqüència pot tenir qualsevol forma: des d'una o més paraules fins a un text complex amb una estructura incorporada.

Aquest mecanisme pot ésser utilitzat de diferents maneres dintre d'un document SGML. La més simple és la representació de caràcters i símbols especials, com poden ésser: ç, ñ, à, α, æ, «, etc. És ben coneguda la dificultat que presenten aquests símbols en el moment de convertir un text d'un entorn a un altre, o bé d'enviar un text per correu electrònic. El problema consisteix bàsicament que la informàtica s'ha desenvolupat a partir d'un conjunt de caràcters tancat (codis ASCII) quan en realitat aquests configuren un conjunt obert. Cada llengua té les seves especificitats, les quals no sempre han estat enteses i reconegudes.

La solució proposada per SGML[3], en el cas de les llengües occidentals, és codificar aquests símbols amb entitats i utilitzant caràcters ASCII que no comportin problemes (0-128). Per exemple, el caràcter **ç** es codifica com '**ç**'; aquesta seqüència de caràcters es tradueix dins cada entorn de treball en la seqüència de *bits* necessària per a la correcta representació. Existeix un conjunt molt ampli de símbols codificats segons aquests principis i la facilitat per definir-ne de nous és clara.

En qualsevol cas, les entitats poden contenir també seqüències de text molt complexes que poden incloure text, marques i altres entitats. De tota manera, el mecanisme és sempre el mateix : quan el programa SGML troba una entitat, la substitueix pel contingut que té associat.

---

[2] Un exemple és el *parser nsgmls* que utilitzem al projecte de Corpus Tècnic Especialitzat (CT) per comprovar el marcatge dels textos. Aquest analitzador és utilitzat en molt projectes arreu del món per analitzar tot tipus de textos. Vegeu l'apartat "Eines SGML" a l'annex VII per obtenir mes informació d'aquest analitzador.

[3] Hi ha una proposta d'estàndard ISO, que es denomina UNICODE, que tot i utilitzar una codificació amb quatre *bytes* vol ésser universal. Actualment es troba en fase de discussió.

La distinció entre text i entitat s'obté fent referència a una sintaxi concreta. En aquest cas s'ha utilitzat el símbol "**&**" per indicar l'inici d'una entitat i el símbol "**;**" per indicar l'acabament. Aquests són els símbols més usuals, però, en cada insta_lació, SGML permet definir-ne d'altres. Aquesta concreció d'aspectes sintàctics com l'esmentat es coneix també com a **sintaxi concreta**.

1.3. Elements i atributs.

Els textos no només es poden veure com una seqüència de caràcters, sinó també com una estructura composta per capítols, títols, paràgrafs, noms, dates, etc. que tot llenguatge de marques intenta representar. Cadascun d'aquests components està caracteritzat per un punt d'inici, un punt de final, el tipus de component i el context en el qual pot aparèixer. SGML resol totes aquestes qüestions amb la definició d'elements i atributs. Els elements s'identifiquen amb etiquetes (*tags*), que no són altra cosa que un nom d'identificació i que, a la vegada, indiquen els punts d'inici i final dels components del text.

Com en el cas de les entitats, és necessari distingir entre dades i elements. Per tant, hem de referir-nos una vegada més a una sintaxi concreta. El més freqüent és utilitzar el símbol "**<**" per indicar l'inici d'una etiqueta d'obertura, el símbol "**</**" per indicar l'inici d'una etiqueta de tancament i, finalment, el símbol "**>**" per indicar el tancament de qualsevol etiqueta. Així, per exemple, en el següent fragment català:

Van demanar un habeas corpus

volem indicar la presència de paraules en un altre idioma (*habeas corpus*). Aquest fet es marca de la següent manera :

Van demanar un <foreign>habeas corpus</foreign>

Però també pot ésser necessari afegir informació addicional associada amb una etiqueta com podria ésser, en l'exemple anterior, l'idioma. Aquesta informació, utilitzant l'estàndard SGML, es reflecteix així :

Van demanar un <foreign lang=LA>habeas corpus</foreign>

on *lang* és el nom d'un atribut associat a l'element *foreign* que té com a valor *LA*. Aquesta informació ens indica que el fragment de text emmarcat per l'element *foreign* és una seqüència de paraules en llatí. Així, doncs, els atributs permeten afegir informació addicional no inclosa explícitament en el text original pero que pot ser d'utilitat conservar[4].

[4] En el cas de l'exemple aqueste informació permet saber que llengues s'han utilitzat en el text, informació que pot ser necessaria en un process posterior, p.ex. un analizador morfològic.

1.4. Definició de Tipus de Document.

La definició d'entitats, elements i atributs és suficient per codificar un text, però per mantenir la consistència del conjunt és necessari conèixer el nom exacte dels elements i atributs permesos així com el context i la seqüència en què poden aparèixer. Per definir tota aquesta informació SGML es recolza en un document addicional anomenat **DTD** (*Document Type Definition*).

Aquest document compleix una funció semblant a la d'una gramàtica del llenguatge natural, és a dir, defineix quines són les construccions permeses en un llenguatge de marques concret. A partir d'una especificació genèrica de document, podem definir quins són els elements permesos, i constituir, així, el que es coneix com a **model de contingut** *(content model*). Aquest tipus d'organització permet, en fase de marcatge, la comprovació de les marques afegidas i en la fase d'explotació facilita la reutilització de blocs de text en diferent documents.

Per exemple, podríem definir un document com un conjunt de capítols, cadascun del quals tindria un títol i un o més paràgrafs. Una DTD hipotètica que reflecteixi aquestes regles és la següent:

```
< !ELEMENT      doc             - -     (chapter+)      >
< !ELEMENT      chapter - -     (title, p+)     >
< !ELEMENT      (title,p) - -   (#PCDATA)   >
< !ATTLIST      title
            type    (main | sub)    main                    >
```

Tanmateix, no és necessari que tot usuari de SGML hagi de definir la seva pròpia DTD. Iniciatives com les esmentades anteriorment (CALS, TEI, etc.) han definit DTDs per emprar en molts casos reals. Només quan les DTDs ofertes per aquestes organitzacions no s'adeqüin a les necessitats de l'usuari, s'ha de pensar a desenvolupar una DTD pròpia. Una vegada obtinguda la DTD que es necessita per codificar un tipus específic de text, pot ésser utilitzada un nombre indefinit de vegades. Els documents obtinguts d'aquesta manera poden fer front al pas del temps raonablement bé, cosa que no es pot garantir amb els processadors de text tradicionals.

1.5. SGML: avantatges i inconvenients.

Els avantatges que genèricament s'obtenen utilitzant aquest estàndard són nombrosos:

- independència dels programes i plataformes utilitzats,
- portabilitat,
- reutilització de tot/part dels documents per molts tipus de presentació,
- acceptació de l'estàndard per grans organitzacions arreu del món i en avanç manifest.

Evidentment, tot avantatge presenta un cost que és necessari assumir i, així, ens trobem amb els següents inconvenients:

- es requereix un estudi aprofundit sobre les necessitats globals de l'organització en tot el que es refereix a l'estructura dels seus documents[5],
- es necessiten eines específiques que encara no tenen una difusió massiva i, per tant, resulten costoses,
- és necessari que les eines existents evolucionin per tal de facilitar-ne l'ús final,
- és inevitable l'augment de la grandària dels fitxers[6].

1.6. SGML i el marcatge de corpus textuals.

Per altra banda, en el camp de la lingüística s'ha fet cada vegada més palesa la necessitat d'emprar corpus voluminosos per conèixer amb més precisió i fonament el comportament real del llenguatge. Per tal d'obtenir el material en format electrònic és habitual recórrer a les fonts, és a dir, a editorials, empreses, universitats i, en general, qualsevol organització que produeixi el tipus de material necessari per al corpus. Com que es probable que aquest material no ens arribi en format SGML sinó en un format intern diferent segons cada organització es convenient d'efectuar una normalització, molt sovint feta a mida, que uniformitzi el sistema de marques.

És evident que l'obtenció i consegüent preparació d'un corpus per a la seva explotació esdevé un procés costós en temps i recursos, motiu pel qual és força atractiu i gairebé necessari contemplar la possibilitat de compartir aquest important material de recerca amb altres organitzacions. A més a més, tenint present el cost que representa desenvolupar els programes d'explotació, és molt desitjable poder utilitzar/adaptar totalment o parcialment aplicacions públiques o estàndards.

Un altre problema, no menyspreable, és la formació que hem de donar a qui necessiti manipular les dades de prop. Si fem servir un únic sistema de marcatge, que a més a més es estándard, aquest problema se simplifica notablement.

Coneguda aquesta problemàtica, és fàcil comprendre la decisió de la Unió Europea d'afavorir la utilització de l'estàndard SGML per als seus projectes de constitució de recursos lingüístics per a la recerca.

Dins d'aquest context, a l'Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra hem adoptat aquest estàndard per al marcatge de tots els textos que formaran part del projecte de **Corpus Tècnic Especialitzat** (CT)[7]. Aquesta aplicació no es limita al marcatge estructural sinó que s'estén al marcatge morfosintàctic i sintàctic, alineament de textos para_lels, etc., tot seguint de prop les indicacions donades pel **CES** (*Corpus Encoding Standard*) dintre de la iniciativa **EAGLES** (*Expert Advisory Group on Language Engineering Standards*). A continuació, aportem una explicació i una justificació de les característiques més rellevants del sistema de marcatge adoptat.

---

[5] L'única manera de garantir la intercanviabilitat de documents és portant l'estandardització al nivell semàntic, és a dir, no només tenint les mateixes regles sintàctiques sinó també donant el mateix sentit a les marques.

[6] L'augment en relació al fitxer ASCII corresponent es pot estimar entre 2 i 10 vegades aproximadament (depèn de la densitat de marques). Amb fitxers de processadors de text l'increment és més difícil d'avaluar però sempre serà inferior al fitxer ASCII corresponent.

[7] Projecte finançat per la CIRIT (codi de refèrencia CS93-4.009).

2.- Marcatge de corpus.

L'objectiu bàsic de tenir un corpus codificat és descriure totes les peces que són lingüísticament rellevants per a la seva anàlisi. En qualsevol text escrit podem distingir objectes:

- d'estructura: les seqüències de text que s'organitzen en forma de frases, paràgrafs, capítols, seccions, etc.
- lingüístics: sintagmes, paraules, morfemes, noms propis, dates, etc.

En codificar un text es fa explícita l'existència d'aquests objectes mitjançant les marques que es defineixin per a cada tipus d'informació. Per realitzar aquest procés de marcatge s'ha de recórrer a eines automàtiques o semiautomàtiques que, a partir de la informació ja inclosa en el text, puguin obtenir la codificació necessària. El grau d'intervenció humana dependrà del tipus de marques que es vulguin incorporar així com del tipus i format del text que es tracti. Malauradament, quan es parteix de textos poc o gens marcats la recuperació d'informació estructural es fa molt difícil. Aquest és el cas dels textos que s'han obtingut en format ASCII o que provenen de processadors de text on l'autor ha marcat molt poc el text (cosa que succeeix molt sovint).

Per a l'obtenció de la informació lingüística és imprescindible recórrer a eines automàtiques: preprocessadors i etiquetadors (de base lingüística o estadística) que poden donar aquesta informació amb un grau d'encert entre el 90 i el 95%.

També hi ha un altre tipus d'informació que pot ésser de gran interès des del punt de vista lingüístic: ens referim a la problemàtica específica de l'alineació de textos paral·lels. Quan es disposa d'un mateix text en més d'una llengua, és de gran utilitat reflectir en un fitxer específic la relació existent entre el mateix text en llengües diferents. Aquesta informació pot ésser de gran utilitat a estudiants i professionals de la Traducció, així com per construir corpus o validar glossaris terminològics ja construïts. En aquest cas també s'ha de recórrer a eines que facilitin l'obtenció de correspondència entre elements, en particular, les frases i paràgrafs.

Una vegada presa la decisió de codificar els textos del CT d'acord amb l'estàndard SGML, el següent punt que havíem de considerar era com obtenir una *gramàtica* (o DTD) que servís per verificar l'aspecte sintàctic del marcatge (lingüístic i estructural) i que s'hagués dissenyat específicament per a aquest tipus de codificació. En el moment de triar disposàvem de la iniciativa TEI i de la proposta CES, la primera de caràcter general i ja consolidada, i la segona encara en una fase embrionària però específicament dissenyada per a la codificació de corpus textuals. En l'actualitat, la iniciativa EAGLES ha publicat una primera versió de la DTD per al marcatge de corpus que l'IULA ha fet seva per iniciar el seu procés de codificació. Presentem ara molt breument les característiques més rellevants de la iniciativa TEI per aprofundir més endavant en la proposta d'EAGLES per a la codificació de Corpus textuals i en les modificacions introduïdes per l'equip de treball de l'IULA.

2.1. Marcatge estructural.

2.1.1 Proposta TEI.1.1 Proposta TEI.

La Iniciativa per a la Codificació de Textos (TEI- Text Enconding Initiative) és patrocinada per diverses associacions internacionals i està dirigida bàsicament a la codificació d'estructures complexes de text per a la investigació lingüística i literària. Els tipus de text que es contemplen són prosa, poesia, teatre, textos orals, terminològics i lexicogràfics; s'inclouen també conjunts

d'etiquetes addicionals per tractar formes de representació especials com, per exemple, transcripcions, crítiques, enllaços, arbres, gràfics, figures, estructures de trets, *corpus*, fórmules, etc.

L'objectiu bàsic que s'havien fixat els patrocinadors d'aquest projecte era crear un sistema de codificació[8]:

- adequat a la investigació,
- simple i clar,
- de fàcil utilització,
- que permetés una definició rigorosa i un processament eficient dels textos,
- extensible i
- que fes ús d'estàndards existents o emergents.

Aquestes característiques que, en principi, semblen fàcils d'assolir, portades a la pràctica, han donat com a resultat un sistema relativament complex i costós, especialment quan es tracta d'aplicar-lo a la creació de corpus textuals per a la recerca lingüística. Un dels motius que ha contribuït a crear aquesta situació és la gran varietat de tipus de text que es pretén codificar amb aquesta iniciativa : des de la prosa fins a diccionaris[9], transcripcions de llengua parlada o definicions d'elements addicionals com hipertextos, estructures de trets, taules, fórmules, gràfics, etc. La seva complexitat ha donat origen recentment a una versió reduïda anomenada TEI-Lite.

Malgrat la declaració explícita d'utilitzar només un sistema de marques per a cada fenomen, en la pràctica es permet descriure una determinada situació de diverses maneres. Per exemple : la codificació de les divisions d'un document es pot fer amb una etiqueta genèrica que pot incrustar-se recursivament o distingint diferents nivells (<div1>, <div2>, ... , <div7>). La codificació de seqüències de text amb característiques (idioma, presentació, etc.) que van més enllà del límit natural d'una frase representen un problema anàleg. Evidentment una organització que apliqui una codificació d'aquest tipus ha d'escollir una única manera d'assenyalar un fenomen i ha d'ésser coherent amb aquesta.

De tota manera, el material proposat per la iniciativa TEI és pres pràcticament sempre com una referència obligada en la codificació de qualsevol tipus de text.

2.1.2 Proposta EAGLES.

Com ja s'ha mencionat, la Unió Europea, a través de la iniciativa EAGLES i, en particular, de la CES, ha definit un conjunt de DTDs aptes per a la codificació de corpus textuals. Aquestes DTDs prenen com a punt de referència l'aportació de la TEI, tot i que en simplifiquen l'esquema general i tenen present la problemàtica específica de textos que no han estat creats amb l'estàndard SGML (en anglès: *legacy texts*). També es reconeix l'especificitat de l'objectiu últim dels textos: la recerca lingüística.

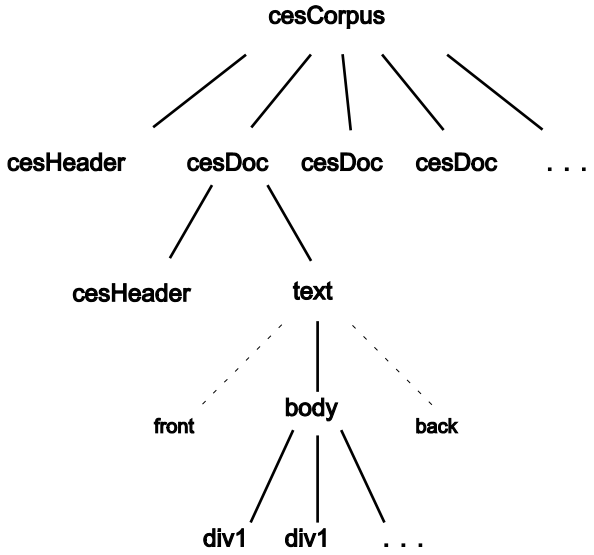
Aquestes consideracions han portat a simplificar notablement el nombre d'elements i atributs i a modificar notablement l'estructura de la DTD. La TEI-DTD preveu una estructura

[8] Per a una descripció exhaustiva d'aquests punts vegeu "*The Design of the TEI Encoding Scheme*" in *Computers and Humanities, Volume 29 nº 1*.

[9] La prosa és un tipus de text lineal mentre que els diccionaris tenen una estructura interna molt precisa i complexa que es pot comparar a la d'una base de dades.

modular que es munta en cada document a partir dels mòduls necessaris per a un text en concret. La CES-DTD, en canvi, preveu una DTD per a cada tipus de document (text, marcatge morfosintàctic i alineació). Cadascun d'aquests tipus de documents es relacionen amb els altres mitjançant enllaços i referències.

L'estructura bàsica d'aquesta proposta preveu dos components principals: el corpus (<cesCorpus>) i els documents que conté (<cesDoc>). Cadascun d'aquests elements té associat un element amb informació que descriu el seu contingut des del punt de vista bibliogràfic i que es denomina **capçalera** (vegeu capítol 4). L'estructura global resultant és la següent[10]:



Per al marcatge estructural la CES defineix una DTD específica (cesDoc-DTD) amb diferents nivells de codificació estructural. El nivell més simple (*level 1*) incorpora només les estructures bàsiques fins al nivell de paràgraf, mentre que el més alt (*level 3*) requereix marcar elements com ara dates, números, paraules estrangeres, etc. dintre de la frase. En els documents del CT de l'IULA pretenem arribar fins a aquest tercer nivell de marcatge.

La DTD en qüestió s'aplica a tots els textos, independentment del seu nivell de codificació. En aquesta DTD s'especifiquen com a característiques més rellevants quins són els elements permesos, en quina seqüència es poden trobar i quina és la seva estructuració.

Tot i que la cesDoc-DTD sigui una gramàtica molt més apropiada a les necessitats reals d'un Corpus que la proposada per la iniciativa TEI, hem cregut necessari fer alguns canvis i modificacions, les raons dels quals, sens dubte, es troben en l'extrema joventut d'aquesta DTD i en les afirmacions dels mateixos autors quan declaren que és un esborrany (*draft*) i no una versió definitiva.

Presentem ara els elements més importants utilitzats en els documents del CT i les

[10] En línia discontínua s'indica la part de l'estructura bàsica d'un document que proposa la iniciativa TEI i que posteriorment desapareix en el disseny del CES. És evident que un del nivells (*text* o *body*) és redundant, però s'ha decidit de mantenir-la perquè això facilita una eventual migració cap al format TEI.

modificacions realitzades des de l'IULA, així com la seva justificació.

2.1.3 Entitats.

A fi d'obtenir la màxima claredat es defineixen un conjunt d'entitats que s'han d'utilitzar en molts punts de la cesDoc-DTD. Aquestes entitats es refereixen a:

- un conjunt bàsic d'atributs que s'apliquen a la majoria dels elements. En aquest cas el codificador pot assignar valors als següents atributs : *id* (identificació única), *n* (identificació no única), *lang* (codi d'idioma, segons les normes ISO), *wsd* (codificació ISO del caràcters especials) i *rend* (presentació tipogràfica en la versió original).
- classes d'elements: agrupació d'elements que poden aparèixer al mateix nivell del paràgraf (llistes, notes, taules, etc.) o bé dintre d'un paràgraf (text, dates, números, termes, noms propis, paraules en una altra llengua, abreviacions, etc.).
- conjunts d'elements que sovint apareixen com a components d'altres elements. L'entitat *par.seq* n'és un exemple, ja que agrupa taules, gràfics i notes que apareixen sempre al mateix nivell dels paràgrafs.

2.1.4 Divisions.

Cada instància de document SGML es defineix amb l'element *cesDoc*. Aquest element utilitza atributs específics per definir el tipus de document (*type*), la DTD que verifica la seva codificació (*version*) i la localització de la capçalera del document (*header.loc*).

En una instància de document SGML han d'aparèixer dos elements: el primer és el *docHead*, que conté una descripció breu del document, i el segon és el *body*, que agrupa el document sencer. En aquest últim element trobarem les divisions al nivell més alt.

Els textos en general s'agrupen sota algun tipus d'estructura, a semblança d'un arbre. Així, per exemple, un llibre es divideix en capítols, els capítols en seccions, les seccions en subseccions, etc. Donat que els textos poden tenir estructures molt complexes i variades, hem decidit codificar-ne l'estructura amb una combinació d'etiquetes i atributs a fi d'evitar una multiplicitat de noms. Cadascuna d'aquestes agrupacions es codifica amb un element, anomenat *div*, més un número que indica la profunditat de la divisió. D'aquesta manera, el llibre del nostre exemple tindria capítols que es marquen amb l'etiqueta *div1*, seccions amb l'etiqueta *div2* i, així, successivament. La DTD preveu identificar fins a quatre nivells. A més, com ja s'ha mencionat, cada element corresponent a una divisió té associat un atribut *type* per indicar-ne el tipus de divisió de què es tracta (ex. type=capítol).

La pràctica ha demostrat que el nombre de divisions necessària per codificar un text de dret va més enllà dels quatre nivells que preveu la cesDoc-DTD. Hem detectat que la necessitat mínima en textos de dret és de nou nivells de divisió, la qual cosa comporta tenir nou etiquetes diferents per indicar un canvi de divisió (<div1>, <div2>,... ,<div9>). Aquesta, evidentment, no és l'única solució perquè el número associat a cada nivell és redundant; l'existència d'un nivell inferior pot inferir-se del fet de tenir una etiqueta incrustada dintre de l'altra. La solució més apropiada i definitiva, ja prevista per la TEI-DTD, és eliminar el número associat a cada nivell de divisió tot deixant una etiqueta genèrica ** ** amb un atribut *type* per indicar el tipus de divisió. D'aquesta manera, la DTD permet qualsevol nivell d'incrustació.

Per raons pràctiques (facilita la comprovació manual), és convenient mantenir

l’estructura de les divisions numerades. El pas cap a una estructura de divisions no numerada és molt simple però el camí invers és molt costós.

Codificació original CES

```
<div1 type=titol>
    <div2 type=llibre >
        <div3 type=article n=1>
            . . .
        </div3>
        <div3 type=article n=2>
            . . .
        </div3>
        . . .
    </div2>
</div1>
```

Codificació alternativa

```
<div type=titol>
   <div type=llibre>
        <div type=article n=1>
            . . .
        </div>
        <div type=article n=2>
            . . .
        </div>
    . . .
  </div>
</div>
```

Mentre que la cesDoc preveu per a les marques de divisions el tret *type*, no s'ha definit cap llista amb els valors concrets que pot tenir. La naturalesa específica dels textos de l'IULA ha permès definir una sèrie molt àmplia de tipus de divisió. A l’Annex I es troben els tipus de text que s'han definit pel domini del dret.

Dintre de cada divisió, en qualsevol nivell, es troben diversos tipus d'encapçalament que introdueixen els paràgrafs i altres elements que s'ha previst que estiguin al seu mateix nivell: llistes, taules, figures, notes, etc.

2.1.5 Llistes.

Per al marcatge de llistes es preveu un element *list* que les englobi i que incorpora un element opcional per al títol de la llista (<*head*>) seguit d'un o més ítems (<*item*>). Si la llista incorpora un element d'enumeració (ex. **a.**), hem preferit codificar-lo com un identificador no exclusiu a través de l'atribut global n. Per exemple en el fragment següent:

> Els contractes a que es refereix la regla anterior que no tinguin caràcter administratiu, perquè no estiguin inclosos en els supòsits que s'hi preveuen, es regiran: A)Quant a la seva preparació i adjudicació, per les seves normes administratives especials, i, quan no n'hi hagi, per les disposicions d'aquesta Lleisobrepreparació i adjudicació dels contractes d'obres, gestió de serveis i subministraments, que s'aplicaran per analogia amb la figura contractual de que es tracti. B)Quant als seus efectes i extinció, per les normes del Dret privat que els siguin aplicables en cada cas, a falta de les seves normes especials, si n'hi hagués.

es codifica com s'indica a continuació:

<s>Els contractes a que es refereix la regla anterior que no tinguin caràcter administratiu, perquè no estiguin inclosos en els supòsits que s'hi preveuen, es regiran:<list><item n='A'>Quant a la seva preparació i adjudicació, per les seves normes administratives especials, i, quan no n'hi hagi, per les disposicions d'aquesta Lleisobrepreparació i adjudicació dels contractes d'obres, gestió de serveis i subministraments, que s'aplicaran per analogia amb la figura contractual de que es tracti.</item><item n=B>Quant als seus efectes i extinció, per les normes del Dret privat que els siguin aplicables en cada cas, a falta de les seves normes especials, si n'hi hagués.</item></list></s>

El tractament que preveu la cesDoc per a aquest tipus d'estructura és decididament molt simple: les llistes només poden existir al mateix nivell de paràgraf o bé incrustar-se totalment dintre de cada ítem. La casuística trobada en documents reals és molt més variada i fa necessària la incorporació de modificacions importants.

Totes les llistes, que tenen en comú una frase introductòria i una sèrie d'ítems, poden diferenciar-se segons el lligam lingüístic entre aquestes dues parts. De fet, s'han observat dos casos possibles:

- aquelles en què el segment introductori és una frase completa (subjecte-verb-objecte) que va seguida d'una enumeració d'ítems (*llista suau*) o
- aquelles en què la frase del fragment introductori es completa en cadascun dels ítems (*llista abrupta*).

La situació es complica encara més si tenim en compte que les llistes poden incrustar-se recursivament, és a dir, que un ítem pot contenir una llista. A més a més, molt sovint es barregen els tipus de llistes que s'incrusten. Per a una discussió més a fons, vegeu Solé & Saurí (1996).

Per intentar reflectir aquesta situació s'ha modificat la cesDoc-DTD convenientment. A més a més, i tenint present que el marcatge de llistes és de difícil automatització, hem deixat oberta la possibilitat de no marcar les suaus que continguin ítems oracionals.

2.1.6 Figures.

L'element *figure* es fa servir per indicar la presència d'una figura o gràfic en el text original. Pot contenir un conjunt de marques opcionals: <*head*> per indicar el títol o capçalera , per a comentaris, <*figDesc*> per descriure l'aspecte o el contingut de la figura i <*body*> per incloure el gràfic.

Si la figura és un fragment de text aquesta marca es pot incloure amb l'element opcional *body* o bé amb l'atribut *entity* de l'element *figure*, el qual indica el nom de l'entitat externa que conté la figura.

En el projecte CT hem utilitzat l'element *figure* només per indicar la presència d'una figura. En cas que la figura compti amb un títol disponible en el format d'entrada, aquest es consigna a través de l'element *head*. Quan en el text només s'indica la presència de la figura sense cap altra informació (p.ex.: [Figura 4]) el codificador pot optar per no codificar-la i esborrar la seva referència.

El següent fragment:

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

es marca com s'indica a continuació:

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

2.1.7 Notes.

En aquest context podríem definir una nota com un fragment de text que es troba fora del cos general del text però que s'hi vincula a través d'una crida, normalment numèrica, que relaciona el fragment en qüestió amb un dels fragments del cos general. Per reflectir aquest caràcter singular de les notes utilitzem dos elements: el primer (*ptr*) fa palès el lloc des d'on es fa la crida i el segon, (*note*), deixa constància de la incorporació efectiva de la nota en el cos textual. Ambdós elements estan lligats unidireccionalment mitjançant un identificador comú.

L'aparició en el text d'aquest tipus d'informació (peu de pàgina, a final del capítol/llibre, etc) es codifica sempre amb l'element *note*. L'atribut *place* permet codificar el lloc d'aparició i l'atribut global *n* el número d'ordre (si n'hi ha). Per exemple, el fragment següent:

1.- La relació de serveis i els contractes sobre personal regulats a la legislació sobre funcionaris, i, en el seu cas, a la laboral.[3]
2.- . . .

. . .

[3] Tingueu en compte l'article 7.è de la Llei de Funcionaris Civils de l'Estat, text articulat aprovat per Decret 315/1964, de 7 de febrer.

es codifica com s'indica a continuació:

<list><item n=1>La relació de serveis i els contractes sobre personal regulats a la legislació sobre funcionaris, i, en el seu cas, a la laboral.<ptr target='d7m2n3'></item>
<item n=2> . . .

<note place=foot n=3 id='d7m2n3'> <s>Tingueu en compte l'article 7.è de la Llei de Funcionaris Civils de l'Estat, text articulat aprovat per Decret 315/1964, de 7 de febrer.</s> </note>

Tot i que el tractament donat per la cesDoc-DTD a aquest tipus d'element és encertat, ens manca la possibilitat d'incorporar dintre d'una nota i al mateix nivell de paràgraf, els elements per a llistes, figures i taules, per la qual cosa hem pres l'opció de modificar l'esmentada DTD.

2.1.8 Taules.

L'element *table* permet reflectir l'existència de text amb format tabular. El contingut es pot expressar només en el format de línia a línia i dintre de cada línia, de ce_la a ce_la. Amb els atributs *rows* i *cols* s'indiquen respectivament el nombre de files i columnes. L'element en qüestió només podrà contenir elements *fila* (*row*); al seu torn, l'element *fila* només podrà contenir elements tipus *ce_la* (*cell*) o bé una altra taula incrustada.

Normalment l'aparició d'una taula porta associada poca o gens d'informació lingüística, ja que la major part sol constar de xifres. Per aquest motiu la persona que insereix las marques pot optar entre una codificació completa o una codificació mínima. En aquest darrer cas s'incorpora la taula amb el títol i tota la resta es marca amb una única etiqueta de ce_la. Només es fa un marcatge detallat quan el seu contingut lingüístic es considera rellevant. Per exemple, el fragment següent:

p	q	p V q
V	V	V
V	F	V
F	V	V
F	F	F

Taula 1. Condicions de veritat del disjuntor

es codifica com s'indica a continuació:

```
<table n=1>
    <head type=main>Condicions de veritat del disjuntor</head>
</table>
```

2.1.9 Fragments amb característiques de presentació especials.

Quan es codifica text heretat, és a dir, text que no ha estat editat amb un processador de textos SGML, és freqüent l'aparició de fragments amb indicacions tipogràfiques (negreta, itàlica, subratllat, etc.) que configuren el mode en què l'autor vol ressaltar un fragment de text.
Quan aquest fragment és simplement una sèrie de paraules dintre de la mateixa frase utilitzem l'element *hi*; en canvi, si la característica a representar s'aplica a tot un element, emprem l'atribut global *rend* de l'element en qüestió. El fragment següent:

```
En aquest cas el mot *replanteig* indica comprovaci&oacute; de la viabilitat material.
```

quedaria marcat com s'indica tot seguit:

```
<p><s>En aquest cas el mot <hi rend=il>replanteig</hi> indica comprovaci&oacute; de la viabilitat material.</s></p>
```

2.1.10 Elements lingüístics.

Per a la detecció i marcatge d'aquest tipus d'elements és molt freqüent recórrer a eines específiques, com ara els analitzadors morfològics i els desambiguadors. Amb tot, també és habitual potenciar el rendiment i l'eficàcia d'aquestes eines amb programes de preprocés que detecten tots aquells elements rellevants lingüísticament, com ara són els noms propis, xifres, dates, abreviatures, etc. Tots aquests fragments es poden detectar automàticament amb un marge d'error relativament petit i tenen associada una etiqueta específica[11].

El projecte del CT incorpora un programa de preprocés, tal com suggereix la iniciativa EAGLES-CES[12]. Tanmateix, els recursos disponibles a l'IULA permeten afegir un tipus de dades addicional: les locucions. Aquest tipus d'element està contingut en una base de dades específica obtinguda directament d'un diccionari en format electrònic.

El fragment:

> Fixar les directrius dels preus dels béns, els productes i els serveis obtinguts com a resultat dels treballs interns, quan aquesta facultat li sigui delegada pel conseller d'acord amb l'article 7, paràgraf 2, de la Llei 4/1985, de 29 de març[22].

es marca de la següent manera:

```
<p><s>Fixar les directrius dels preus dels b&amp;eacute;ns, els productes i els serveis obtinguts com a resultat dels treballs interns, quan aquesta facultat li sigui delegada pel conseller <loc pos='P'>d'acord amb</loc> l'article <num>7</num>, par&amp;agrave;graf <num>2</num>, de la <name>Llei</name> <num>4/1985</num>, de <date>29 de mar&amp;ccedil;</date><ptr target='d130c2n22'> .</s></p>
```

2.2. Marcatge morfosintàctic.

La necessitat d'afegir informació morfosintàctica als textos d'un corpus és clara. Des d'un bon començament, *corpus* com Brown (Kucera i Francis, 1967) o LOB, pioners en aquest tipus de recerques, han afegit marques morfosintàctiques. Un text amb aquesta informació afegida permet recerques molt més complexes i útils.

Tradicionalment aquesta informació s'ha afegit associant a cada paraula un codi que representa l'única anàlisi possible per a aquesta peça lèxica. Per exemple, la frase:

The victim's friends told police that Krueger drove into the quarry and never surfaced

pot ésser codificada com s'indica a continuació :

The$AT victim$NN1's$GEN friends$NN1 told$VVD police$NN2 that$CST Krueger$NP1 drove$VVD into$I the$AT quarry$NN1 and$CC never$R surfaced$VVD

[11] Vegeu Vivaldi et al (1997) per a l'explicació d'aquesta eina.

[12] Vegeu M. Pujol et al (1996) per a l'explicació d'aquesta eina.

2.2.1 Utilització de l'estàndard SGML.

Amb aquest tipus de codificació s'afegeix al text original tota la informació necessària per saber quina és la seqüència de categories morfosintàctiques. De tota manera, cal tenir present les limitacions que comporta:

- el nivell de detall està limitat pel sistema de codificació i qualsevol canvi en aquest sistema afecta tot el conjunt i
- no té sentit associar un codi només a una paraula quan es marquen seqüències de mots amb estructures més complexes que la paraula (les locucions, termes multimot, paraules compostes, sintagmes, etc.).

Una de les solucions possibles d'aquest problema és identificar i marcar els límits de les paraules tot afegint-hi després la informació morfosintàctica en forma d'elements i atributs. Trobarem una descripció detallada d'aquesta possibilitat a *Computers and Humanities, Volume* 29 No 1.

La codificació de la paraula *roda* en la interpretació com a nom és la següent :

```
<w type=token>
    <str>roda</str>
    <lema pos='N5FS' status='+'>roda</lema>
    <lema pos='VDR2S' status='-'>roda</lema>
</w>
```

Aquest sistema dóna més informació que no pas el tradicional. La codificació de la categoria ve donada per l'atribut *pos* (*part of speech*). Malgrat tot, donada l'ambigüitat d'aquesta paraula (nom comú i verb), és necessari indicar la interpretació vàlida en cada context d'aparició. Aquesta informació s'indica amb l'atribut *status*, que pot tenir com a valor un '+' (vàlid) o un '-' (descartat en aquest context).

Cadascuna de les possibles categories té associada una estructura de trets que dóna la informació morfosintàctica corresponent. Per exemple, l'estructura de trets que es podria donar per a un nom comú femení singular seria la següent:

```
<fsLib>
 . . .
<fs id='N5FS'>
    <f name=cat><sym value=nom></f>
    <f name=type><sym value=com></f>
    <f name=agreement>
    <fs>
        <f name=gen><sym value=fem></f>
        <f name=num><sym value=sing></f>
    </fs>
</fs>
 . . .
</fsLib>
```

L'associació entre els dos conjunts d'informació ve donada pel fet que els valor dels atributs *pos* (element lema) i *id* (element fs) són els mateixos (*N5FS*). De la mateixa manera que s'ha codificat una paraula aïllada es podrien codificar seqüències més complexes com ara:

- una data:

```
<w type=date>
        <str>25 de maig de 1810</str>
        <lema pos='D' status='+'>1810/05/25</lema>
</w>
```

- un nom propi:

```
<w type=plex>
        <str>Institut Universitari de Ling&uuml;&iacute;stica Aplicada</str>
        <lema pos='N4MS' status='+'>=</lema>
</w>
```

2.2.2 Proposta EAGLES.

Aquest tipus de representació suposa una millora considerable en relació als primers sistemes ja que permet codificar qualsevol tipus d'estructura, però té un punt feble relativament important: les marques s'han d'incorporar directament sobre les dades primàries. Això ens priva d'aplicar diferents esquemes de codificació a un mateix text sense haver de duplicar-ne l'original.

El sistema escollit per la iniciativa EAGLES-CES preveu una separació absoluta entre el text original (independentment del nivell de marcatge que tingui) i les anotacions que s'hi facin. Per aconseguir-ho les anotacions s'incorporen a un document SGML lligat a les dades originals mitjançant enllaços unidireccionals, els quals es basen en una variació del mecanisme anomenat localitzador (*locators*), definit per la iniciativa TEI i per l'estàndard ISO/IEC DIS 10744:1992 HyTime (*Hypermedia/Time based Document Structuring Language*). D'aquesta manera, diferents usuaris poden aplicar diversos marcatges utilitzant el mateix text.

En qualsevol cas, els autors de la proposta CES són conscients que en aquesta solució, ja de per si tècnicament vàlida, queden alguns detalls a perfilar i per això consideren que aquesta part de la proposta en qüestió és encara un esborrany. El caràcter no definitiu d'aquesta part de l'estàndard fa que la implementació que descrivim pugui quedar afectada per possibles canvis.

2.2.3 Localitzadors.

A fi de resoldre la dificultat de separar les dades de les marques que els apliquem és imprescindible trobar un mecanisme que permeti relacionar el contingut de dos fitxers: un fitxer és el text que volem marcar i l'altre fitxer el que conté les marques. El lligam, entre aquests dos fitxers, ha de permetre indicar, quan sigui necessari, el número dels caràcters d'inici i final de la paraula a què volem associar una marca.

Per resoldre aquesta qüestió és necessari identificar cada element i, fins i tot, cada paraula. En aquest sentit hi ha almenys dues possibilitats: accés absolut i accés relatiu. En el primer cas es fixa la posició de cada element i de cada paraula a identificar en relació a l'inici del fitxer, mentre que en el segon cas primer s'identifica l'element i després la paraula en relació a l'inici de l'element que la conté. El primer mètode és més directe però també més vulnerable a canvis en el document, en canvi, el segon és més resistent a canvis en el document però més complex.

L'estàndard SGML permet definir mecanismes per establir aquest tipus de relació: la

manera de marcar les notes n'és un exemple. En aquest cas, el problema és més complex perquè volem establir una relació entre dos punts de fitxers diferents.

Per aconseguir-ho primer hem d'identificar unívocament cada element de l'arbre sense haver de marcar-lo explícitament. Per obtenir aquesta identificació hem de recórrer al direccionament en arbre. Aquest mecanisme identifica numèricament, d'esquerra a dreta i de l'arrel cap avall, cada branca. Per exemple, podríem referir-nos al quart fill del segon fill de l'arrel de la següent manera: "1.2.4". Amb aquest mecanisme, podem definir el valor d'un atribut com el direccionament en arbre (o localització) de l'element que volem relacionar. Per exemple, amb l'element següent:

<relaciona element1='1.4.2.1' element2='1.8.1.1'>

establim una relació entre els elements 1 i 2 d'un document SGML. El significat d'aquest lligam queda en mans de la definició que l'usuari hagi fet de l'element *relaciona*.

Aquest mecanisme permet apuntar directament un element del text a marcar. De tota manera, és necessari arribar a identificar una forma[13] dintre d'un element, la qual cosa fa imprescindible un mecanisme d'*offset* (desplaçament) que ens indiqui, en la cadena de caràcters, el número dels caràcters d'inici i acabament d'una forma. Aquest desplaçament s'indica afegint aquest número al localitzador de l'element (*localització estesa*). Si tenim la següent frase:

<s>L'altre dia vaig anar al mercat</s>,

localitzada a '1.2.4', la segona paraula (altra) queda identificada per la marca següent:

<tok from='1.2.4\3' to='1.2.4\7'>

Utilitzarem aquests mecanismes tant per indicar la informació morfosintàctica com per a l'alineació dels textos para_lels del CT.

2.2.4 Marcatge morfosintàctic d'acord amb la proposta EAGLES.

La informació de les marques de cada paraula s'incorpora a un fitxer separat que té totes les característiques d'un document SGML, la qual cosa fa que hagi de comptar, igual que els fitxers de text, amb una DTD associada i que pugui ésser visualitzat com un arbre i pugui ésser validat amb un procediment anàleg al del document SGML.

Per facilitar el processament dels fitxers de text i els de marques se'ns ha imposat la restricció següent: l'ordre dels elements ha d'ésser el mateix en cada fitxer. Tanmateix, és possible marcar només les frases, de manera que quedin al marge els elements que puguin necessitar un tractament especial, com són ara títols, taules, llistes, figures, etc.

L'estructura prevista per a aquests documents és simple. L'element que serveix d'arrel, que s'anomena *cesAna*, té com a fills la capçalera del document (*cesHeader*) i un conjunt de grups de dades (*chunkList*). Amb l'element *cesAna* es defineixen atributs que descriuen el tipus de marcatge (*type*), la posició del fitxer de text (*doc*) i la versió (*version*) de la DTD.

L'element *capçalera* s'afegeix en forma d'entitat amb la seva crida, mentre que l'element

[13] Vegeu de Yzaguirre (1996) per una descripció de la terminologia emprada a l'IULA.

chunkList agrupa elements *chunk*. Aquest últim element pot contenir *dades* (*tok*) o elements *frase* (*s*) o *paràgraf* (*p*). Sota aquests dos últims elements també podem trobar dades.

Per a cada forma s'inclou tota la informació associada sota l'element *tok*. Aquest element té els atributs següents:

- classe (*class*): indica si la paraula té alguna característica especial (nom propi, número, data, locució, etc.).
- document origen (*doc*): nom del fitxer de text.
- punt d'inici (*from*): número del primer caràcter de la paraula en la cadena de text en el fitxer origen.
- punt d'acabament (*to*): nombre de l'últim caràcter de la cadena de text en el fitxer origen.

Els elements que poden dependre de *tok* són:

- forma ortogràfica (*orth*): la paraula tal com apareix en el text original
- etiqueta (*disamb*): marca morfosintàctica resultant de l'aplicació del desambiguador
- formes lèxiques (*lex*): totes les marques morfosintàctiques possibles de la forma ortogràfica, incloent-hi per a cada una el lema i l'etiqueta

A l'annex IV mostrem un exemple d'aplicació d'aquest esquema a un document del CT.

2.2.5 Millores proposades per l'IULA.

Encara que la solució proposada satisfà plenament l'objectiu previst, si observem l'exemple de l'annex IV, ens adonarem que la informació associada a una mateixa forma es repeteix tantes vegades com aquesta apareix.

De fet, podríem pensar en una agrupació de la informació de manera tal que evités aquesta multiplicitat. Així, es podrien obtenir fitxers més compactes i avantatges en l'explotació de les dades (vegeu de Yzaguirre (1996) per a una descripció detallada d'una proposta de l'IULA en aquesta línia).

2.3 Alineació de textos.

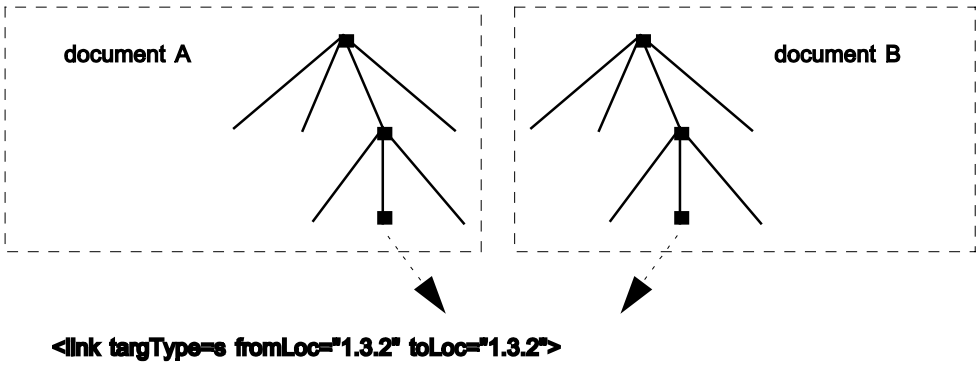
Amb l'alineació de textos es poden relacionar dos documents iguals, un del quals està traduït a una altra llengua. Aquest lligam es realitza a nivell de dades textuals, tot i que es podria fer entre els documents que contenen la informació morfosintàctica. En principi, la relació es fa tant a nivell d'elements sencers, que poden ésser frases o paràgrafs, com de *paraules gramaticals* o *segments* (*tokens*).

El mecanisme adoptat és semblant a l'emprat per la informació morfosintàctica, és a dir, el del lligam de dos fitxers mitjançant enllaços. La diferència radica en el fet que ara és necessari crear un fitxer que contingui les relacions entre els dos fitxers que apleguen les dades. Aquestes relacions poden ésser de dos tipus:

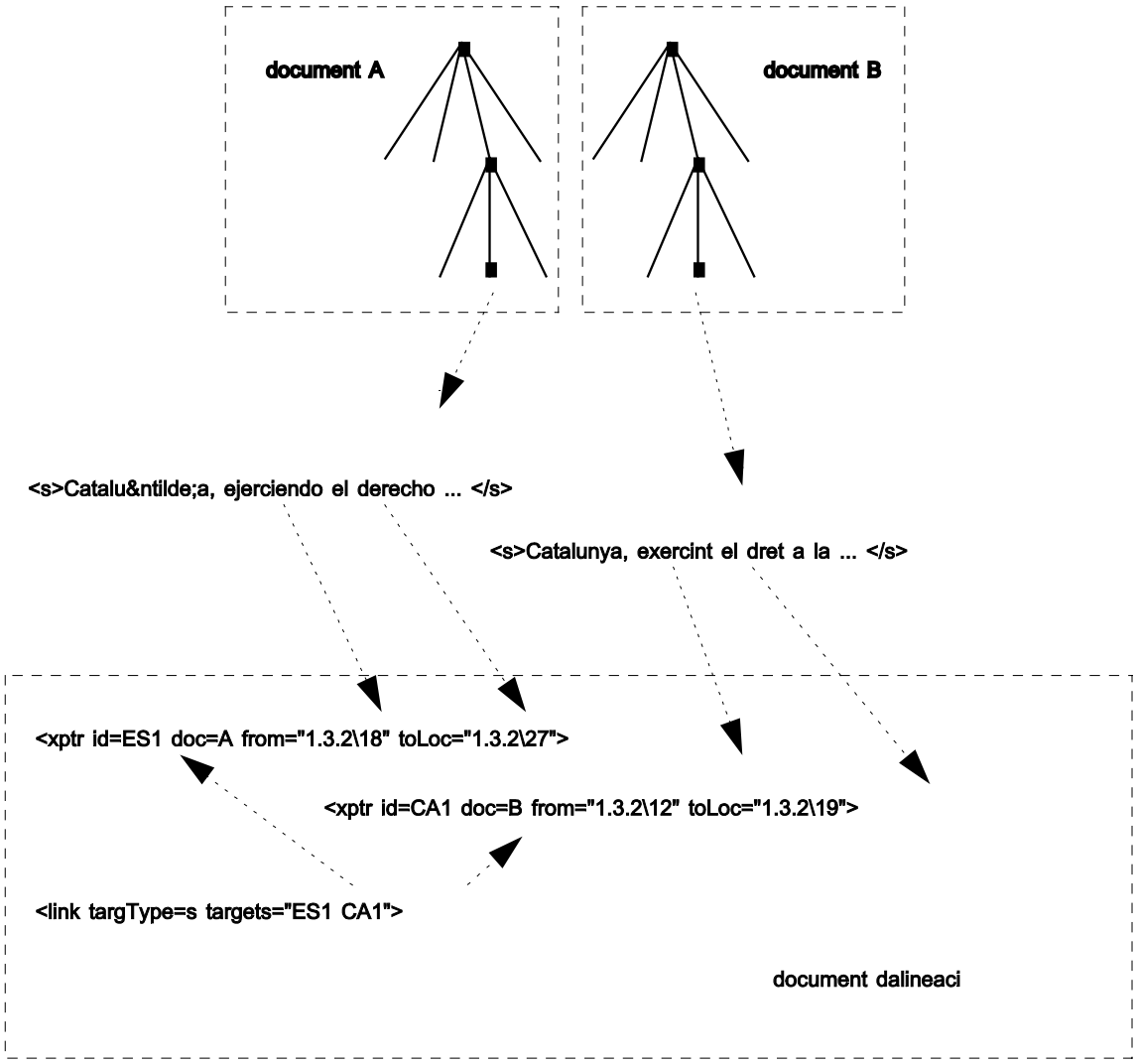
- elements sencers
- porcions d'un element

En el primer tipus de relació es vinculen elements com ara dues frases o dos *tokens*. Donats dos textos A i B, per establir un vincle d'un element del text A amb l'element corresponent del text B cal tenir present el mecanisme de localització. Aquesta relació s'explicita a través d'un element *enllaç* (*link*) que té com a atributs la localització dels elements que es volen relacionar (*fromLoc* i *toLoc*). D'aquesta manera, la relació entre el segon fill del tercer fill

de l'arrel de dos documents s'expressa de la següent manera:



En el segon tipus de relació cal definir els punts d'inici i final del segment de text que es vol alinear, la qual cosa requereix utilitzar el mecanisme de localització estesa ja mencionat en la secció 2.3.2. El procediment consisteix a definir l'element *xptr* (*extended pointer*) que ens permet identificar cadascun dels dos segments a relacionar. Per a cada segment s'indica la localització estesa del fragments a relacionar amb els atributs *from* i *to*. A més a més cada element té assignat un identificador exclusiu (atribut *id*). Finalment, es relacionen aquest dos apuntadors amb l'element d'enllaç *link* ja definit per al primer tipus de relació amb la diferència que es relacionen identificadors i no localitzadors. Vegeu-ne el procediment esquematitzat a continuació:


document A
document B
<s>Cataluña, ejerciendo el derecho ... </s>
<s>Catalunya, exercint el dret a la ... </s>
<xptr id=ES1 doc=A from="1.3.2\18" toLoc="1.3.2\27">
<xptr id=CA1 doc=B from="1.3.2\12" toLoc="1.3.2\19">
<link targType=s targets="ES1 CA1">
document dalineaci

3.- Capçalera.

Els investigadors manifesten, cada vegada més, la necessitat de treballar amb corpus de volum sempre creixent. Volums que inicialment ascendien a alguns milions de paraules (Brown: 1M, LOB: 1M, Penn Treebank: 4.5M, ...) actualment poden arribar a centenars de milions (CoBuild: 200 M, CREA: 200 M, Univ. de Michigan: 800 M). L'aplicació de tècniques estadístiques, el fet que cada vegada sigui més fàcil obtenir text en format electrònic i que les tecnologies del tractament de la informació facilitin el processament de grans volums d'informació han contribuït a aquesta expansió.

En aquest context es fa imprescindible saber quins textos formen part d'un corpus i quines són les seves característiques fonamentals. Aquesta informació no es limita merament a la bibliogràfica (nom i autor del text, editor, data de publicació, codi d'identificació, etc.) sinó que s'ha d'ampliar a altres característiques importants com ara:

- grandària: nombre de *bytes*, paraules i marques.
- convencions de codificació: nivell i tipus de codificació, autor, etc.
- classificació del text: arbre de camp, utilització del text, tipologia textual, etc.
- origen: text electrònic de l'editor/autor, escànner, distribució lliure (BBS, Internet, ...), transcripció manual, etc.
- idioma: original/traducció, disponibilitat del text en un altra llengua.
- nivell de revisió.

Aquesta informació, molt útil per als investigadors que utilitzin el textos en les seves recerques, es recull en un bloc d'informació específic anomenat *capçalera* (de l'anglès *header*).

La necessitat d'aquest bloc d'informació fou considerada per primera vegada en el marc de la iniciativa TEI i, posteriorment, aquest bloc d'informació fou adoptat en el marc del projecte EAGLES amb una diferència d'estratègia important: mentre que TEI defineix un element *header* com a part integral del text, el CES manté aquest bloc d'informació en un fitxer independent, amb la seva DTD específica. Tant el text com la capçalera mantenen informació recíproca.

Amb aquesta informació es poden conèixer exactament quins són els documents del corpus sense haver de visualitzar el cos del text. D'altra banda, aquest bloc d'informació també facilita la feina a les eines de tractament, ja que, un cop escollit un text, el contingut de la capçalera és irrellevant.

A continuació, exposem quin tipus d'informació s'afegeix a les capçaleres. En l'annex II es presenta un esborrany de la capçalera del CT, mentre que a l'annex III consignem la capçalera típica d'un document del CT.

3.1.- Estructura.

Com ja hem remarcat en l'apartat 2, la CES defineix dos elements bàsics, el corpus (*cesCorpus*) i el document (*cesDoc*), cadascun dels quals té associada una capçalera que descriu tota la informació específica. Donat que la informació que es descriu en aquests dos elements és molt semblant, aquesta es defineix només en una DTD i s'ha adaptat a cada cas. Així, en la capçalera del corpus es donen les característiques, convencions i declaracions generals que afecten tots els documents del CT, mentre que en la capçalera d'un document s'afegeix només la informació que li és pròpia.

L'element que serveix d'arrel a tota la capçalera es denomina *cesHeader* i té associats una sèrie d'atributs que permeten definir:

- el tipus de capçalera (*type*): textual o del corpus.
- el creador (*creator*): organisme responsable de la capçalera.
- la versió de la DTD associada a la capçalera (*version*).
- la condició (*status*) que especifica si és o no la versió original de la capçalera.
- la data de creació (*date.created*).
- la data de la darrera actualització (*date.updated*).

La informació que s'incorpora a la capçalera es divideix en quatre blocs, en els quals donarem compte de:
- la descripció bibliogràfica del document (vegeu 3.2).
- la relació entre el text electrònic i l'original (vegeu 3.3).
- el perfil del text (vegeu 3.4).
- el nivell de revisió (vegeu 3.5).

Una capçalera pot ésser relativament complexa o molt simple, ja que gran part de la informació és opcional. Tot seguit descrivim breument el contingut de cadascun dels blocs que s'inclouen en el CT de l'IULA.

3.2.- Descripció bibliogràfica (*fileDesc*).

La descripció bibliogràfica és la primera part i la més important de la capçalera i per això el CES considera que ha d 'ésser l'única obligatòria. El format establert és semblant a l'adoptat pels sistemes de catalogació electrònica ja existents i la seva funció és descriure el text en suport electrònic. Conté els elements següents:
- Títol (<*titleStmt*>): informació sobre el nom donat al corpus (o a la versió electrònica del document) així com la persona o institució responsable de la codificació.
- Edició (<*editionStmt*>): només s'aplica a la capçalera del corpus per indicar el nombre de versió i/o revisió.
- Extensió (<*extent*>): grandària del fitxer en nombre de *bytes* (*byteCount*) i nombre de paraules (*wordCount*). Entenem per paraula qualsevol seqüència de caràcters separada per blancs.
- Publicació (<*publicationStmt*>): s'utilitza per indicar les dades de l'editor i la disponibilitat pública dels fitxers. En el cas del CT sempre s'utilitza per indicar que la difusió és restringida.
- Origen (<*sourceDesc*>): descripció del text que ha donat origen al fitxer electrònic que s'està descrivint. Les dades que s'hi inclouen són les següents: títol, autor, edició, editor, codi d'identificació, data i lloc de publicació.

En aquest apartat l'element <*titleStmt*> només té sentit quan ens referim al corpus, mentre que l'element <*sourceDesc*> només es justifica per a cada document individual.

3.3.- Codificació (*encodingDesc*).

En aquesta secció presentem la relació entre el text en suport electrònic i el text en el seu suport original. Es tracta d'una informació que es refereix als processos de normalització, correcció, nivell de codificació i, en general, a totes les decisions que s'han pres en l'edició del textos en suport electrònic per al CT. Els elements que formen part d'aquest grup són:

- descripció del projecte (<*projectDesc*>): objectiu amb què es codifica aquest corpus
- mostreig (<*samplingDecl*>): criteri que s'ha seguit per definir les mostres dels documents
- decisions editorials (<*editorialDecl*>): criteri seguit en el CT per a la correcció del textos, tractament donat als guionets, nivell de segmentació, normalització, nivell de codificació dintre del CES, etc.
- etiquetes (<*tagsDecl*>): aquest element conté tants elements *tagUsage* com marques s'utilitzin. Té un ús diferent segons es tracta de capçaleres del CT o de les contingudes en cada document. En la capçalera del CT es dóna una descripció per a cada etiqueta, mentre que en cada document s'indica el nombre de vegades que apareix l'element, amb l'atribut *occurs*. Amb l'atribut *gi* sempre s'identifica l'etiqueta.
- construcció de referències (<*refsDecl*>): descriu el sistema emprat per codificar l'atribut ID de frases i paràgrafs per la seva utilització en la para_lelització de textos.
- classificació del text (<*classDecl*>): es representen els tres sistemes de classificació dels textos del CT mitjançant elements *category* que s'incrusten un dintre l'altre tot formant un arbre. Cada nus terminal de l'arbre té assignat un codi d'identificació en l'atribut global *id*. Es fa referència a aquest codi en l'element *textClass* de cada document (veure secció 3.1.3).

En general, la major part d'aquesta informació s'aplica a tot el CT, per tant, es reflecteix només a la capçalera del corpus. L'excepció ve donada per l'element *tagsDecl* que s'utilitza en ambdues capçaleres.

3.4.- Perfil (*profileDesc)*.

Aquesta secció dóna compte del aspectes no bibliogràfics de cada text: idioma, conjunt de caràcters utilitzats, classificació del text, disponibilitat del mateix text en altres llengües i anotacions lingüístiques. En aquest grup trobem:

- creació (<*creation*>): s'especifica com i en quina data s'ha obtingut el text que s'està codificant. S'utilitzen els codis següents: "scanner", "bd", "internet", "transcr" i "fitxer".
- idioma (<*langUsage*>): s'utilitza un o més elements per identificar cadascun dels idiomes utilitzat en el text que s'ha codificat. Amb l'atribut *type* (definit per l'IULA) s'assenyala si aquest és un text en la versió original, una traducció, etc.
- classe de text (<*textClass*>): s'utilitza en els documents per fer referència als sistemes de classificació del CT (vegeu 3.1.2).
- traduccions (<*translation*s>): dóna informació sobre possibles versions del text en una altra llengua inclosa en el CT. Conté l'element *translator*, que indica el traductor; de tota manera hem modificat la declaració per fer-lo opcional, donat que aquesta informació no sempre és disponible.
- anotacions lingüístiques (<*anotations*>): amb aquest element s'indica el nom dels documents associats a aquest text que contenen informació lingüística, alineació, etc.

3.5.- Nivell de revisió (*revisionDesc)*.

L'últim element de la capçalera dóna un mecanisme per enregistrar qualsevol canvi realitzat en els documents del CT. Cada vegada que es realitzi una modificació es consignarà tot inserint un element *change* que inclou la data, el responsable i el motiu del canvi.

4.- Estructura dels fitxers associats a un document del CT.

Com s'ha pogut observar fins ara, cada document està compost de molts fitxers (mostres, capçaleres, marcatge morfosintàctic, etc.) Tots s'han de relacionar en un primer nivell, per formar els documents i, en un segon nivell, per formar el corpus. Per realitzar aquesta vinculació és necessari recórrer a una sèrie de fitxers auxiliars.

Estudiem en primer lloc un document individual. Per poder enllaçar correctament tota la informació es necessiten fitxers amb la informació següent:

- instància SGML del document.
- cos del document: capçalera i referències a les mostres.
- definició de les entitats que formen el document.
- mostres.

Considerem ara un document hipotètic constituït per tres mostres. El fitxer que defineix les entitats ha d'incloure la definició de les entitats que fan referència a cadascuna d'elles. A més, cal definir una entitat per a la instància del document. Aquest fitxer, que podem anomenar *index*, es caracteritza per tenir un nom amb l'extensió *idx* (p.ex. d0001.idx) i té l'aspecte següent:

```
<!ENTITY doc1r    SYSTEM    "doc1.ist"    >
<!ENTITY doc1m1 SYSTEM    "doc1.1"    >
<!ENTITY doc1m2 SYSTEM    "doc1.2"    >
<!ENTITY doc1m3 SYSTEM    "doc1.3"    >
```

El següent fitxer a considerar és el que conté el cos de la instància del document, és a dir, el que conté l'element *cesHeader* i el cos del document. En el primer element s'inclou tota la informació que s'ha definit en l'apartat tres per als documents del CT. En el segon trobem les mostres del document com a part dels elements necessaris segons la DTD del CES (<text> i <body>). Aquestes mostres s'inclouen a través d'entitats que s'han definit en el fitxer *index*. Aquest fitxer té el següent aspecte:

```
<cesHeader>
    ... dades de la capçalera ...
</cesHeader>
<text>
    <body>
            &doc1m1;
            &doc1m2;
            &doc1m3;
    </body>
</text>
```

Amb aquests fitxers ja podem definir el que serà un document SGML complet. Per arribar a aquest punt hem de definir i cridar el fitxer d'índex així com cridar l'entitat que defineix el cos del document. La instància presenta l'aspecte següent:

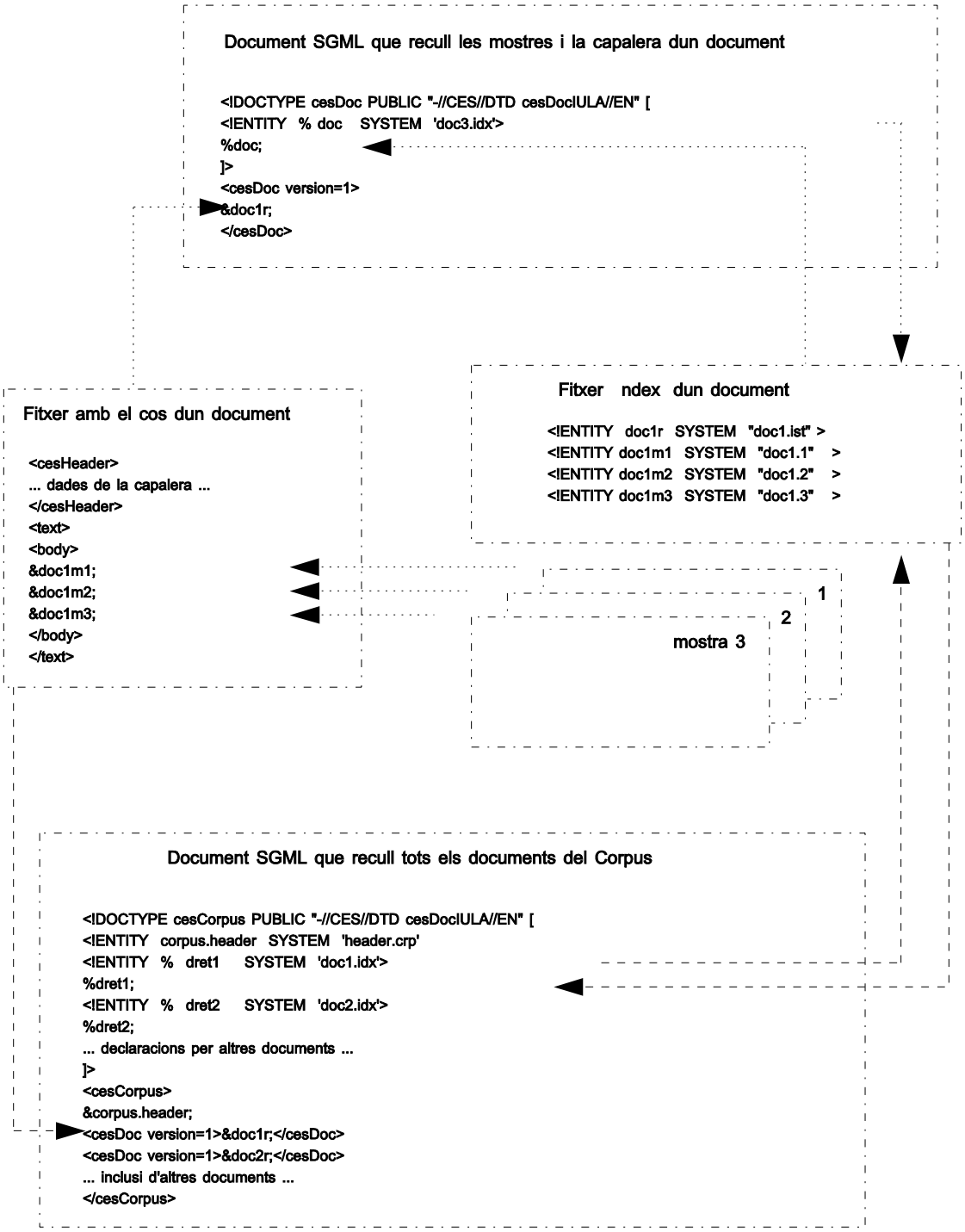
```
<!DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDocIULA//EN" [
      <!ENTITY     % doc    SYSTEM 'c:\datos\sgml\corpus\doc3.idx'>
      %doc;
]>
<cesDoc version=1>
      &doc1r;
</cesDoc>
```

D'aquesta manera, quan un programa especialment preparat per tractar fitxers SGML processa un document com el ja mencionat, substitueix les entitats pel contingut que li correspongui en cada cas, tot reconstruint el document d'origen.

L'estructura que hem esmentat permet també visualitzar el document com una part més d'un document SGML de segon nivell que recull tots els documents del corpus. Per obtenir aquesta altra visualització només hem de definir el corpus com un document SGML, tot afegint-hi tantes entitats com documents en formin part. Un fitxer amb aquestes característiques podria ésser el següent:

```
<!DOCTYPE cesCorpus PUBLIC "-//CES//DTD cesDocIULA//EN" [
      <!ENTITY     corpus.header  SYSTEM  'c:\datos\sgml\corpus\header.crp'
      <!ENTITY     % dret1    SYSTEM  'c:\datos\sgml\corpus\doc1.idx'>
      %dret1;
      <!ENTITY     % dret2    SYSTEM  'c:\datos\sgml\corpus\doc2.idx'>
      %dret2;
     ... declaracions per a altres documents ...
]>
<cesCorpus>
     &corpus.header;
     <cesDoc version=1>&doc1r;</cesDoc>
     <cesDoc version=1>&doc2r;</cesDoc>
     ... inclusió d'altres documents ...
</cesCorpus>
```

Esquemàticament podem representar la interrelació entre aquests fitxers de la següent manera:


Document SGML que recull les mostres i la capalera dun document
<!DOCTYPE cesDoc PUBLIC "-//CES//DTD cesDocIULA//EN" [
<!ENTITY % doc SYSTEM 'doc3.idx'>
%doc;
]>
<cesDoc version=1>
&doc1r;
</cesDoc>
Fitxer amb el cos dun document
<cesHeader>
... dades de la capalera ...
</cesHeader>
<text>
<body>
&doc1m1;
&doc1m2;
&doc1m3;
</body>
</text>
Fitxer ndex dun document
<!ENTITY doc1r SYSTEM "doc1.lst" >
<!ENTITY doc1m1 SYSTEM "doc1.1" >
<!ENTITY doc1m2 SYSTEM "doc1.2" >
<!ENTITY doc1m3 SYSTEM "doc1.3" >
1
2
mostra 3
Document SGML que recull tots els documents del Corpus
<!DOCTYPE cesCorpus PUBLIC "-//CES//DTD cesDocIULA//EN" [
<!ENTITY corpus.header SYSTEM 'header.crp'
<!ENTITY % dret1 SYSTEM 'doc1.idx'>
%dret1;
<!ENTITY % dret2 SYSTEM 'doc2.idx'>
%dret2;
... declaracions per altres documents ...
]>
<cesCorpus>
&corpus.header;
<cesDoc version=1>&doc1r;</cesDoc>
<cesDoc version=1>&doc2r;</cesDoc>
... inclusi d'altres documents ...
</cesCorpus>

Bibliografia

1. Bryan M., (1988): *SGML an author's guide*. Addison-Wessley Publishing Company, New York.

2. de Yzaguirre L. et al (1996): *Proposta de codificació de la informació morfosintàctica per al Corpus Tècnic de l'IULA*. En preparació.

3. de Yzaguirre L. ed. (1996): *Terminologia de la lingüística de corpus*. http://rc46.upf.es/termcorp/termcorp.htm. En preparació.

4. Garside, G., Leech, N. & Sampson G. R. (1991): *The Computational Analysis of English: a Corpus Based Approach*. Oxford University Press, Oxford.

5. Goldfarb C. F. (1990): *The SGML Handbook*. Oxford University Press, Oxford.

6. van Herwijnen E.(1995): *Practical SGML*. Kluwer Academic Publishers.

7. Ide N. & Véronis J., eds (1995): *Computers and Humanities*, Volume 29 No 1-3. Kluwer Academic Publishers.

8. Ide N. & Véronis J. (1996): *MULTEXT/EAGLES - Corpus Encoding Standart. V. 1.2*.

9. Kucera, H. & Nelson F. (1967): *Computational Analysis of Present-day American English*. Brown University Press, Providence, RI.

10. Solé X. & Saurí R. (1996): *Marcatge de llistes*. Document intern IULA. c0005.w52.

11. Sperberg-McQueen C.M. & Burnard L. (1994): *Guidelines for Electronic Text Encoding and Interchange* (TEI P3). Text Encoding Initiative. Chicago and Oxford.

12. Sperberg-McQueen C.M. & Burnard L. (1995) *TEI Lite: An introduction to text encoding for interchange*. (Document No: TEI U 5).

13. Morel J. & Bach C.: *Etiquetari per al Corpus Tècnic de l'IULA*. En preparació.

14. Pujol M. et al (1996): *Un sistema de preprocés multilingüe*. En preparació.

15. TechnoTeacher Inc., Fujitsu Ltd., (1995): *HyTime Application Development Guide*. Version 1.2.

16. Vivaldi J. et al (1997): *Adaptació d'un tagger de base estadística per al català*. En preparació.

Annex I

Codis de tipus de divisions previstos per als textos de dret

Divisió	Codi
llibre	lbr
títol preliminar	tipr
títol	tit
capítol	cap
secció	sec
subsecció	ssec
article	art
disposicions addicionals	diad
disposicions transitòries	ditr
disposicions derogatòries	dide
disposicions finals	difi
annex	anx
programa	prg
clàusula	clau
norma	nrm
acte	act
acord	acd
conveni	conv
decret	decr
llei	llei
ordre	ord
tractat	trac
reglament	reg
formulari	frm
paràgraf	par
protocol	prot
apèndix	apen
declaració	decl
preàmbul	prea
part dispositiva	padi
part final	pafi
avantprojecte de llei	avpr
part	part
dictamen	dict
decisió	deci

Annex II

Esborrany de la capçalera del Corpus Tècnic

```
<cesHeader version='2.1' type=corpus creator=IULA>
<!-- capçalera del Corpus de Dret -->
   <fileDesc>
      <titleStmt>
            <h.title>Corpus de Dret </h.title>
            <respStmt>
                  <respType>Director</respType>
                  <respName>M. T. Cabr&eacute;</respName>
            </respStmt>
      </titleStmt>
      <publicationStmt>
            <distributor>IULA</distributor>
            <pubAddress>Rambla Santa M&oacute;nica 32
                  08002 Barcelona Spain</pubAddress>
            <availability status=restricted>Us intern</availability>
            <pubDate>1996</pubDate>
      </publicationStmt>
   </fileDesc>
   <encodingDesc>
      <projectDesc></projectDesc>
      <samplingDecl>mostres entre 3000 i 5000 paraules cadascuna; si el
          document origen és en suport paper el nombre de mostres es limita a deu
            aproximadament, cadascuna en un fitxer independent
      </samplingDecl>
      <EditorialDecl>
            <conformance>CES nivell 3</conformance>
               <correction status=medium>diccionari de WordPerfect</correction>
            <quotation marks=all form=nonstd>les cometes es mantenen en la forma original,
                        és a dir, sense marcar</quotation>
         <hyphenation>els guionets de final de línia s'han eliminat, mentre que
                        els guionets associats a paraules (fer-se) s'han mantingut</hyphenation>
            <segmentation>nivell de frase</segmentation>
            <normalization>cap</normalization>
      </EditorialDecl>
      <tagsDecl>
          <tagUsage gi=cesDoc>arrel de tos els documents del CT</tagUsage>
          <tagUsage gi=text>text</tagUsage>
          <tagUsage gi=body>cos del document</tagUsage>
          <tagUsage gi=div1>divisió a nivell u</tagUsage>
          <tagUsage gi=div2>divisió a nivell dos</tagUsage>
          <tagUsage gi=div3>divisió a nivell tres</tagUsage>
          <tagUsage gi=div4>divisió a nivell quatre</tagUsage>
          <tagUsage gi=div5>divisió a nivell cinc</tagUsage>
          <tagUsage gi=div6>divisió a nivell sis</tagUsage>
```

```
    <tagUsage gi=div7>divisió a nivell set</tagUsage>
    <tagUsage gi=div8>divisió a nivell vuit</tagUsage>
    <tagUsage gi=div9>divisió a nivell nou</tagUsage>
    <tagUsage gi=opener>fragment introductori</tagUsage>
    <tagUsage gi=closer>fragment de tancament</tagUsage>
    <tagUsage gi=head>títol d'una divisió, llista, taula, etc.</tagUsage>
    <tagUsage gi=dateline>encapçalament que indica lloc, data, etc.</tagUsage>
    <tagUsage gi=p>paràgraf</tagUsage>
    <tagUsage gi=s>frase</tagUsage>
    <tagUsage gi=list>llista</tagUsage>
    <tagUsage gi=label>etiqueta d'una llista</tagUsage>
    <tagUsage gi=item>element d'una llista</tagUsage>
    <tagUsage gi=note>nota a peu de pàgina, final de capítol, etc.</tagUsage>
    <tagUsage gi=ptr>apuntador a una nota, figura, taula, etc.</tagUsage>
    <tagUsage gi=table>taula</tagUsage>
    <tagUsage gi=row>fila d'una taula</tagUsage>
    <tagUsage gi=cell>cel·la d'una taula</tagUsage>
    <tagUsage gi=figure>figura</tagUsage>
    <tagUsage gi=gap>fragment no inclós</tagUsage>
    <tagUsage gi=hi>fragment amb una característica tipogràfica:
            negreta, cursiva, etc.</tagUsage>
    <tagUsage gi=foreign>paraules en una llengua diferent a la
            pròpia del document</tagUsage>
</tagsDecl>
<classDecl>
  <taxonomy>
    <category id=DretPrivat>
        <catDesc>Dret privat</catDesc>
        <category id=DretPC>
            <catDesc>Dret civil</catDesc>
        </category>
        <category id=DretPM>
            <catDesc>Dret mercantil</catDesc>
        </category>
        <category id=DretPL>
            <catDesc>Dret laboral</catDesc>
        </category>
        <category id=DretPP>
            <catDesc>Dret penal</catDesc>
        </category>
        <category id=DretPA>
            <catDesc>Dret canònic</catDesc>
        </category>
    </category>
    <category id=DretPublic>
        <catDesc>Dret Públic</catDesc>
        <category id=DretUC>
            <catDesc>Dret Constitucional i estatutari</catDesc>
        </category>
        <category id=DretUA>
            <catDesc>Dret Administratiu</catDesc>
```

```
            </category>
            <category id=DretUF>
                <catDesc>Dret Financer i Tributari</catDesc>
            </category>
            <category id=DretUI>
                <catDesc>Dret Internacional Públic</catDesc>
            </category>
        </category>
        <category id=DretTeoria>
            <catDesc>Teoria del Dret</catDesc>
        </category>
      </taxonomy>
      <taxonomy>
        <category id=DretTLegal>
            <catDesc>Textos legals (lleis i reglaments)</catDesc>
        </category>
        <category id=DretTJudicial>
            <catDesc>Textos judicials(escrits de les parts,
            dilig&egrave;ncies, sent&egrave;ncies, etc.)</catDesc>
        </category>
        <category id=DretTeoric>
            <catDesc>Textos te&ograve;rics (manuals, monografies,
            articles, etc.)</catDesc>
        </category>
        <category id=DretTInstr>
            <catDesc>Textos instrumentals (diccionaris,
            vocabularis, glossaris, etc.)</catDesc>
        </category>
      </taxonomy>
      <taxonomy>
        <category id=descriptiu>
            <catDesc>descriptiu</catDesc>
        </category>
        <category id=narratiu>
            <catDesc>narratiu</catDesc>
        </category>
        <category id=expositiu>
            <catDesc>expositiu</catDesc>
        </category>
        <category id=argumentatiu>
            <catDesc>argumentatiu</catDesc>
        </category>
        <category id=instructiu>
            <catDesc>instructiu</catDesc>
        </category>
      </taxonomy>
    </classDecl>
  </encodingDesc>
</cesHeader>
```

Annex III

Capçalera d'un document del Corpus Tècnic

```
<cesHeader creator=IULA status=new date.created="06/13/96" version='2.1i'>    <filedesc>
   <editionStmt>1</editionStmt>
    <extent>
     <wordcount>75883</wordcount>
     <bytecount>653849</bytecount>
    </extent>
   <publicationStmt>
    <distributor> </distributor>
    <pubAddress> </pubAddress>
    <availability status=restricted> </availability>
    <pubDate> </pubDate>
   </publicationStmt>
   <sourceDesc>
    <biblStruct>
     <monogr>
       <h.title> La redacci&oacute; de les Lleis </h.title>
       <h.author> Oriol Camps </h.author>
       <edition> 1
       </edition>
       <imprint>
     <publisher type=org> Escola d'Administraci&oacute; P&uacute;blica de Catalunya
           </publisher>
        <pubDate>11/1/95</pubDate>
        <pubPlace> Barcelona </pubPlace>
       </imprint>
       <idno type=isbn> 84-393-3691-8 </idno>
     </monogr>
    </biblStruct>
   </sourceDesc>
  </filedesc>
  <encodingDesc>
    <tagsdecl>
     <tagUsage gi=cell n=3></tagUsage>
     <tagUsage gi=label n=2></tagUsage>
     <tagUsage gi=table n=3></tagUsage>
     <tagUsage gi=foreign n=350></tagUsage>
     <tagUsage gi=item n=211></tagUsage>
     <tagUsage gi=p n=1538></tagUsage>
     <tagUsage gi=hi n=953></tagUsage>
     <tagUsage gi=head n=239></tagUsage>
     <tagUsage gi=s n=2458></tagUsage>
     <tagUsage gi=div1 n=8></tagUsage>
     <tagUsage gi=div2 n=49></tagUsage>
```

```
    <tagUsage gi=ptr n=228></tagUsage>
    <tagUsage gi=note n=228></tagUsage>
    <tagUsage gi=div3 n=85></tagUsage>
    <tagUsage gi=corr n=8></tagUsage>
    <tagUsage gi=div4 n=153></tagUsage>
    <tagUsage gi=div5 n=5></tagUsage>
    <tagUsage gi=row n=3></tagUsage>
    <tagUsage gi=list n=54></tagUsage>
   </tagsdecl>
  </encodingDesc>
  <profileDesc>
   <creation date="06/13/96">diskette</creation>
   <langUsage status=orig  docLangs="c----">
    <language id=CA iso639='CA'>Catalan</language>
    <language id=EN iso639='EN'>English</language>
    <language id=RU iso639='RU'>Russian</language>
    <language id=ES iso639='ES'>Spanish</language>
    <language id=FR iso639='FR'>French</language>
    <language id=DE iso639='DE'>German</language>
    <language id=LA iso639='LA'>Latin</language>
    <language id=IT iso639='IT'>Italian</language>
    <language id=JE iso639='JE'></language>
    <language id=PO iso639='PO'></language>
   </langUsage>

   <textClass>
     <catRef target="duc teor inf">
       <h.keywords>
         <keyTerm>dret constitucional i estatutari</keyTerm>
         <keyTerm>te&ograve;ric</keyTerm>
         <keyTerm>informatiu</keyTerm>
       </h.keywords>
   </textClass>
  </profileDesc>
  <revisionDesc>
   <change>
    <changeDate>06/13/96</changeDate>
    <respName>vivaldi</respName>
    <h.item>marcatge nivel 2</h.item>
   </change
  </revisionDesc>
 </cesHeader>
```

Annex IV

Fragment d'un document del Corpus Tècnic codificat a nivell 3

```
<div1 n=1 complete=n>
<head type=main rend=bo><num>1</num> LLEI <num>19/1985</num>,
DE <date>16 DE JULIOL</date>, CANVI&Agrave;RIA I DEL XEC</head>
<head type=sub>(<name>BOE</name> dia <date>19 de juliol</date>)
</head>

<div2 n=1 complete=y>
<head type=main rend=bo>PRE&Agrave;MBUL</head>

<div3 n=1 complete=y>
<p><s>L'adaptaci&oacute; de l'ordenament sobre la lletra de canvi,
el xec i el pagar&eacute; a l'anomenada legislaci&oacute; uniforme
de <name>Ginebra</name> suposa donar un pas decisiu encaminat a
la renovaci&oacute; del nostre <name>Dret Mercantil</name>, tan
necessitat de reforma.</s><s>Si aquesta necessitat &eacute;s
predicable d'altres sectors de l'ordenament mercantil, en pocs
es fa tan evident com en el d'aquests t&iacute;tols valors, la
regulaci&oacute; dels quals, quasi centen&agrave;ria, ha estat
denunciada repetidament perqu&egrave; no serveix per protegir
adequadament els cr&egrave;dits incorporats als dits documents.
</s></p>
<p><s>La regulaci&oacute; de la lletra de canvi, continguda en
el t&iacute;tol <num>X</num> del llibre <num>II</num> del <name>Codi
de Comer&ccedil;</name>, est&agrave; inspirada directament en
la del seu hom&ograve;nim franc&egrave;s, dominat, quan aquell
es va promulgar, per una concepci&oacute; instrumental de la
lletra de canvi, sobre la qual incidien directament totes les
vicissituds del negoci causal.</s><s>Aquesta concepci&oacute;
xoca obertament amb les necessitats del tr&agrave;fic jur&iacute;dic
contemporani, en qu&egrave; la circulaci&oacute; dels t&iacute;tols
no pot quedar sotmesa al mateix r&egrave;gim que la simple cessi&oacute;
de cr&egrave;dits.</s><s>Aquestes insufici&egrave;ncies estan
vinculades directament al sistema d'excepcions oposables pel
deutor canviari, del qual ha fet un problema eminentment processal
la circumst&agrave;ncia d'&eacute;sser la <name>Llei d'Enjudiciament
Civil</name> anterior al <name>Codi de Comer&ccedil;</name>,
quan, <loc pos='D'>al contrari</loc>, la seva soluci&oacute;
&eacute;s determinant del r&egrave;gim jur&iacute;dic substantiu
d'aquests t&iacute;tols.</s><s>Dit amb d'altres paraules: del
r&egrave;gim d'excepcions que s'adopti dep&egrave;n que es perpetu&iuml;
la configuraci&oacute; causalista de la lletra, o <loc pos='L'>b&eacute;
que</loc> s'inici&iuml; la tend&egrave;ncia a l'abstracci&oacute;
del t&iacute;tol.</s></p>
</div3>
```

<div3 n=2 complete=y>
<s>Aquestes insuficiències no són, <loc pos='D'>en
canvi</loc>, l'únic factor determinant de la reforma que
es proposa.</s><s>S'hi ha d'afegir la voluntat política
d'incorporar <name>Espanya</name> al conjunt d'<name>Estats</name>
que estan contribuint a fer realitat el propòsit, explícit,
per exemple en l'article <lab>3,h)\=:B#</lab>, del <name>Tractat
de Roma</name>, de constitució de la <name>Comunitat Econòmica
Europea</name>, d'aproximar les legislacions nacionals en la
mesura necessària per al funcionament del <name>Mercat
Comú</name>.</s>
<s>Ja se sap que el <name>Dret Mercantil</name> ha reivindicat
històricament la nota d'universalitat, molt <loc pos='L'>abans
que</loc> les relacions de tot tipus entre els pobles i entre
els <name>Estats</name> assolissin el grau de fluïdesa que
tenen en l'actualitat.</s><s><loc pos='D'>De fet</loc>, l'intercanvi
empresarial entre <name>Estats</name> dotats de sistemes polítics
similars, que reconeixen, <loc pos='D'>al seu torn</loc>, sistemes
similars d'organització econòmica, requereix l'existència
de regulacions homogènies en un bon nombre d'institucions.
</s>
<s>Una de les categories de l'entorn institucional comú
-l'autonomia de la voluntat- ha permès que els sectors
interessats acudissin a l'autoregulació i a la unificació
de practiques negociables en força ocasions.</s><s>Però
quan l'autoregulació no és possible, han estat
els <name>Estats</name> i les <name>Organitzacions Internacionals</name>
els que han tingut cura d'accentuar els perfils comuns de les
institucions necessàries perquè el tràfic
jurídic es desenvolupi adequadament.</s><s>Un d'aquests
casos, el constitueix l'ordenament de la lletra de canvi, del
pagaré a l'ordre i del xec, contingut en les <name>Lleis
Uniformes</name> annexes als <name>Convenis de Ginebra</name>
de <date ISO8601='06/07/1930'>7 de juny de 1930</date> i de <date
ISO8601='03/19/1931'>19 de març de 1931</date>.</s><s>
Aquesta <name>Llei</name> recull, substancialment, la regulació
ginebrina.</s>
</div3>

<div3 n=3 complete=y>
<s>L'opció manifestada pel sistema de les <name>Lleis
de Ginebra</name> es fonamenta, bàsicament, en la superioritat
tècnica d'aquesta normativa enfront del nostre <name>Codi
de Comerç</name>.</s>
<s>Les novetats que incorpora la <name>Llei</name> tenen manifestacions
múltiples i comencen per la senzillesa amb què
es delimiten els requisits formals dels títols regulats
i el vigor amb què es defensa la validesa genèrica
de cada una de les declaracions que s'hi incorporen <loc pos='L'>encara
que</loc> algunes de les altres estigui afectada per vicis invalidants.
</s><s><loc pos='D'>En definitiva</loc>, es tracta de facilitar

la circulació d'aquests documents sense imposar a l'adquirent
la càrrega d'examinar, <loc pos='D'>a més</loc>
de la regularitat formal dels endossaments, la validesa intrínseca
de totes les declaracions precedents.</s><s>També acull
situacions que es produeixen, en la pràctica, tals com
els títols <loc pos='D'>en blanc</loc>, que no tenen regulació
en els textos vigents, la subscripció d'aquests documents
al·legant una representació inexistent (problema
per a la solució del qual s'ha d'acudir avui a categories
extracanviàries), el xec per abonar en compte o el xec
certificat o conformat.</s><s>Quan es refereix als requisits
formals del títol s'ha de ressaltar la desaparició
de la menció de la «clàusula de valor»
en la lletra de canvi, rastre evident de la concepció
causal que domina, si bé no amb absoluta exclusivitat,
el sistema vigent.</s>
<s>La superioritat tècnica dels textos ginebrins ressalta
especialment en els articles <num>17</num> de la <name>Llei de
la Lletra de Canvi</name> i <num>22</num> de la del <name>Xec</name>,
dels quals són transsumpte fidel els articles <num>20</num>
i <num>128</num>, respectivament, d'aquesta llei.</s><s>S'hi
delimita negativament i amb una senzillesa encomiable el règim
d'excepcions, sense necessitat d'acudir a llistes taxades, cosa
que contrasta amb la dicció tallant del vigent article
<num>480</num> del <name>Codi de Comerç</name>, que tantes
matisacions jurisprudencials ha rebut en els seus <num>cent</num>
anys de vigència.</s>
<s>També mereix una menció especial la configuració
de l'aval.</s><s>El text tracta de posar fi a la polèmica
doctrinal i jurisprudencial sobre la naturalesa jurídica
d'aquesta relació canviaria, optant per la seva definició
com a obligació autònoma, vàlida <loc pos='L'>encara
que</loc> sigui nul la l'obligació garantida per un motiu
distint dels vicis de forma.</s><s>Les normes sobre la presentació
a l'acceptació, en el cas de la lletra, i al pagament
de les <num>tres</num> classes de títols regulats denoten
la flexibilitat amb què s'aborda aquesta matèria;
cal destacar l'ampliació dels terminis per presentar a
l'acceptació les lletres de canvi girades <loc pos='D'>a
la vista</loc> i, <loc pos='D'>en general</loc>, per a la presentació
al pagament d'aquests títols.</s>
</div3>

<div3 n=4 complete=y>
<s>Les <name>Lleis</name> uniformes tenen el propòsit
manifest d'enfortir la posició jurídica del creditor
canviari.</s><s>Aquest propòsit té el seu reflex
en aquesta <name>Llei</name>, no només en la formulació
d'excepcions oposables, a que ja s'ha fet menció per subratllar
el seu caràcter substantiu, sinó en altres àmbits.
</s><s>S'ha de destacar en primer lloc la flexibilitat amb què

s'aborda el règim de protest, permetent-ne la substitució
per declaracions del lliurat o de la <name>Cambra de Compensació</name>
o eliminant-lo.</s><s>També suposa una novetat, almenys
com a formulació normativa, el que la rigorosa obligació
de l'acceptant de la lletra de canvi i dels seus avaladors no
quedi sotmesa a condició de protest o a una declaració
equivalent.</s><s>Un altre mecanisme fonamental per reforçar
la garantia del tenidor és l'establiment de la solidaritat
passiva absoluta dels deutors canviaris, els quals, amb independència
de la seva posició en el títol, podran ser demandats
conjuntament o separadament.</s><s>També es poden emmarcar
entre les mesures que han de suposar, indirectament, una millor
situació del creditor, l'establiment d'un interès
de demora més adequat a la situació <loc pos='D'>del
moment</loc> en què es produeixi l'impagament d'un dels
títols.</s><s>Per al cas concret del xec es preveu una
clàusula penal que jugarà contra el lliurador que
emeti un xec sense tenir provisió de fons en poder del
lliurat.</s><s>Un nou camí procedimental per al judici
executiu canviari completa les mesures tendents a reforçar
la posició del tenidor, <loc pos='D'>a més</loc>
de la reforma de l'article <num>1.429</num> de la <name>Llei
d'Enjudiciament Civil</name>, per tal d'incloure en l'enunciat
dels títols executius el pagaré i el xec.</s>
<s>La <name>Llei</name> dedica, finalment, <num>dos</num>
capítols a resoldre els problemes derivats del conflicte
de <name>Lleis</name>.</s>
</div3>

<div3 n=5 complete=y>
<s>No es pot negar el descrèdit relatiu que envolta
avui la lletra de canvi en la nostra societat; és cert
que tal actitud no deriva exclusivament, ni tan sols principalment,
de les insuficiències normatives que han estat exposades.
</s><s>La situació critica que viu la nostra economia
i una utilització desmesurada de la lletra de canvi, tant
en el mercat de béns i serveis de consum com en el mercat
financer, i unes lleis processals obsoletes, no són factors
estanys al nombre elevadíssim d'impagats que recullen
les estadístiques.</s><s>La nova <name>Llei</name>, rigorosa
amb el deutor, vol canviar certs usos que han contribuït
a aquest descrèdit, restablint la confiança en
l'<name>Ordenament</name> jurídic i en un dels valors
fonamentals de la vida empresarial, la bona fe.</s>
<s>La normativa jurídica que introdueix aquesta <name>Llei</name>,
absolutament necessària i convenient, no impedeix que,
després dels estudis oportuns i quan les circumstàncies
econòmiques i socials ho requereixin, es pugui abordar
l'elaboració d' un text legal complementari i específic
que estableixi les normes que hagin de regir per a les lletres
emeses en operacions realitzades pels consumidors i usuaris.

</s><s>Les diferents orientacions dels ordenaments jurídics
d'altres països europeus, i també la inexistència
de normativa uniforme en aquesta matèria, aconsellen de
no introduir en la present <name>Llei</name> la seva regulació
definitiva, sens perjudici que això pugui fer-se i hagi
de fer-se en el moment oportú.</s></p>
</div3>
</div2>
</div1>

Annex V

Informació morfosintàctica associada a un fragment d'un document del Corpus Tècnic

```
<chunkList>
  <chunk doc="d0004.1" from='1.1.1\1'>
    <tok doc="d0004.1" class=tok from='1.1.1.1\1'>
        <orth>1</orth>
        <disamb><ctag>X</ctag></disamb>
        <lex>
            <base>1</base>
            <ctag>X</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.1\14'>
        <orth>llei</orth>
        <disamb><ctag>N5FS</ctag></disamb>
        <lex>
            <base>llei</base>
            <ctag>N5FS</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.1.2\1'>
        <orth>19/1985</orth>
        <disamb><ctag>X</ctag></disamb>
        <lex>
            <base>19/1985</base>
            <ctag>X</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.1\40'>
        <orth>de</orth>
        <disamb><ctag>P</ctag></disamb>
        <lex>
            <base>de</base>
            <ctag>P</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.1.3\1'>
        <orth>16 DE JULIOL</orth>
        <disamb><ctag>T</ctag></disamb>
        <lex>
            <base>16 DE JULIOL</base>
            <ctag>T</ctag>
        </lex>
    </tok>
```

```
  <tok doc="d0004.1" class=tok from='1.1.1\70'>
      <orth>canvi&agrave;ria</orth>
      <disamb><ctag>JQMS</ctag></disamb>
      <lex>
          <base>canviari</base>
          <ctag>JQMS</ctag>
      </lex>
  </tok>
  <tok doc="d0004.1" class=tok from='1.1.1\87'>
      <orth>i</orth>
      <disamb><ctag>C</ctag></disamb>
      <lex>
          <base>i</base>
          <ctag>C</ctag>
      </lex>
  </tok>
  <tok doc="d0004.1" class=pgr from='1.1.1\89' to='1.1.1\89'>
      <orth>d</orth>
      <disamb><ctag></ctag></disamb>
      <lex>
          <base>de</base>
          <ctag>P</ctag>
      </lex>
  </tok>
  <tok doc="d0004.1" class=pgr from='1.1.1\90' to='1.1.1\91'>
      <orth>el</orth>
      <disamb><ctag>AMS</ctag></disamb>
      <lex>
          <base>el</base>
          <ctag>AMS</ctag>
      </lex>
  </tok>
  <tok doc="d0004.1" class=tok from='1.1.1\93'>
      <orth>xec</orth>
      <disamb><ctag>N5MS</ctag></disamb>
      <lex>
          <base>xec</base>
          <ctag>N5MS</ctag>
      </lex>
  </tok>
 </chunk>

<chunk doc="d0004.1" from='1.1.2\1'>
  <tok doc="d0004.1" class=tok from='1.1.2.1\1'>
      <orth>BOE</orth>
      <disamb><ctag>N4</ctag></disamb>
      <lex>
          <base>BOE</base>
          <ctag>N4</ctag>
      </lex>
  </tok>
```

```
   <tok doc="d0004.1" class=tok from='1.1.2\1'>
         <orth>dia</orth>
         <disamb><ctag>N5MS</ctag></disamb>
         <lex>
               <base>dia</base>
               <ctag>N5MS</ctag>
         </lex>
   </tok>
   <tok doc="d0004.1" class=tok from='1.1.2.2\1'>
         <orth>19 de juliol</orth>
         <disamb><ctag>T</ctag></disamb>
         <lex>
               <base>19 de juliol</base>
               <ctag>T</ctag>
         </lex>
   </tok>
 </chunk>

<chunk doc="d0004.1" from='1.1.3.1\1'>
   <tok doc="d0004.1" class=tok from='1.1.3.1\1'>
         <orth>pre&agrave;mbul</orth>
         <disamb><ctag>N5MS</ctag></disamb>
         <lex>
               <base>pre&agrave;mbul</base>
               <ctag>N5MS</ctag>
         </lex>
   </tok>
 </chunk>

<chunk doc="d0004.1" from='1.1.3.2.1\1'>
   <tok doc="d0004.1" class=tok from='1.1.3.2.1\1'>
         <orth>I</orth>
         <disamb><ctag>X</ctag></disamb>
         <lex>
            <base>i</base>
            <ctag>C</ctag>
         </lex>
         <lex>
            <base>I</base>
            <ctag>X</ctag>
         </lex>
   </tok>
 </chunk>

<chunk doc="d0004.1" from='1.1.3.2.2.1\1'>
   <tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\1' to='1.1.3.2.2.1\1'>
         <orth>L'</orth>
         <disamb><ctag>AMS</ctag></disamb>
         <lex>
             <base>el</base>
             <ctag>AMS</ctag>
```

```
        </lex>
    </tok>
    <tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\3' to='1.1.3.2.2.1\18'>
        <orth>adaptaci&oacute;</orth>
        <disamb><ctag>N5FS</ctag></disamb>
        <lex>
            <base>adaptaci&oacute;</base>
            <ctag>N5FS</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.3.2.2.1\20'>
        <orth>de</orth>
        <disamb><ctag>P</ctag></disamb>
        <lex>
            <base>de</base>
            <ctag>P</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\23' to='1.1.3.2.2.1\23'>
        <orth>l'</orth>
        <disamb><ctag>AMS</ctag></disamb>
        <lex>
            <base>el</base>
            <ctag>AMS</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\24' to='1.1.3.2.2.1\33'>
        <orth>ordenament</orth>
        <disamb><ctag>N5MS</ctag></disamb>
        <lex>
            <base>ordenament</base>
            <ctag>N5MS</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.3.2.2.1\35'>
        <orth>sobre</orth>
        <disamb><ctag>P</ctag></disamb>
        <lex>
            <base>sobre</base>
            <ctag>D4</ctag>
        </lex>
        <lex>
            <base>sobre</base>
            <ctag>P</ctag>
        </lex>
        <lex>
            <base>sobre</base>
            <ctag>N5MS</ctag>
        </lex>
    </tok>
    <tok doc="d0004.1" class=tok from='1.1.3.2.2.1\41'>
```

```
        <orth>la</orth>
        <disamb><ctag>AFS</ctag></disamb>
        <lex>
            <base>el</base>
            <ctag>AFS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\44'>
        <orth>lletra</orth>
        <disamb><ctag>N5FS</ctag></disamb>
        <lex>
            <base>lletra</base>
            <ctag>N5FS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\51'>
        <orth>de</orth>
        <disamb><ctag>P</ctag></disamb>
        <lex>
            <base>de</base>
            <ctag>P</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\54'>
        <orth>canvi</orth>
        <disamb><ctag>N5MS</ctag></disamb>
        <lex>
            <base>canvi</base>
            <ctag>N5MS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\62'>
        <orth>el</orth>
        <disamb><ctag>AMS</ctag></disamb>
        <lex>
            <base>el</base>
            <ctag>AMS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\65'>
        <orth>xec</orth>
        <disamb><ctag>N5MS</ctag></disamb>
        <lex>
           <base>xec</base>
            <ctag>N5MS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\69'>
        <orth>i</orth>
        <disamb><ctag>C</ctag></disamb>
        <lex>
```

<base>i</base>
<ctag>C</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\71'>
<orth>el</orth>
<disamb><ctag>AMS</ctag></disamb>
<lex>
<base>el</base>
<ctag>AMS</ctag>
</lex>
<lex>
<base>el</base>
<ctag>REEC3MS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\74'>
<orth>pagaré</orth>
<disamb><ctag>N5MS</ctag></disamb>
<lex>
<base>pagar</base>
<ctag>VDU1S</ctag>
</lex>
<lex>
<base>pagaré</base>
<ctag>N5MS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\88'>
<orth>a</orth>
<disamb><ctag>P</ctag></disamb>
<lex>
<base>a</base>
<ctag>P</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\89' to='1.1.3.2.2.1\89'>
<orth>l'</orth>
<disamb><ctag>AMS</ctag></disamb>
<lex>
<base>el</base>
<ctag>AMS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\90' to='1.1.3.2.2.1\98'>
<orth>anomenada</orth>
<disamb><ctag>N5FS</ctag></disamb>
<lex>
<base>anomenada</base>
<ctag>N5FS</ctag>
</lex>

```
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\100'>
    <orth>legislaci&oacute;</orth>
    <disamb><ctag>N5FS</ctag></disamb>
    <lex>
        <base>legislaci&oacute;</base>
        <ctag>N5FS</ctag>
    </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\118'>
    <orth>uniforme</orth>
    <disamb><ctag>JQ6S</ctag></disamb>
    <lex>
        <base>uniforme</base>
        <ctag>JQ6S</ctag>
    </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\128'>
    <orth>de</orth>
    <disamb><ctag>P</ctag></disamb>
    <lex>
        <base>de</base>
        <ctag>P</ctag>
    </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1.1\137'>
    <orth>Ginebra</orth>
    <disamb><ctag>N4</ctag></disamb>
    <lex>
        <base>ginebra</base>
        <ctag>N5FS</ctag>
    </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\152'>
    <orth>suposa</orth>
    <disamb><ctag>V8R6S</ctag></disamb>
    <lex>
        <base>suposar</base>
        <ctag>V8R6S</ctag>
    </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\159'>
    <orth>donar</orth>
    <disamb><ctag>VI</ctag></disamb>
    <lex>
        <base>donar</base>
        <ctag>VI</ctag>
    </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\165'>
    <orth>un</orth>
```

```
        <disamb><ctag>AMS</ctag></disamb>
        <lex>
            <base>un</base>
            <ctag>ENMS</ctag>
        </lex>
        <lex>
            <base>un</base>
            <ctag>AMS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\168'>
        <orth>pas</orth>
        <disamb><ctag>N5MS</ctag></disamb>
        <lex>
            <base>pas</base>
            <ctag>N5MS</ctag>
        </lex>
        <lex>
            <base>pas</base>
            <ctag>D4</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\172'>
        <orth>decisiu</orth>
        <disamb><ctag>JQMS</ctag></disamb>
        <lex>
            <base>decisiu</base>
            <ctag>JQMS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\180'>
        <orth>encaminat</orth>
        <disamb><ctag>VCMS</ctag></disamb>
        <lex>
            <base>encaminar</base>
            <ctag>VCMS</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\192'>
        <orth>a</orth>
        <disamb><ctag>P</ctag></disamb>
        <lex>
            <base>a</base>
            <ctag>>P</ctag>
        </lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\195'>
        <orth>la</orth>
        <disamb><ctag>AFS</ctag></disamb>
        <lex>
            <base>el</base>
```

<ctag>AFS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\198'>
<orth>renovació</orth>
<disamb><ctag>N5FS</ctag></disamb>
<lex>
<base>renovació</base>
<ctag>N5FS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\215' to='1.1.3.2.2.1\215'>
<orth>d</orth>
<disamb><ctag>P</ctag></disamb>
<lex>
<base>de</base>
<ctag>P</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=pgr from='1.1.3.2.2.1\216' to='1.1.3.2.2.1\217'>
<orth>el</orth>
<disamb><ctag>AMS</ctag></disamb>
<lex>
<base>el</base>
<ctag>AMS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\219'>
<orth>nostre</orth>
<disamb><ctag>JP21MS</ctag></disamb>
<lex>
<base>nostre</base>
<ctag>JP21MS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1.2\1'>
<orth>Dret Mercantil</orth>
<disamb><ctag>N4MS</ctag></disamb>
<lex>
<base>Dret Mercantil</base>
<ctag>N4MS</ctag>
</lex>
</tok>
<tok doc="d0004.1" class=tok from='1.1.3.2.2.1\255'>
<orth>tan</orth>
<disamb><ctag>D4</ctag></disamb>
<lex>
<base>tan</base>
<ctag>D4</ctag>
</lex>
</tok>

```
        <tok doc="d0004.1" class=tok from='1.1.3.2.2.1\260'>
            <orth>necessitat</orth>
            <disamb><ctag>VCMS</ctag></disamb>
            <lex>
                <base>necessitar</base>
                <ctag>VCMS</ctag>
            </lex>
            <lex>
                <base>necessitat</base>
                <ctag>N5FS</ctag>
            </lex>
            <lex>
                <base>necessitat</base>
                <ctag>JQMS</ctag>
            </lex>
        </tok>
        <tok doc="d0004.1" class=tok from='1.1.3.2.2.1\271'>
            <orth>de</orth>
            <disamb><ctag>D</ctag></disamb>
            <lex>
                <base>de</base>
                <ctag>P</ctag>
            </lex>
        </tok>
        <tok doc="d0004.1" class=tok from='1.1.3.2.2.1\274'>
            <orth>reforma</orth>
            <disamb><ctag>N5FS</ctag></disamb>
            <lex>
                <base>reforma</base>
                <ctag>N5FS</ctag>
            </lex>
        </tok>
    </chunk>
</chunkList>
```

Annex VI

Llista de recursos Internet

Afegim ara les adreces Internet d'aquells recursos que trobem relacionats amb SGML i la codificació de corpus més útils. Aquesta llista no pretén d'esser exhaustiva sinó donar punts de partida interessants i engrescadors.

- CES

Corpus Encoding Standart
<URL http://www.lpl.univ-aix.fr/projects/multext/CES/>

EAGLES Home Page
<URL: http://www.ilc.pi.cnr.it/EAGLES/home.html>

- Eines SGML

The Whirlwind Guide to SGML Tools and Vendors
<URL: http://www.falch.no/~pepper/sgmltool/>

Public SGML software
<URL: http://www.sil.org/sgml/publicSW.html>

NSGMLS (Parser documentos SGML)
<URL http://www.jclark.com/sp.html">

EAGLES subgroup on Tools
<URL: http://www.univ-aix.lpl/projects/eagles/Tools/>

Fred: The SGML Grammar Builder: Automatic DTD creation service
<URL: http://www.oclc.org/fred/>

- Empreses privades

Electronic Book Technologies
<URL: http://www.ebt.com>

SoftQuad
<URL: http://www.sq.com>

Open Text Corp
<URL: http://www.opentext.com>

- Projectes de compilació de corpus

English-Norwegian Parallel Corpus Project
<URL: http://www.hd.uib.no/enpc.html>

The Lingua Parallel Concordancing Project
<URL: http://www.loria.fr/exterieur/equipe/dialogue/lingua/TT/tt.html>

British National Corpus
<URL: http://info.ox.ac.uk/bnc/>

Victorian Women WrittersProject
<URL: http://www.indiana.edu/~letrs/vwwp/>

- SGML i estàndards relacionats

SGML Web Page
<URL: http://www.sil.org/sgml/sgml.html>

SGML Open home page
<URL: http://www.sgmlopen.org/>

HyTime: ISO 10744 Hypermedia/Time-based Structuring Language
<URL: http://www.sil.org/sgml/gen-apps.html#hytime>

Yahoo
<URL: http://www.yahoo.com/Computers/Languages/SGML>

- TEI

Text Encoding Initiative
<URL: http://www-tei.uic.edu/orgs/tei/>

The Electronic Text Center Introduction to TEI and Guide to Document Preparation
<URL: http://www.lib.virginia.edu/etext/tei/uvatei3.html>

TEI Lite DTD
<URL: http://www.lib.virginia.edu/etext/tei/teilite-dtd.html>

Annex VII

Llista d'abreviacions

ASCII	*American Stantard Code for Information Interchange*
CALS	*Computer-aided Acquisition and Logistic Support*
CES	*Corpus Encoding Standard*
CREA	*Corpus de Referencia del Español Actual*
CT	*Corpus Tècnic Especialitzat*
DTD	*Document Type Definition*
EAGLES	*Expert Advisory Group on Language Engineering Standards*
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*
HyTime	*Hypermedia/Time based Document Structuring Language*)
ISO	*International Standard Organization*
IULA	*Institut Universitari de Lingüística Aplicada*
LOB	*London Oslo/Bergen*
SGML	*Standard Generalized Markup Language*
TEI	*Text Encoding Initiative*
WWW	*World Wide Web*
Unicode	*Universal Multiple-Octet Coded Character Set*
UPF	*Universitat Pompeu Fabra*
URL	*Universal Resource Locators*