EL CORPUS DE L'IULA: DESCRIPCIÓ

C. Bach, R. Saurí, J. Vivaldi, M.T. Cabré

Sèrie Informes, 17

Barcelona
Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
1997

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

Primera edició: 1997

La Rambla, 30-32
08002 Barcelona

Dipòsit legal: B-34.227-2002

Índex

1. El marc: l'Institut Universitari de Lingüística Aplicada (IULA) 1
2. Utilitat dels corpus textuals 2
3. El corpus de l'IULA 4
3.1 Objectius 4
3.2 Disseny 4
3.3 Fase prèvia: Estructuració dels camps conceptuals i de la tipologia textual 5
3.4 Etapes de processament 6
3.4.1 Fase de tractament dels materials 6
3.4.1.1 Selecció, localització i adquisició dels materials 6
3.4.1.2 Automatització dels textos que no es troben en suport electrònic 7
3.4.1.3 Marcatge estructural 7
3.4.2 Fase de processament lingüístic 8
3.4.2.1 Preprocés 9
3.4.2.2 Anàlisi morfològica 9
3.4.2.3 Desambiguació 10
3.5 Explotació 13
4. Bibliografia 15
Annexos 16
Annex 1: Fragment de document no processat 16
Annex 2: Inserció de marques estructurals: procediment manual 17
Annex 3: Fragment d'un document del corpus tècnic etiquetat estructuralment 18
Annex 4: Fragment de document preprocessat 19
Annex 5: Fragment de document analitzat morfològicament 20
Annex 6: Fragment de document castellà analitzat morfològicament i desambiguat lingüísticament 25
Annex 7: Fragment d'un document d'entrada al desambiguador estadístic per al català 29
Annex 8: Fragment de document desambiguat estadísticament 34
Annex 9: Eina principal d'explotació del Corpus Tècnic 39
Annex 10: Estat actual (maig, 1997) 43
Annex 11: Membres del projecte 58

1. El marc: l'Institut Universitari de Lingüística Aplicada (IULA)

L'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra realitza, com correspon als instituts universitaris, activitats en dues vessants ben definides: d'una banda la formació de tercer cicle, i de l'altra la investigació científica i tècnica. És en el marc d'aquesta darrera on s'inscriuen les activitats científiques i professionals dirigides a la cobertura de necessitats lingüístiques d'organismes i de particulars, i a la resolució de problemes lingüístics i de comunicació. Aquestes activitats requereixen un treball tant en recerca fonamental com en la construcció d'eines i recursos relacionats amb les aplicacions del llenguatge.

Els membres de l'IULA constitueixen un equip amb una línia de recerca comuna, la lingüística aplicada al tractament del llenguatge. Aquest equip, identificat amb el nom de *grup LATRAL*, s'organitza internament en cinc unitats: la *Unitat de recerca de Terminologia i Neologia*, ocupada en la investigació en els camps de la morfologia lèxica, la neologia, la descripció dels llenguatges d'especialitat i la terminologia; la *Unitat de recerca de Lexicografia*, centrada en el disseny i la constitució de materials lexicogràfics; la *Unitat de Lingüística Computacional*, dedicada a la representació computacional del coneixement lingüístic (gramàtiques i lèxics computacionals, aplicacions diverses en el camp de la traducció automàtica, recuperació d'informació, etc.)*;* la *Unitat d'Enginyeria Lingüística*, encarregada del disseny i elaboració d'eines informàtiques destinades a l'explotació dels corpus, i la *Unitat de Variació Lingüística*, centrada en la variació i el canvi lingüístic. Cada unitat de recerca està coordinada per un responsable i té adscrits projectes de recerca finançats per organismes públics o empreses privades.

A l'IULA es desenvolupa igualment un projecte de recerca prioritari, el *Projecte Llenguatges Especialitzats*, a l'entorn del qual s'articulen aquestes cinc unitats. Aquest projecte, en el qual hi participen tots els membres i el personal tècnic de l'Institut, a més d'altres col·laboradors externs, té com a objectiu la constitució d'un corpus textual plurilingüe, especialitzat en els camps del dret, l'economia, la medicina, la informàtica i el medi ambient, que faciliti els materials adequats per realitzar recerca lingüística sobre aquests àmbits.

La infraestructura amb què compta l'IULA per al desenvolupament d'aquestes activitats està constituïda per dos entorns operatius diferents: MS-DOS/Windows i UNIX. L'amigabilitat de tractament que ofereix l'entorn Windows, així com la potència de càlcul de l'entorn UNIX, són les característiques més destacables de cada un d'aquests sistemes que en justifiquen la seva elecció. A més d'aquests avantatges, els dos entorns estan connectats en xarxa per tal de poder aprofitar totes les aplicacions informàtiques disponibles per a cada un d'ells.

2. Utilitat dels corpus textuals

Un corpus lingüístic de dades textuals és una col·lecció organitzada de textos seleccionats sobre la base d'uns criteris determinats. Aquests criteris s'estableixen prèviament de manera acurada, per tal que els materials que formin part del corpus siguin representatius del conjunt d'actualitzacions, en un domini concret, d'una llengua -en el cas dels corpus monolingües- o de diverses llengües -en el cas dels corpus multilingües. A partir de l'anàlisi de les dades dels corpus s'intenten inferir les regularitats que regeixen el comportament de cada llengua en el domini escollit.

Es pot establir una tipologia dels corpus lingüístics en funció de la combinació de tres paràmetres:
a) la finalitat per a la qual s'han constituït (corpus amb finalitats generals *versus* corpus amb finalitats específiques)
b) el canal de producció dels textos emmagatzemats (corpus orals *versus* corpus escrits), i
c) el contingut dels corpus (corpus de llengua general *versus* corpus de subllenguatges determinats).

Mentre que els corpus amb finalitats generals constitueixen una font d'informació textual per a diversos usos, els corpus amb finalitats específiques intenten donar resposta a pròposits clarament definits, com ara l'estudi d'aspectes concrets de la gramàtica o del lèxic, l'extracció de dades estadístiques, l'estudi del comportament lingüístic d'una determinada població de parlants, l'anàlisi comparativa de diverses varietats lingüístiques, el desenvolupament i evolució de sistemes de processament del llenguatge o l'elaboració de models estadístics per millorar el rendiment d'aplicacions de reconeixement de la veu. De totes maneres, un corpus amb finalitats específiques pot ser utilitzat amb finalitats diferents a les previstes, sempre que les seves característiques i el seu disseny en permetin la reutilització.

Pel que fa al segon paràmetre de classificació de corpus, el canal de producció dels textos emmagatzemats, és important fer notar que la complexitat de tractament i emmagatzematge dels textos dels corpus està directament relacionada amb el canal per mitjà del qual aquests s'han produït. Així, si el tractament d'un corpus de textos escrits és una tasca complexa, el tractament dels corpus orals comporta encara més dificultat.

Per últim, quant al contingut dels corpus, cal destacar que un corpus que abasti la llengua general també pot utilitzar-se per a usos específics i que un corpus d'un determinat subllenguatge pot utilitzar-se també amb finalitats generals.

A tall d'exemple, es presenten a continuació les finalitats que es contemplen en alguns dels corpus dels àmbits del català i l'espanyol:

- Corpus de llengua general amb finalitats generals:

Corpus textual informatitzat de la llengua catalana (CTILC), Institut d'Estudis Catalans. Es concep com un corpus de referència general del català, amb textos de tots els àmbits.

- Corpus de llengua general amb finalitats específiques:

Corpus del proyecto SISCOOR, Universitat Politècnica de València. Aquest corpus conté textos científics i tècnics, i textos orals dels mitjans de comunicació. Ha estat concebut amb l'objectiu d'abordar diversos problemes sintàctics.

- Corpus d'un subllenguatge amb finalitats generals:

Corpus Escrit del Català Actual (CECA), Universitat de Barcelona. Està constituït per textos procedents de l'àmbit periodístic, però no se circumscriu a cap finalitat específica.

- Corpus d'un subllenguatge amb finalitats específiques:

Corpus de Lengua escrita por aspirantes a estudios universitarios, P.A.A.U. junio (1992), Universitat Pompeu Fabra. A través d'aquest corpus és vol assolir la caracterització del text acadèmic escrit per estudiants que han cursat l'ensenyament secundari i, alhora, estudiar la llengua escrita formal no destinada a ser publicada.

L'explotació de dades d'un corpus és útil en tres àmbits diferents: en el terreny de la recerca, en l'activitat professional i en el camp de l'ensenyament. És en el primer d'aquests àmbits, la recerca, on la utilització de corpus en suport automàtic està més difosa, per raó de l'enorme ventall de possibilitats que ofereix el fet de disposar d'una gran quantitat de dades lingüístiques i de poder-les manejar de forma versàtil i eficient. A més, els estudis que es porten a terme en aquest camp incideixen directament en la creació de productes d'ús generalitzat, com ara els processadors de textos o els diccionaris electrònics.

En l'activitat professional, els corpus constitueixen sistemes d'ajut que permeten, entre altres possibilitats, la consulta de mots de la llengua general o de termes d'una determinada especialitat, localitzant-los en els diferents contextos en què poden aparèixer així com facilitant els seus equivalents en d'altres llengües. Un corpus multilingüe pot ser, doncs, una eina de gran ajut en el camp de la traducció en general i, en cas d'estar circumscrit a un subllenguatge determinat, en el camp de la traducció de textos d'especialitat.

En darrer lloc, pel que fa a l'àmbit de l'ensenyament, els corpus poden utilitzar-se en activitats dins el camp de la filologia o relacionades amb l'aprenentatge de llengües. Igualment, un corpus especialitzat concebut com a sistema d'ajut i de consulta és una forma pràctica i senzilla d'introduir-se en la utilització de la terminologia d'una determinada especialitat.

3. El corpus de l'IULA[1]

El projecte de constitució del corpus *Llenguatges Especialitzats* de l'IULA, que està finançat per la CIRIT (CS93-4.009) i per la Universitat Pompeu Fabra, aglutina la totalitat dels investigadors de l'Institut Universitari de Lingüística Aplicada, que en el seu conjunt constitueixen el grup LATRAL, grup de recerca reconegut com a grup consolidat en el Pla de Recerca de Catalunya.

3.1 Objectius

El corpus de l'IULA es proposa de facilitar tres objectius principals:

a) *L'estudi de la llengua escrita dels textos especialitzats*. Aquest estudi, útil per a l'ensenyament, pot ser també una eina de suport en l'àmbit de la filologia, de l'aprenentatge d'idiomes o de la traducció especialitzada. Igualment, és d'interès en l'activitat professional ja que permet consultes terminològiques específiques en diferents contextos.

b) *La creació d'una eina de suport a les línies d'investigació de l'IULA*,[2] per tal com és capaç de proporcionar mostres àmplies de material lingüístic. Aquesta eina és la base de moltes recerques: a partir del Corpus de l'IULA es pot treballar en la paral·lelització de textos de diferents llengües o extreure índexs de paraules amb la seva freqüència d'aparició així com concordances basades en informació morfosintàctica, on es mostren les paraules en el seu context, etc. El corpus suposa també una eina de gran utilitat per a l'estudi dels canvis en el llenguatge en diferents registres o en l'anàlisi de determinats fenòmens sintàctics.

c) *El desenvolupament d'aplicacions lingüístiques útils per a la investigació*, com ara la detecció de terminologia, la identificació de neologia, la identificació dels diferents significats lexicogràfics d'una mateixa paraula, l'elaboració de tesaurus o de glossaris terminològics i fraseològics. D'altres aplicacions possibles relacionades amb la lingüística computacional o l'enginyeria lingüística són el desenvolupament de models estadístics del llenguatge, de correctors ortogràfics i d'analitzadors sintàctics.

3.2 Disseny

Un corpus no és un conjunt de textos recollits sense cap mena d'ordre sinó que ha d'estar estructurat a partir d'uns criteris clars. Així, el disseny del corpus *Llenguatges Especialitzats* és suficientment flexible perquè pugui adaptar-se a les diferents necessitats dels membres investigadors de l'Institut Universitari de Lingüística Aplicada i també a les necessitats que puguin tenir els potencials usuaris d'aquest corpus.

[1]Per obtenir més informació d'aquest projecte consulteu el següent URL: http://www.iula.upf.es/corpus/corpus.htm
[2]Vegeu el primer apartat d'aquest article.

Per al Corpus de l'IULA els criteris s'han establert tenint en compte, d'una banda, els objectius assenyalats en l'apartat anterior; de l'altra, la necessària reutilització del corpus; i per últim, considerant la relació dels membres de l'IULA amb la Facultat de Traducció i Interpretació d'aquesta universitat, on la traducció té un enfoc especialitzat en les àrees que cobreix el corpus.

Els criteris que s'apliquen són els següents:

- El corpus de l'IULA comprèn textos escrits especialitzats en els àmbits del dret, l'economia, la medicina, el medi ambient i la informàtica.
- El corpus de l'IULA és multilingüe i, en la mesura que és possible, conté textos paral·lels en diferents llengües. Les llengües del corpus són català, castellà, anglès, francès i alemany.
- El corpus està marcat a partir de l'estàndard SGML (ISO 8879). Particularment, el projecte segueix les recomenacions de la iniciativa EAGLES: tots els documents estan marcats segons aquest estàndard internacional.[3]
- Seguint la prosposta de J. Sinclair (1992), el corpus pretén assolir un grau òptim de representativitat recollint un nombre de paraules suficientment elevat de cada àrea d'especialitat, per tal que sigui possible estudiar-ne les característiques pròpies.

3.3 Fase prèvia: Estructuració dels camps conceptuals i de la tipologia textual

Atès que la finalitat del corpus és servir de material de base per a estudis sobre el comportament lingüístic real en cada un dels dominis d'especialitat seleccionats, és imprescindible que el corpus sigui una base fiable d'aquests subllenguatges i, doncs, que sigui representatiu de la població de textos que els correspon. És per això que s'ha buscat l'establiment de tipologies que permetin classificar els materials, tant pel que fa a les seves característiques textuals com en funció de l'estructuració interna de cada un dels diferents dominis d'especialitat.

Per a l'elaboració de la tipologia textual s'ha buscat una classificació molt general que distingeix els textos segons la finalitat a la qual estan adreçats. Es diferencien tres grans conjunts: el dels *textos normatius*, que engloba els materials de caràcter legislatiu i regulador de cada domini d'especialitat; el dels *textos instrumentals*, com ara diccionaris, vocabularis i glossaris; i, en darrer lloc, el dels *textos propis de l'àrea*, ja siguin teòrics o relatius a la pràctica professional.

Quant a l'estructuració dels diferents camps conceptuals, s'ha comptat amb la col·laboració d'especialistes de cada una de les àrees escollides. En el camp del dret hi han intervingut Xavier Bernardí, Ester Cabrera i Carles Duarte; en el d'economia Miquel Centelles i Vicent Ortun. Horacio Rodríguez i Jordi Vivaldi han participat en l'organització de l'àrea d'informàtica; Pau Serra en la de medi ambient, i Antoni Valero en la de medicina. Aquests especialistes s'han encarregat d'elaborar l'arbre de camp que permet classificar els textos del corpus en les diferents subàrees que conformen cada una de les disciplines i de seleccionar els documents més representatius de la matèria que constituiran el recull textual de l'àrea. Vegeu l'annex 10 on es presenta l'arbre de camp que els especialistes han establert per a cada àrea.

[3]Vegeu *Papers de l'IULA*, sèrie informes, 1, pàg. 5.

Combinant les etiquetes resultants de l'arbre de camp de cada àrea amb les tres etiquetes de la tipologia textual (textos normatius, textos instrumentals i textos propis de l'àrea), s'arriba a una classificació creuada dels materials que fa possible un control força rigorós de l'equilibri de les dades en el corpus en funció dels textos processats.

3.4 Etapes de processament

Aquest apartat s'estructura en dues parts clarament diferenciades: a 3.4.1 es presenta la fase de tractament dels materials i a 3.4.2 s'explica la fase següent, on es fa un processament lingüístic dels documents prèviament tractats.

3.4.1 Fase de tractament dels materials

Aquesta etapa de treball es desenvolupa en quatre subetapes diferents:
1) la fase de selecció, localització i adquisició dels materials que han de conformar el corpus
2) la fase d'automatització dels textos que no es troben en suport magnètic
3) la tasca d'etiquetatge estructural.

3.4.1.1 Selecció, localització i adquisició dels materials

En aquesta fase se seleccionen, amb l'ajuda dels especialistes, els materials més representatius de cada àrea que han de passar a formar part del corpus. Per portar a terme aquesta tasca es facilita als especialistes un llistat exhaustiu de documents elaborat a partir de catàlegs de publicacions d'institucions públiques i d'editorials que produeixen obres de cada disciplina específica del corpus. El llistat conté un primer bloc de bibliografia procedent de fonts prioritzades per la seva facilitat d'accés. El segon bloc està constituït per fonts també interessants, com documents (llibres i revistes) existents al catàleg de la biblioteca de la UPF o la bibliografia extreta de programes universitaris.

El grau de prioritització de les fonts d'origen de la bibliografia, però, no és l'únic paràmetre de classificació del llistat que es passa als especialistes, sinó que cada una de les fonts s'estructura internament seguint la tipologia textual comentada anteriorment: documents normatius, documents propis de l'àrea i documents instrumentals. A partir del llistat de treball, l'especialista indica, per a cada obra que s'hi detalla, si és *essencial*, *important* o *complementària*, ponderant-la amb un 1, un 2 o un 3 respectivament.

Un cop aquesta tasca acomplerta, correspon al servei de documentació de l'IULA de localitzar i adquirir els materials. Els documents indicats pels especialistes es poden obtenir per mitjà de dues vies diferents: bé en suport electrònic, bé en suport paper. Els documents obtinguts per aquesta segona via són sotmesos a un procés d'escanneig i reconeixement, procés que suposa un tractament llarg. D'aquí que es prioritzin, sempre que és possible, els documents en format electrònic per sobre de les versions en paper. Vegeu a l'annex 1 un exemple d'un fragment d'un document seleccionat per al corpus que encara no ha estat processat.

3.4.1.2 Automatització dels textos que no es troben en suport electrònic

Hi ha, però, una quantitat important de documents que no es troben en suport electrònic. Això converteix en indispensable la utilització de l'escànner com a sistema d'adquisició de materials, encara que això suposi un cost considerable d'hores de treball. El procediment per a la incorporació dels textos a través de l'escànner segueix les següents fases:

a) Selecció de les porcions o mostres del document que s'han de reconèixer -que cal que siguin d'una extensió mitjana d'entre 3000 i 4000 mots.[4]
b) Elaboració de les fotocòpies corresponents a les dades d'identificació de l'obra (títol, autor, ISBN,...) i del text d'aquesta que s'incorpora al corpus.
c) Procés d'escanneig i reconeixement textual.
d) Correcció ortogràfica del document utilitzant correctors automàtics.

3.4.1.3 Marcatge estructural

Un cop realitzat el reconeixement dels documents impresos a través de l'escànner, o en el seu cas, un cop obtingut el document en format electrònic, es realitza el procés de marcatge estructural en dues fases i dos entorns diferents:

a) La primera fase de marcatge estructural es duu a terme en l'entorn *Windows*, treballant en el processador de textos *WordPerfect*, des de les diferents estacions de treball de cada una de les persones que constitueixen l'equip del projecte *Corpus*.

En aquesta etapa es marquen les divisions que componen el document, els títols i subtítols, les característiques tipogràfiques (ús de la negreta, de la cursiva, etc.), les notes a peu de pàgina o a final de document, les llistes, les taules, les figures, les fórmules, les comandes, i els fragments redactats en una llengua diferent a la del document. Totes aquestes marques s'insereixen utilitzant una sèrie de *macros,* definides dins el sistema *WordPerfect,* que agiliten la feina. És també en aquesta etapa que es substitueixen per entitats SGML caràcters especials com els següents: les vocals accentuades o amb dièresi, *ç* (la ce trencada), *l·l* (l'ela geminada), *ñ* (l'enya espanyola), etc. Vegeu, com a exemple, l'annex 2, on es mostra l'aspecte de la pantalla de treball en el decurs de la fase de marcatge estructural. Concretament, és possible observar algunes marques en format SGML que ja han estat inserides al fragment, així com la caixa de diàleg de la *macro* que assigna una marca estructural a les notes del text.

b) La segona fase del marcatge estructural es desenvolupa en l'entorn *MS-DOS*. En una primera etapa s'incorporen les marques d'inici i de final de paràgraf i frase (* , *; i *<s>, </s>*). Aquest procés es realitza de manera automàtica a partir d'una operació de cerca de totes les seqüències formades per un punt gràfic i un espai. En cas que aquesta seqüència estigui continuada

[4]Amb aquesta segmentació, tot i que es perd la possibilitat de tractar documents sencers, s'intenta obviar els problemes de drets d'autor.

per una lletra en majúscula, s'interpretarà com una frontera entre dues frases, i s'inseriran les marques </s><s>. En cas que la seqüència 'punt + espai' continuï amb un salt de línia, s'inseriran les marques *</s> <s>* indicant canvi de paràgraf.

Per tal que les marques inserides senyalin únicament els límits entre frases i paràgrafs, cal diferenciar els punts gràfics d'acabament d'unitat frasal dels que presenten altres funcions, com passa, per exemple, en contextos com topònims (*St. Andreu*), antropònims (*O. Amat i Salas*), sigles (*U. P. F.*) i abreviatures en general (*S. XIX*). Aquests casos són detectats en l'operació de prepocessament mateix i, doncs, no reben el tractament que es comentava més amunt.

L'annex 3 presenta un fragment de document del corpus completament etiquetat pel que fa a les marques estructurals: divisions (* *), títols (*<head>*), inici i final de frase o paràgraf (*<s>, </s>, , *) i entitats que substitueixen els caràcters especials (*ó à &ccedilt;*).

Per acabar, s'efectua el procés d'anàlisi sintàctica del resultat de l'operació de marcatge, per tal de detectar i corregir els problemes i errors generats en aquesta tasca. Aquesta operació de comprovació es realitza a partir d'un document addicional, anomenat DTD (*Document Type Definition*), en el qual s'especifica quins són els elements i les construccions permeses per al marcatge lingüístic del corpus (*model de contingut*). Per a una explicació més detallada d'aquest aspecte, vegeu Vivaldi, J. *et al.* (1996).

3.4.2 Fase de processament lingüístic

Aquesta etapa del processament del corpus consta de tres fases diferenciades, a) el preprocés, b) l'anàlisi morfològica, i c) la desambiguació.

3.4.2.1 Preprocés

La finalitat d'aquesta fase de treball és tractar les unitats lèxiques contingudes en els diferents documents marcats estructuralment, que, per les seves característiques lingüístiques, admeten una detecció automàtica prèvia a l'anàlisi morfològica: locucions, dates, noms propis, abreviacions, identificadors, etc.

El preprocés encapsula aquestes unitats per tal d'agilitar la fase de processament lingüístic, que ja no analitzarà aquestes unitats marcades pel preprocés. En el cas dels noms propis, de les locucions i dels números, el preprocés els assigna també l'etiqueta morfosintàctica que els correspon. Vegeu l'annex 4, en el qual, a més de les marques estructurals, s'han afegit també les pròpies de la fase de preprocés: <*loc*> per a les locucions, <*name*> per als noms propis i <*num*> per a les xifres, entre d'altres.

Totes aquelles unitats lèxiques que no puguin ser tractades en la fase de preprocés passaran a la següent fase de treball, on se'ls aplicarà l'analitzador morfològic.

3.4.2.2 Anàlisi morfològica

En aquesta fase els documents que prèviament han estat marcats estructuralment i preprocessats, es lematitzen i s'etiqueten gramaticalment mitjançant els analitzadors morfològics de què l'IULA disposa. Les paraules dels documents que es processen apareixen amb tots els lemes i etiquetes gramaticals possibles per la qual cosa, en moltes ocasions, s'assigna més d'un lema i etiqueta a la mateixa paraula. Així, a mode d'exemple, al mot *en* se li assignen en aquesta etapa tres lemes i tres etiquetes diferents: *en* [lema *en*, etiqueta *P* (preposició)], *en* [lema *en*, etiqueta *AMS* (determinant masculí singular] i *en* [lema *en*, etiqueta *REE7---* (pronom personal feble, cas pendent d'especificar -acusatiu/oblic-)].

Per a l'assignació de lemes i d'etiquetes es parteix dels diccionaris amb què compta cada eina, dels criteris de lematització establerts per al corpus l'IULA[5] i dels corresponents etiquetaris, confeccionats també en el marc del projecte i que a grans línies segueixen les directrius del projecte EAGLES.[6]

[5]Document intern del projecte Corpus c0021.doc.
[6]Vegeu Morel J. *et al.* (1997) "El corpus de l'IULA: Etiquetaris", *Papers de l'IULA,* sèrie informes, 18.

Es presenten a continuació les eines de marcatge morfològic utilitzades per al català i el castellà:

a) Català
La lematització i l'anàlisi morfològica dels documents catalans es fa a partir d'un analitzador que extreu la informació necessària per a l'etiquetatge i la lematització d'un mòdul integrat de tres diccionaris: el *Diccionari de la Llengua Catalana* (1983) d'Enciclopèdia Catalana, que serveix de diccionari de base, el *Diccionari de la Llengua Catalana* (1993) i el *Diccionari de la Llengua Catalana* (1995) de l'Institut d'Estudis Catalans.

b) Castellà
Per a l'assignació de lemes i informació gramatical a les paraules de cada document en castellà, s'ha partit del *Diccionario Actual de la Lengua Española (DALE)*, cedit per l'empresa Vox-Bibliograf exclusivament per a finalitats de recerca, que conté 88.442 lemes amb la corresponent informació categorial.

A l'annex 5 es presenta un fragment del corpus analitzat morfològicament. Després de l'anàlisi morfològica i la lematització, resten encara algunes paraules a les quals no s'ha assignat cap lema ni etiqueta per tal com no han estat reconegudes pel diccionari (neologismes, manlleus, errors del document original, errors derivats del procés d'escanneig, etc.). A aquestes paraules se'ls aplica un programa previ a l'etapa de desambiguació estadística, que, sempre que és possible, els assigna un lema i una categoria en funció de la seva terminació: d'aquesta manera, per exemple, una paraula no reconeguda acabada en *-itzar* serà etiquetada com a *VI* (verb infinitiu) i se li assignarà com a lema la forma d'infinitiu. És també aquest el moment en què s'assigna lema i etiqueta a les seqüències anteriorment marcades pel preprocés que havien quedat sense etiquetar.

3.4.2.3 Desambiguació

La següent fase del processament lingüístic és la de desambiguació dels mots "sobreetiquetats", en la qual es fa la tria d'una etiqueta gramatical entre totes les proposades. La desambiguació es realitza mitjançant dues eines en curs que es complementen: un desambiguador lingüístic, que tria l'etiqueta correcta mitjançant regles de caràcter lingüístic,[7] i un desambiguador estadístic, que resol les ambigüitats que encara queden després d'haver passat el desambiguador lingüístic. Vegeu l'annex 6 on es presenta una mostra d'un document en castellà desambiguat lingüísticament, l'annex 7 on es presenta un fragment de document que serveix d'entrada al desambiguador estadístic per al català i l'annex 8, en el qual apareix la mateixa mostra desambiguada al 100% després d'haver passat el desambiguador estadístic.

Un cop el document està totalment desambiguat, un programa indexa totes les paraules amb la informació del lema i l'etiqueta que duen associades. Aquestes informacions ja indexades passen a

[7]Mentre que la desambiguació lingüística per al castellà es fa a través d'un nombre elevat de regles lingüístiques, en el cas de la desambiguació lingüística per al català només ha calgut d'aplicar-hi un paquet molt reduït de regles en el postprocés.

incoporar-se a una base de dades a partir de la qual es realitzarà l'explotació del corpus. Vegeu l'apartat 3.5 d'aquest article.

El projecte preveu d'entrar ja des d'ara, però de manera progressiva, en altres estadis del processament lingüístic del corpus:

d) processament sintàctic
e) processament semàntic
f) processament pragmàtic

En el gràfic que segueix, a mode de resum, es veu un esquema de les diferents etapes de treball del corpus de l'IULA, des de la fase prèvia de selecció dels documents per part dels especialistes fins a l'etapa d'explotació que es presenta en el proper apartat.

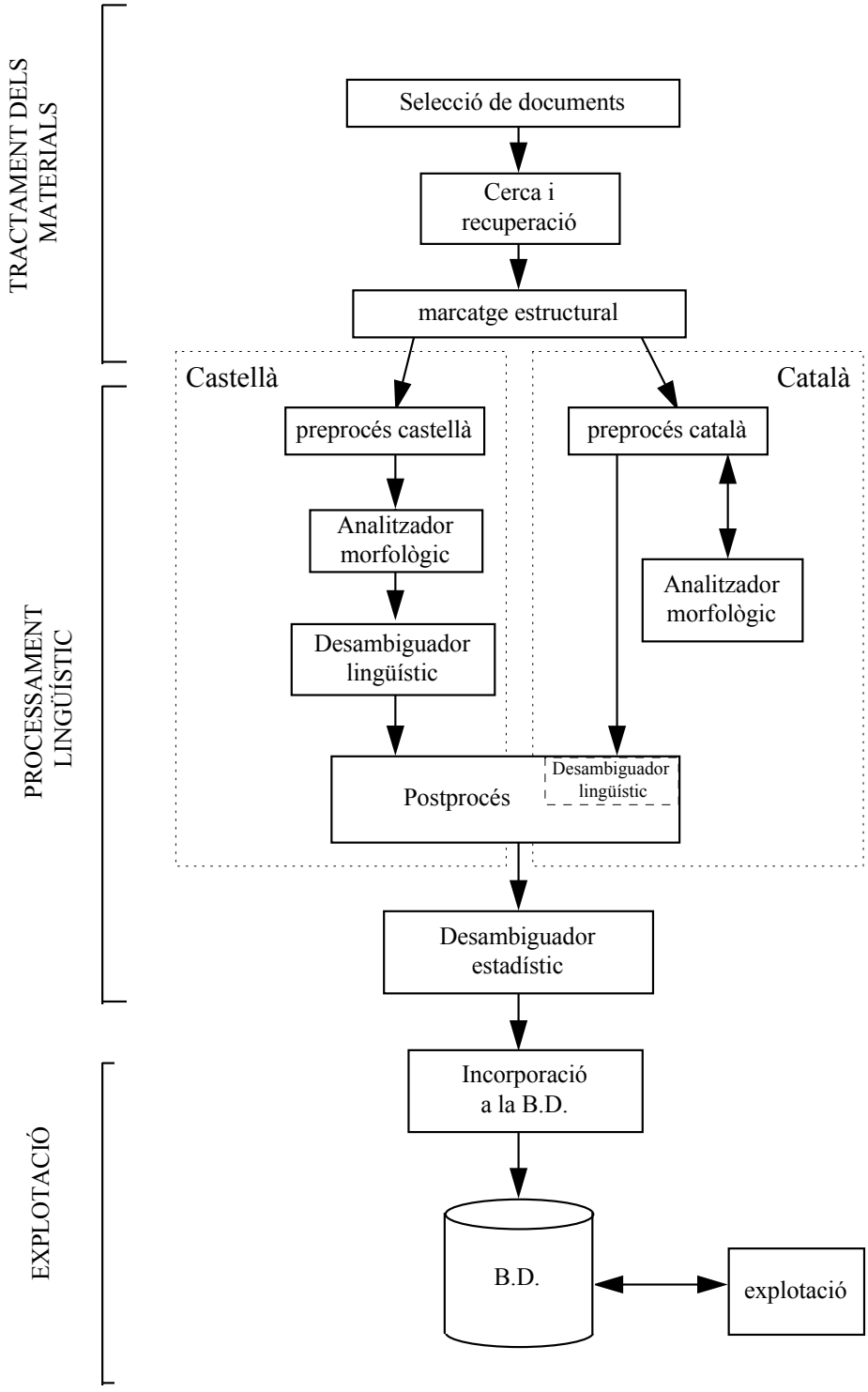


Fig. 2: Etapes de treball del corpus de l'IULA

3.5 Explotació

En aquest àmbit es treballa en la constitució d'eines que facilitin l'extracció de tota la informació necessària per a les diferents investigacions dels membres de l'IULA o dels seus col·laboradors externs. Així, amb l'aplicació del desambiguador estadístic als documents del corpus, i un cop els mots han estat incorporats a la base de dades que conté els mots indexats amb la seva etiqueta i el seu lema, es poden iniciar les recerques, des d'un entorn Windows o des d'un entorn Unix. Les dades del corpus han de servir per al màxim nombre d'estudis possible, per la qual cosa el sistema d'explotació tipus pretén cobrir uns mínims que siguin comuns per a tota recerca.

Extracció de concordances:
- càlcul de freqüències absolutes i relatives,
- selecció de grups de documents a partir dels paràmetres d'àrea, tipologia o traducció/original,
- recerques sobre seqüències o segments de paraules, i
- cerca de patrons en funció de la forma, lema i categoria (amb opcionalitat i negació)

Vegeu l'annex 9.a. com a exemple de les pantalles de diàleg per realitzar algunes d'aquestes consultes.

En aquests moments s'està treballant en la presentació de les dades segons les consultes efectuades i en la generació de diferents presentacions per pantalla. L'usuari podrà escollir a través d'un sistema de diàleg el nivell de presentació dels resultasts de les cerques:

(a) només les paraules o seqüències trobades (es mostra una simple llista de les paraules),
(b) les seqüències amb un context d'amplada seleccionable,
(c) les seqüències amb el context sencer (tot l'element SGML que conté la seqüència trobada),
(d) document sencer amb les paraules o seqüències de paraules trobades ressaltades,
(e) etc.

Cadascun d'aquests formats de presentació de les consultes pot contenir alhora diferents tipus d'informació, que es pot mostrar directament o en un format hipertextual amb finestres desplegables o amb la inclusió del marcatge estructural del text. L'annex 9.d. ofereix un exemple de pantalla de presentació de resultats. Finalment, els resultats de les cerques es poden guardar en el seu format original per ser recuperats en una sessió posterior, constituint d'aquesta manera un arxiu propi per a cada estudi, o poden ser tractats de manera similar a les presentacions i guardats per ser utilitzats per a bases de dades, processadors de textos o d'altres aplicacions.

En un futur immediat es preveu disposar d'eines que permetran d'altres aplicacions com la detecció de terminologia o la paral·lelització de textos en diferents llengües.

En el gràfic que segueix es presenten, a mode de resum, les diferents vies d'explotació del corpus que estan previstes en aquest moment:

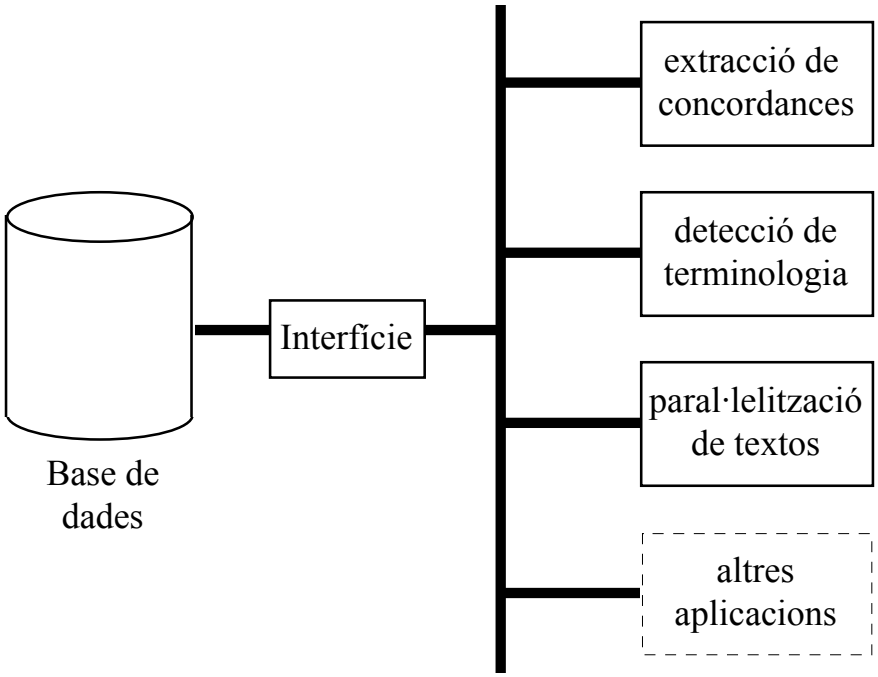


Fig. 3: Vies d'explotació del corpus de l'IULA

Fins aquí s'ha fet referència a les diferents etapes de processament del corpus i s'ha parlat també de les seves possibilitats d'explotació. En l'annex 10 es fa un balanç de l'estat en què es troba el projecte.

4. Bibliografia

BACH, C. (1997) *Criteris de lematització*. Document intern IULA, c0021.doc.

CABRÉ, M. T. (1992, [en premsa]) "Sobre la diversidad y la terminología", III Simposio Iberoamericano de Terminología, S. Millán de la Cogolla, desembre 1992.

_____ (1995) "Corpus para la observación terminológica y neológica. Lexicografía y tecnologías de la lengua: situación y perspectivas de las lenguas románicas", ms., Escuela Internacional de Altos Estudios en Lingüística Aplicada, San Millán de la Cogolla, septiembre de 1995.

Diccionari de la llengua catalana (1993), Barcelona: Enciclopèdia Catalana.

Diccionari de la llengua catalana (1995), Barcelona: Institut d'Estudis Catalans.

INSTITUTO CERVANTES (ed.) (1996) *Informe sobre recursos lingüísticos para el español (II): Corpus escritos y orales disponibles y en desarrollo en españa*, Alcalá de Henares.

MOREL, J.; TORNER, S.; VIVALDI, J. i CABRÉ M.T. (1997) "El corpus de l'IULA: Etiquetaris", *Papers de l'IULA,* sèrie informes, 18, Barcelona: IULA, Universitat Pompeu Fabra.

SINCLAIR, J. (1992) "The automatic analysis of corpora" a SVARTVIK J. (ed) *Directions in corpus linguistics*, Berlin/Nova York: Mouton de Gruyter, 379-497.

VIVALDI, J. (1996) "Proyectos del IULA: Corpus técnico", a FORCADA, V. C.; DE CARRASCO, A. G.; SAGER, J. C. (ed.), *Estudios computacionales del español y el inglés*, Madrid: Instituto Cervantes, 227-241.

VIVALDI, J.; DE YZAGUIRRE, Ll.; SOLÉ, X. i CABRÉ, M. T. (1996) "Marcatge estructural i morfosintàctic del corpus tècnic amb l'estàndard SGML", *Papers de l'IULA,* sèrie informes, 1, Barcelona: IULA, Universitat Pompeu Fabra.

Annexos

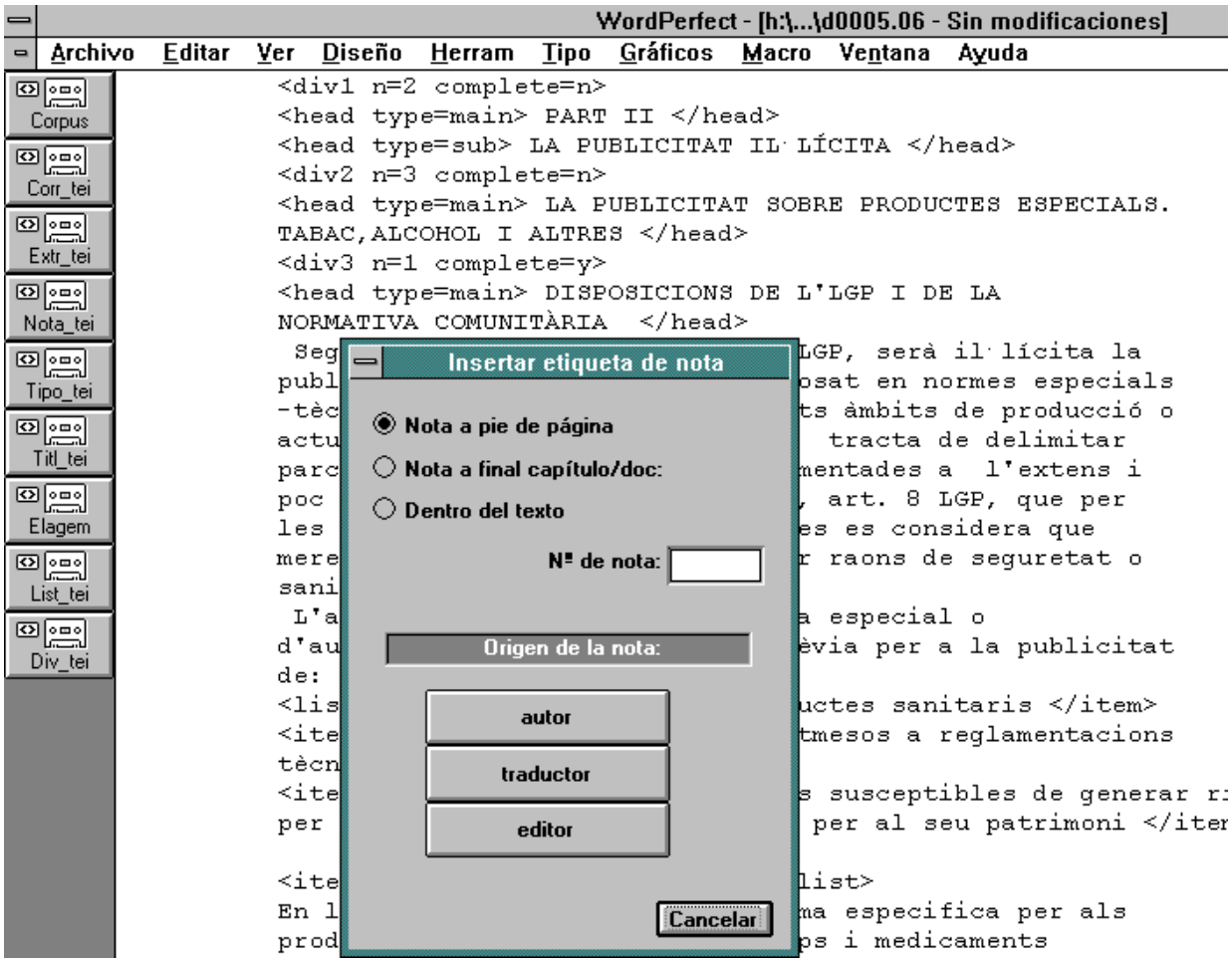
Annex 1: Fragment de document no processat

(fragment de d00024.sgm, mostra 14 -*Dret Mercantil*-)

DRET DE LA NAVEGACIÓ
ESTATUT JURÍDIC DEL VAIXELL I DE L'AERONAU EL VAIXELL

Concepte i naturalesa jurídica
En sentit tècnic parlem de vaixell per referir-nos a qualsevol construcció destinada a la navegació marítima o fluvial. A aquesta idea atenen també, en general, els ordenaments positius moderns en determinar, amb major o menor amplitud, la noció jurídica del vaixell.
En el nostre ordenament legal, l'article 146 del Reglament del Registre Mercantil de 1956, transitòriament vigent (veg. disposició transitòria sisena del Reglament del Registre Mercantil de 29 de desembre de 1989), suplint la llacuna del Codi, estableix, també en aquest sentit, que "es reputaran vaixells, per als efectes del Codi de comerç i d'aquest Reglament, no només les embarcacions destinades a la navegació de cabotatge i altura, sinó també els dics flotants, pontons, dragues, gànguils i qualsevol altre aparell flotant destinat o que pugui destinar-se a serveis de la indústria o comerç marítim o fluvial".

Annex 2: Inserció de marques estructurals: procediment manual



Annex 3: Fragment d'un document del corpus tècnic etiquetat estructuralment

(fragment de d00024.sgm, mostra 14 -*Dret Mercantil*-)
(div1, div2, div3, head, p, s, etc.)

```
<div1 n=6 complete=n>
<head type=main>DRET DE LA NAVEGACI&Oacute;</head>
<div2 n=72>
<head type=main>ESTATUT JUR&Iacute;DIC DEL VAIXELL I DE
L'AERONAU</head>
<div3 n=1>
<head type=main>EL VAIXELL</head>
<div4 n=1.107>
<head type=main>Concepte i naturalesa jur&iacute;dica</head>
<p><s>En sentit t&egrave;cnic parlem de vaixell per referir-nos
a qualsevol construcci&oacute; destinada a la navegaci&oacute;
mar&iacute;tima o fluvial.</s><s>A aquesta idea atenen
tamb&eacute;, en general, els ordenaments positius moderns en
determinar, amb major o menor amplitud, la noci&oacute;
jur&iacute;dica del vaixell.</s></p>
<p><s>En el nostre ordenament legal, l'article 146 del
Reglament del Registre Mercantil de 1956,
transit&ograve;riament vigent (veg. disposici&oacute;
transit&ograve;ria sisena del Reglament del Registre Mercantil
de 29 de desembre de 1989), suplint la llacuna del Codi,
estableix, tamb&eacute; en aquest sentit, que "es reputaran
vaixells, per als efectes del Codi de comer&ccedil; i d'aquest
Reglament, no nom&eacute;s les embarcacions destinades a la
navegaci&oacute; de cabotatge i altura, sin&oacute;
tamb&eacute; els dics flotants, pontons, dragues,
g&agrave;nguils i qualsevol altre aparell flotant destinat o
que pugui destinar-se a serveis de la ind&uacute;stria o
comer&ccedil; mar&iacute;tim o fluvial".</s></p>
```

Annex 4: Fragment de document preprocessat

(fragment de d00024.sgm, mostra 14 -*Dret Mercantil*-)
(loc, name, date, num, etc.)

```
<div1 n=6 complete=n>
<head type=main>DRET DE LA NAVEGACI&Oacute;</head>
<div2 n=72>
<head type=main>ESTATUT JUR&Iacute;DIC DEL VAIXELL I DE
L'AERONAU
</head>
<div3 n=1>
<head type=main>EL VAIXELL</head>
<div4 n=1.107>
<head type=main>Concepte i naturalesa jur&iacute;dica</head>
<p><s>En sentit t&egrave;cnic parlem de vaixell per referir-nos
a qualsevol construcci&oacute; destinada a la navegaci&oacute;
mar&iacute;tima o fluvial.</s><s>A aquesta idea atenen
tamb&eacute;, <loc pos='D'>en general</loc>, els ordenaments
positius moderns en determinar, amb major o menor amplitud, la
noci&oacute; jur&iacute;dica
del vaixell.</s></p>
<p><s>En el nostre ordenament legal, l'article <num
pos='X'>146</num> del <name>Reglament del Registre
Mercantil</name> de <num pos='X'>1956</num>,
transit&ograve;riament vigent (<abbr>veg.</abbr>
disposici&oacute; transit&ograve;ria sisena del <name>Reglament
del Registre Mercantil</name> de <date ISO8601='12/29/1989'>29
de desembre de 1989</date>), suplint la llacuna del
<name>Codi</name>, estableix, tamb&eacute; en aquest sentit,
que "es reputaran vaixells, per als efectes del
<name>Codi</name> de comer&ccedil; i d'aquest
<name>Reglament</name>, no nom&eacute;s les embarcacions
destinades a la navegaci&oacute; de cabotatge i altura,
sin&oacute; tamb&eacute; els dics flotants, pontons, dragues,
g&agrave;nguils i qualsevol altre aparell flotant destinat o
que pugui destinar-se a serveis de la ind&uacute;stria o
comer&ccedil; mar&iacute;tim o fluvial".</s>
```

Annex 5: Fragment de document analitzat morfològicament

(fragment de d00024.sgm, mostra 14 -*Dret Mercantil*-)

```
##   TAG  <div1 n=6 complete=n>
##   TAG  <head type=main>
1    TOK  DRET BOS  dret\JQ--MS|dret\N5-MS
2    TOK  DE        de\P
3    TOK  LA        el\AFS|pr\REEC3FS
4    TOK  NAVEGACI&Oacute;    EOS  navegaci&oacute;\N5-FS
##   TAG  </head>
##   TAG  <div2 n=72>
##   TAG  <head type=main>
5    TOK  ESTATUT   BOS  estatut\N5-MS
6    TOK  JUR&Iacute;DIC      jur&iacute;dic\JQ--MS
7    TOK  DE        de\P
8    PGR  L         el\AMS
9    TOK  VAIXELL       vaixell\N5-MS
10   TOK  I         i\C
11   TOK  DE        de\P
12   TOK  L         el\A6S|pr\REE636S
##-  DLI  '         =\DELIM
13   PGR  AERONAU   EOS  aeronau\N5-FS
##   TAG  </head>
##   TAG  <div3 n=1>
##   TAG  <head type=main>
14   TOK  EL   BOS  el\AMS|pr\REEC3MS
15   TOK  VAIXELL   EOS  vaixell\N5-MS
##   TAG  </head>
##   TAG  <div4 n=1.107>
##   TAG  <head type=main>
16   TOK  Concepte  BOS  concepte\N5-MS
17   TOK  i         i\C
18   TOK  naturalesa          naturalesa\N5-FS
19   TOK  jur&iacute;dica     EOS  jur&iacute;dic\JQ--FS
##   TAG  </head>
##   TAG  <p>
##   TAG  <s>
20   TOK  En   BOS  en\P|en\AMS|pr\REE7---
21   TOK  sentit         sentir\HMS|sentit\N5-MS
22   TOK  t&egrave;cnic       t&egrave;cnic\JQ--
MS|t&egrave;cnic\N5-MS
23   TOK  parlem         parlar\V7R1P-
24   TOK  de        de\P
25   TOK  vaixell        vaixell\N5-MS
26   TOK  per       per\P
27   TOK  referir        referir\VI----
##-  DLI  -         =\DELIM
```

```
28   PGR  nos       pr\REE616P
29   TOK  a         a\P
30   TOK  qualsevol      qualsevol\EN--6S|qualsevol\N5-6S
31   TOK  construcci&oacute;       construcci&oacute;\N5-FS
32   TOK  destinada      destinar\VC--SF
33   TOK  a         a\P
34   TOK  la        el\AFS|pr\REEC3FS
35   TOK  navegaci&oacute;         navegaci&oacute;\N5-FS
36   TOK  mar&iacute;tima          mar&iacute;tim\JQ--FS
37   TOK  o         o\C
38   TOK  fluvial        fluvial\JQ--6S
---  DLD  .    EOS  =\DELS
##   TAG  </s>
##   TAG  <s>
39   TOK  A    BOS  a\P
40   TOK  aquesta        aquest\ED--FS
41   TOK  idea      idea\N5-FS|idear\V8R6S-
42   TOK  atenen         atenir-se\VDR3P-|atendre\VDR3P-
43   TOK  tamb&eacute;         tamb&eacute;\D4
---  DLD  ,         =\DELIM
##   TAG  <loc pos='D'>
44   TOK  en general          =\D4
##   TAG  </loc>
---  DLS  ,         =\DELIM
45   TOK  els       el\AMP|pr\REE636P
46   TOK  ordenaments         ordenament\N5-MP
47   TOK  positius       positiu\N5-MP|positiu\JQ--MP
48   TOK  moderns        modern\JQ--MP
49   TOK  en        en\P|en\AMS|pr\REE7---
50   TOK  determinar          determinar\VI----
---  DLD  ,         =\DELIM
51   TOK  amb       amb\P
52   TOK  major          major\N5-MS|major\JQ--6S
53   TOK  o         o\C
54   TOK  menor          menor\JQ--6S|menor\N5-6S
55   TOK  amplitud       amplitud\N5-FS
---  DLD  ,         =\DELIM
56   TOK  la        el\AFS|pr\REEC3FS
57   TOK  noci&oacute;         noci&oacute;\N5-FS
58   TOK  jur&iacute;dica          jur&iacute;dic\JQ--FS
59   TOK  de        de\P
60   PGR  l         el\AMS
61   TOK  vaixell        vaixell\N5-MS
---  DLD  .    EOS  =\DELS
##   TAG  </s>
##   TAG  </p>
##   TAG  <p>
##   TAG  <s>
```

```
62   TOK  En   BOS  en\P|en\AMS|pr\REE7---
63   TOK  el        el\AMS|pr\REEC3MS
64   TOK  nostre         nostre\EP21MS
65   TOK  ordenament          ordenament\N5-MS
66   TOK  legal          legal\JQ--6S
---  DLD  ,         =\DELIM
67   TOK  l         el\A6S|el\REE636S
##-  DLI  '         =\DELIM
68   PGR  article        article\N5-MS
##   TAG  <num pos='X'>
69   TOK  146       num\X
##   TAG  </num>
70   TOK  de        de\P
71   PGR  l         el\AMS
##   TAG  <name>
72   TOK  Reglament del Registre Mercantil   reglament del
     registre mercantil\N4666
##   TAG  </name>
73   TOK  de        de\P
##   TAG  <num pos='X'>
74   TOK  1956      num\X
##   TAG  </num>
---  DLS  ,         =\DELIM
75   TOK  transit&ograve;riament   transit&ograve;riament\D4
76   TOK  vigent         vigent\JQ--6S
---  DLS  (         =\DELIM
##   TAG  <abbr>
77   TOK  veg.      =\N5-66
##   TAG  </abbr>
78   TOK  disposici&oacute;        disposici&oacute;\N5-FS
79   TOK  transit&ograve;ria       transitori\JQ--FS
80   TOK  sisena         sis&egrave;\EO--FS
81   TOK  de        de\P
82   PGR  l         el\AMS
##   TAG  <name>
83   TOK  Reglament del Registre Mercantil   reglament          del
     registre mercantil\N4666
##   TAG  </name>
84   TOK  de        de\P
##   TAG  <date ISO8601='12/29/1989'>
85   TOK  29 de desembre de 1989       =\T
##   TAG  </date>
---  DLS  ),        =\DELIM
86   TOK  suplint        suplir\VG----
87   TOK  la        el\AFS|pr\REEC3FS
88   TOK  llacuna        llacuna\N5-FS
89   TOK  de        de\P
90   PGR  l         el\AMS
```

```
##   TAG  <name>
91   TOK  Codi      codi\N4666
##   TAG  </name>
---  DLS  ,         =\DELIM
92   TOK  estableix      establir\V8R6S-
---  DLD  ,         =\DELIM
93   TOK  tamb&eacute;        tamb&eacute;\D4
94   TOK  en        en\P|en\AMS|pr\REE7---
95   TOK  aquest         aquest\ED--MS
96   TOK  sentit         sentir\HMS|sentit\N5-MS
---  DLD  ,         =\DELIM
97   TOK  que       que\C|que\RR---66|que\D4
##-  DLE  "         =\DELIM
98   TOK  es        pr\R6-----
99   TOK  reputaran      reputar\VDU3P-
100  TOK  vaixells       vaixell\N5-MP
---  DLD  ,         =\DELIM
101  TOK  per       per\P
102  TOK  a         a\P
103  PGR  ls        el\AMP
104  TOK  efectes        efecte\N5-MP
105  TOK  de        de\P
106  PGR  l         el\AMS
##   TAG  <name>
107  TOK  Codi      codi\N4666
##   TAG  </name>
108  TOK  de        de\P
109  TOK  comer&ccedil;       comer&ccedil;\N5-MS
110  TOK  i         i\C
111  TOK  d         de\P
##-  DLI  '         =\DELIM
112  PGR  aquest         aquest\ED--MS
##   TAG  <name>
113  TOK  Reglament      reglament\N4666
##   TAG  </name>
---  DLS  ,         =\DELIM
114  TOK  no        no\D4|no\N5-MS
115  TOK  nom&eacute;s        nom&eacute;s\D4
116  TOK  les       el\AFP|pr\REEC3FP
117  TOK  embarcacions        embarcaci&oacute;\N5-FP
118  TOK  destinades          destinar\VC--PF
119  TOK  a         a\P
120  TOK  la        el\AFS|pr\REEC3FS
121  TOK  navegaci&oacute;         navegaci&oacute;\N5-FS
122  TOK  de        de\P
123  TOK  cabotatge      cabotatge\N5-MS
124  TOK  i         i\C
125  TOK  altura         altura\N5-FS
```

```
---  DLD  ,          =\DELIM
126  TOK  sin&oacute;          sin&oacute;\C
127  TOK  tamb&eacute;         tamb&eacute;\D4
128  TOK  els        el\AMP|pr\REE636P
129  TOK  dics       dic\N5-MP
130  TOK  flotants        flotant\JQ--6P
---  DLD  ,          =\DELIM
131  TOK  pontons         pont&oacute;\N5-MP
---  DLD  ,          =\DELIM
132  TOK  dragues         draga\N5-FP|dragar\VDR2S-
---  DLD  ,          =\DELIM
133  TOK  g&agrave;nguils          g&agrave;nguil\N5-MP
134  TOK  i          i\C
135  TOK  qualsevol       qualsevol\EN--6S|qualsevol\N5-6S
136  TOK  altre           altre\EN--MS
137  TOK  aparell         aparell\N5-MS
138  TOK  flotant         flotant\JQ--6S|flotar\VG----
139  TOK  destinat        destinar\VC--SM
140  TOK  o          o\C
141  TOK  que        que\C|que\RR---66|que\D4
142  TOK  pugui           poder\V9R6S-
143  TOK  destinar        destinar\VI----
##-  DLI  -          =\DELIM
144  PGR  se         es\R6-----
145  TOK  a          a\P
146  TOK  serveis         servei\N5-MP
147  TOK  de         de\P
148  TOK  la         el\AFS|pr\REEC3FS
149  TOK  ind&uacute;stria         ind&uacute;stria\N5-FS
150  TOK  o          o\C
151  TOK  comer&ccedil;        comer&ccedil;\N5-MS
152  TOK  mar&iacute;tim       mar&iacute;tim\JQ--MS
153  TOK  o          o\C
154  TOK  fluvial         fluvial\JQ--6S
---  DLD  ".   EOS   =\DELS
##   TAG  </s>
```

Annex 6: Fragment de document castellà analitzat morfològicament i desambiguat lingüísticament

(fragment de d00063.sgm, mostra 3 -*Dret constitucional i estatutari*-)

```
##  TAG  <div2 type=tit n=4>
##  TAG  <head type=main>
1   TOK  TÍTULO                BOStítulo\N5-MS
##  TAG  <num pos='X'>
2   TOK  IV                    EOSnum\X
##  TAG  </num>
##  TAG  </head>
##  TAG  <head type=sub>
3   TOK  De                    BOSde\P
4   TOK  la                       el\AFS
5   TOK  composición              ?\N5-66|?\JQ--66
6   TOK  y                        y\C
7   TOK  atribuciones             atribución\N5-FP
8   TOK  de                       de\P
9   TOK  los                      el\AMP
10  TOK  órganos                  órgano\N5-MP
11  TOK  jurisdiccionales      EOSjurisdiccional\JQ--6P
##  TAG  </head>
##  TAG  <div3 type=cap n=1>
##  TAG  <head type=main>
12  TOK  CAPÍTULO              BOScapítulo\N5-MS
13  TOK  i                     EOS?\N5-66|?\JQ--66
##  TAG  </head>
##  TAG  <head type=sub>
14  TOK  DE                    BOSde\P
15  PGR  L                        el\AMS
16  TOK  TRIBUNAL                 tribunal\N5-MS
17  TOK  SUPREMO               EOSsupremo\JQ--MS
##  TAG  </head>
##  TAG  <div4 type=art n=53>
##  TAG  <p>
##  TAG  <s>
18  TOK  El                    BOSel\AMS
##  TAG  <name>
19  TOK  Tribunal Supremo         tribunal supremo\N4666
##  TAG  </name>
--- DLS  ,                        =\DELIM
20  TOK  con                      con\P
21  TOK  sede                     sede\N5-FS
22  TOK  en                       en\P
23  TOK  la                       el\AFS
24  TOK  villa                    villa\N5-FS
```

```
25  TOK  de                        de\P
##  TAG  <name>
26  TOK  Madrid                    madrid\N4666
##  TAG  </name>
--- DLS  ,                         =\DELIM
27  TOK  es                        ser\VDR3S-
28  TOK  el                        el\AMS
29  TOK  órgano                    órgano\N5-MS
30  TOK  jurisdiccional            jurisdiccional\JQ--6S
31  TOK  superior                  superior\JQ--MS
32  TOK  en                        en\P
33  TOK  todos                     todo\EN--MP
34  TOK  los                       el\AMP
35  TOK  órdenes                   orden\N5-6P
--- DLD  ,                         =\DELIM
36  TOK  salvo                     salvo\D4|salvo\JQ--
MS|salvar\VDR1S-
37  TOK  lo                        pr\REEC3MS
38  TOK  dispuesto                 disponer\HMS
39  TOK  en                        en\P
40  TOK  materia                   materia\N5-FS
41  TOK  de                        de\P
42  TOK  garantías                 garantía\N5-FP
43  TOK  constitucionales          constitucional\JQ--
6P|constitucional\N5-6P
--- DLD  .                       EOS=\DELS
##  TAG  </s>
##  TAG  </p>
##  TAG  <p>
##  TAG  <s>
44  TOK  Tendrá                  BOStener\VDU3S-
45  TOK  jurisdicción              jurisdicción\N5-FS
46  TOK  en                        en\P
47  TOK  toda                      todo\EN--FS
##  TAG  <name>
48  TOK  España                    españa\N4666
##  TAG  </name>
49  TOK  y                         y\C
50  TOK  ningún                    ningún\JN--MS
51  TOK  otro                      otro\EN--MS
52  TOK  podrá                     poder\VDU3S-
53  TOK  tener                     tener\VI----
54  TOK  el                        el\AMS
55  TOK  título                    título\N5-MS
56  TOK  de                        de\P
##  TAG  <name>
57  TOK  Supremo                   supremo\N4666
##  TAG  </name>
```

```
--- DLS  .                    EOS=\DELS
##  TAG  </s>
##  TAG  </p>
##  TAG  </div4>
##  TAG  <div4 type=art n=54>
##  TAG  <p>
##  TAG  <s>
58  TOK  El                   BOSel\AMS
##  TAG  <name>
59  TOK  Tribunal Supremo        tribunal supremo\N4666
##  TAG  </name>
60  TOK  se                      pr\R6
61  TOK  compondrá               componer\VDU3S-
62  TOK  de                      de\P
63  TOK  su                      su\JP636S
##  TAG  <name>
64  TOK  Presidente              presidente\N4666
##  TAG  </name>
--- DLS  ,                       =\DELIM
65  TOK  los                     el\AMP
##  TAG  <name>
66  TOK  Presidentes de Sala     presidentes de sala\N4666
##  TAG  </name>
67  TOK  y                       y\C
68  TOK  los                     el\AMP
##  TAG  <name>
69  TOK  Magistrados             magistrados\N4666
##  TAG  </name>
70  TOK  que                     que\RR---66
71  TOK  determine               determinar\VJR6S-
72  TOK  la                      el\AFS
##  TAG  <name>
73  TOK  Ley                     ley\N4666
##  TAG  </name>
74  TOK  para                    para\P
75  TOK  cada                    cada\JN--6S|cada\N5-MS
76  TOK  una                     un\E6--FS|unir\V9R6S-
77  TOK  de                      de\P
78  TOK  las                     el\AFP
##  TAG  <name>
79  TOK  Salas                   salas\N4666
##  TAG  </name>
80  TOK  y                       y\C
--- DLD  ,                       =\DELIM
81  TOK  en                      en\P
82  TOK  su                      su\JP636S
83  TOK  caso                    caso\N5-MS
--- DLD  ,                       =\DELIM
```

```
##  TAG  <name>
84  TOK  Secciones              secciones\N4666
##  TAG  </name>
85  TOK  en                     en\P
86  TOK  que                    que\RR---66
87  TOK  las                    el\AFP
88  TOK  mismas                 mismo\EN--FP
89  TOK  puedan                 poder\VJR3P-
90  TOK  articular              articular\VI----
91  PGR  se                     pr\R6
--- DLD  .                    EOS=\DELS
##  TAG  </s>
##  TAG  </p>
##  TAG  </div4>
```

Annex 7: Fragment d'un document d'entrada al desambiguador estadístic per al català

(fragment de d00024.sgm, mostra 14 -*Dret Mercantil*-)

```
##   TAG  <div1 n=6 complete=n>
##   TAG  <head type=main>
1    TOK  DRET BOS  dret\JQ--MS|dret\N5-MS
2    TOK  DE        de\P
3    TOK  LA        el\AFS
4    TOK  NAVEGACI&Oacute;    EOS  navegaci&oacute;\N5-FS
##   TAG  </head>
##   TAG  <div2 n=72>
##   TAG  <head type=main>
5    TOK  ESTATUT   BOS  estatut\N5-MS
6    TOK  JUR&Iacute;DIC      jur&iacute;dic\JQ--MS
7    TOK  DE        de\P
8    PGR  L         el\AMS
9    TOK  VAIXELL        vaixell\N5-MS
10   TOK  I         i\C
11   TOK  DE        de\P
12   TOK  L         el\A6S|pr\REE636S
##-  DLI  '         =\DELIM
13   PGR  AERONAU   EOS  aeronau\N5-FS
##   TAG  </head>
##   TAG  <div3 n=1>
##   TAG  <head type=main>
14   TOK  EL   BOS  el\AMS
15   TOK  VAIXELL   EOS  vaixell\N5-MS
##   TAG  </head>
##   TAG  <div4 n=1.107>
##   TAG  <head type=main>
16   TOK  Concepte  BOS  concepte\N5-MS
17   TOK  i         i\C
18   TOK  naturalesa         naturalesa\N5-FS
19   TOK  jur&iacute;dica    EOS  jur&iacute;dic\JQ--FS
##   TAG  </head>
##   TAG  <p>
##   TAG  <s>
20   TOK  En   BOS  en\P
21   TOK  sentit         sentir\HMS|sentit\N5-MS
22   TOK  t&egrave;cnic      t&egrave;cnic\JQ--
MS|t&egrave;cnic\N5-MS
23   TOK  parlem         parlar\V7R1P-
24   TOK  de        de\P
25   TOK  vaixell        vaixell\N5-MS
26   TOK  per       per\P
27   TOK  referir        referir\VI----
```

```
##-  DLI  -          =\DELIM
28   PGR  nos        pr\REE616P
29   TOK  a          a\P
30   TOK  qualsevol       qualsevol\EN--6S|qualsevol\N5-6S
31   TOK  construcci&oacute;       construcci&oacute;\N5-FS
32   TOK  destinada       destinar\VC--SF
33   TOK  a          a\P
34   TOK  la         el\AFS
35   TOK  navegaci&oacute;         navegaci&oacute;\N5-FS
36   TOK  mar&iacute;tima          mar&iacute;tim\JQ--FS
37   TOK  o          o\C
38   TOK  fluvial         fluvial\JQ--6S
---  DLD  .    EOS   =\DELS
##   TAG  </s>
##   TAG  <s>
39   TOK  A    BOS   a\P
40   TOK  aquesta         aquest\ED--FS
41   TOK  idea       idea\N5-FS|idear\V8R6S-
42   TOK  atenen          atenir-se\VDR3P-|atendre\VDR3P-
43   TOK  tamb&eacute;         tamb&eacute;\D4
---  DLD  ,          =\DELIM
##   TAG  <loc pos='D'>
44   TOK  en general           =\D4
##   TAG  </loc>
---  DLS  ,          =\DELIM
45   TOK  els        el\AMP
46   TOK  ordenaments          ordenament\N5-MP
47   TOK  positius        positiu\N5-MP|positiu\JQ--MP
48   TOK  moderns         modern\JQ--MP
49   TOK  en         en\P
50   TOK  determinar           determinar\VI----
---  DLD  ,          =\DELIM
51   TOK  amb        amb\P
52   TOK  major           major\N5-MS|major\JQ--6S
53   TOK  o          o\C
54   TOK  menor           menor\JQ--6S|menor\N5-6S
55   TOK  amplitud        amplitud\N5-FS
---  DLD  ,          =\DELIM
56   TOK  la         el\AFS
57   TOK  noci&oacute;         noci&oacute;\N5-FS
58   TOK  jur&iacute;dica          jur&iacute;dic\JQ--FS
59   TOK  de         de\P
60   PGR  l          el\AMS
61   TOK  vaixell         vaixell\N5-MS
---  DLD  .    EOS   =\DELS
##   TAG  </s>
##   TAG  </p>
##   TAG  <p>
```

```
##   TAG  <s>
62   TOK  En   BOS  en\P
63   TOK  el        el\AMS
64   TOK  nostre         nostre\EP21MS
65   TOK  ordenament          ordenament\N5-MS
66   TOK  legal          legal\JQ--6S
---  DLD  ,         =\DELIM
67   TOK  l         el\A6S
##-  DLI  '         =\DELIM
68   PGR  article        article\N5-MS
##   TAG  <num pos='X'>
69   TOK  146       num\X
##   TAG  </num>
70   TOK  de        de\P
71   PGR  l         el\AMS
##   TAG  <name>
72   TOK  Reglament del Registre Mercantil   reglament           del
     registre mercantil\N4666
##   TAG  </name>
73   TOK  de        de\P
##   TAG  <num pos='X'>
74   TOK  1956      num\X
##   TAG  </num>
---  DLS  ,         =\DELIM
75   TOK  transit&ograve;riament   transit&ograve;riament\D4
76   TOK  vigent         vigent\JQ--6S
---  DLS  (         =\DELIM
##   TAG  <abbr>
77   TOK  veg.      =\N5-66
##   TAG  </abbr>
78   TOK  disposici&oacute;        disposici&oacute;\N5-FS
79   TOK  transit&ograve;ria       transitori\JQ--FS
80   TOK  sisena         sis&egrave;\EO--FS
81   TOK  de        de\P
82   PGR  l         el\AMS
##   TAG  <name>
83   TOK  Reglament del Registre Mercantil   reglament           del
     registre mercantil\N4666
##   TAG  </name>
84   TOK  de        de\P
##   TAG  <date ISO8601='12/29/1989'>
85   TOK  29 de desembre de 1989        =\T
##   TAG  </date>
---  DLS  ),        =\DELIM
86   TOK  suplint        suplir\VG----
87   TOK  la        el\AFS
88   TOK  llacuna        llacuna\N5-FS
89   TOK  de        de\P
```

```
90   PGR  l          el\AMS
##   TAG  <name>
91   TOK  Codi       codi\N4666
##   TAG  </name>
---  DLS  ,          =\DELIM
92   TOK  estableix       establir\V8R6S-
---  DLD  ,          =\DELIM
93   TOK  tamb&eacute;         tamb&eacute;\D4
94   TOK  en         en\P
95   TOK  aquest          aquest\ED--MS
96   TOK  sentit          sentir\HMS|sentit\N5-MS
---  DLD  ,          =\DELIM
97   TOK  que        que\C|que\RR---66|que\D4
##-  DLE  "          =\DELIM
98   TOK  es         pr\R6-----
99   TOK  reputaran       reputar\VDU3P-
100  TOK  vaixells        vaixell\N5-MP
---  DLD  ,          =\DELIM
101  TOK  per        per\P
102  TOK  a          a\P
103  PGR  ls         el\AMP
104  TOK  efectes         efecte\N5-MP
105  TOK  de         de\P
106  PGR  l          el\AMS
##   TAG  <name>
107  TOK  Codi       codi\N4666
##   TAG  </name>
108  TOK  de         de\P
109  TOK  comer&ccedil;        comer&ccedil;\N5-MS
110  TOK  i          i\C
111  TOK  d          de\P
##-  DLI  '          =\DELIM
112  PGR  aquest          aquest\ED--MS
##   TAG  <name>
113  TOK  Reglament       reglament\N4666
##   TAG  </name>
---  DLS  ,          =\DELIM
114  TOK  no         no\D4|no\N5-MS
115  TOK  nom&eacute;s         nom&eacute;s\D4
116  TOK  les        el\AFP
117  TOK  embarcacions         embarcaci&oacute;\N5-FP
118  TOK  destinades           destinar\VC--PF
119  TOK  a          a\P
120  TOK  la         el\AFS
121  TOK  navegaci&oacute;          navegaci&oacute;\N5-FS
122  TOK  de         de\P
123  TOK  cabotatge       cabotatge\N5-MS
124  TOK  i          i\C
```

```
125  TOK  altura          altura\N5-FS
---  DLD  ,          =\DELIM
126  TOK  sin&oacute;          sin&oacute;\C
127  TOK  tamb&eacute;         tamb&eacute;\D4
128  TOK  els        el\AMP
129  TOK  dics       dic\N5-MP
130  TOK  flotants        flotant\JQ--6P
---  DLD  ,          =\DELIM
131  TOK  pontons         pont&oacute;\N5-MP
---  DLD  ,          =\DELIM
132  TOK  dragues         draga\N5-FP|dragar\VDR2S-
---  DLD  ,          =\DELIM
133  TOK  g&agrave;nguils          g&agrave;nguil\N5-MP
134  TOK  i          i\C
135  TOK  qualsevol       qualsevol\EN--6S|qualsevol\N5-6S
136  TOK  altre           altre\EN--MS
137  TOK  aparell         aparell\N5-MS
138  TOK  flotant         flotant\JQ--6S|flotar\VG----
139  TOK  destinat        destinar\VC--SM
140  TOK  o          o\C
141  TOK  que        que\C|que\RR---66|que\D4
142  TOK  pugui           poder\V9R6S-
143  TOK  destinar        destinar\VI----
##-  DLI  -          =\DELIM
144  PGR  se         pr\R6-----
145  TOK  a          a\P
146  TOK  serveis         servei\N5-MP
147  TOK  de         de\P
148  TOK  la         el\AFS
149  TOK  ind&uacute;stria         ind&uacute;stria\N5-FS
150  TOK  o          o\C
151  TOK  comer&ccedil;        comer&ccedil;\N5-MS
152  TOK  mar&iacute;tim       mar&iacute;tim\JQ--MS
153  TOK  o          o\C
154  TOK  fluvial         fluvial\JQ--6S
---  DLD  ".   EOS   =\DELS
##   TAG  </s>
```

Annex 8: Fragment de document desambiguat estadísticament

(fragment de d00024.sgm, mostra 14 -*Dret Mercantil*-)
(informació morfològica sense ambigüitat)

```
##   TAG  <div1 n=6 complete=n>
##   TAG  <head type=main>
1    TOK  DRET BOS  dret\JQ--MS
2    TOK  de        de\P
3    TOK  la        el\AFS
4    TOK  NAVEGACI&Oacute;    EOS  navegaci&oacute;\N5-FS
##   TAG  </head>
##   TAG  <div2 n=72>
##   TAG  <head type=main>
5    TOK  ESTATUT   BOS  estatut\N5-MS
6    TOK  JUR&Iacute;DIC      jur&iacute;dic\JQ--MS
7    TOK  de        de\P
8    PGR  l         el\AMS
9    TOK  VAIXELL        vaixell\N5-MS
10   TOK  i         i\C
11   TOK  de        de\P
12   TOK  L         el\A6S
##-  DLI  '         =\DELIM
13   PGR  AERONAU   EOS  aeronau\N5-FS
##   TAG  </head>
##   TAG  <div3 n=1>
##   TAG  <head type=main>
14   TOK  el   BOS  el\AMS
15   TOK  VAIXELL   EOS  vaixell\N5-MS
##   TAG  </head>
##   TAG  <div4 n=1.107>
##   TAG  <head type=main>
16   TOK  Concepte  BOS  concepte\N5-MS
17   TOK  i         i\C
18   TOK  naturalesa         naturalesa\N5-FS
19   TOK  jur&iacute;dica     EOS  jur&iacute;dic\JQ--FS
##   TAG  </head>
##   TAG  <p>
##   TAG  <s>
20   TOK  En   BOS  en\P
21   TOK  sentit         sentit\N5-MS
22   TOK  t&egrave;cnic       t&egrave;cnic\JQ--MS
23   TOK  parlem         parlar\V7R1P-
24   TOK  de        de\P
25   TOK  vaixell        vaixell\N5-MS
26   TOK  per       per\P
27   TOK  referir        referir\VI----
```

```
##-  DLI  -         =\DELIM
28   PGR  nos       pr\REE616P
29   TOK  a         a\P
30   TOK  qualsevol      qualsevol\EN--6S
31   TOK  construcci&oacute;       construcci&oacute;\N5-FS
32   TOK  destinada      destinar\VC--SF
33   TOK  a         a\P
34   TOK  la        el\AFS
35   TOK  navegaci&oacute;         navegaci&oacute;\N5-FS
36   TOK  mar&iacute;tima          mar&iacute;tim\JQ--FS
37   TOK  o         o\C
38   TOK  fluvial        fluvial\JQ--6S
---  DLD  .    EOS  =\DELS
##   TAG  </s>
##   TAG  <s>
39   TOK  A    BOS  a\P
40   TOK  aquesta        aquest\ED--FS
41   TOK  idea      idear\V8R6S-
42   TOK  atenen         atenir-se\VDR3P-
43   TOK  tamb&eacute;         tamb&eacute;\D4
---  DLD  ,         =\DELIM
##   TAG  <loc pos='D'>
44   TOK  en general          =\D4
##   TAG  </loc>
---  DLS  ,         =\DELIM
45   TOK  els       el\AMP
46   TOK  ordenaments         ordenament\N5-MP
47   TOK  positius       positiu\JQ--MP
48   TOK  moderns        modern\JQ--MP
49   TOK  en        en\P
50   TOK  determinar          determinar\VI----
---  DLD  ,         =\DELIM
51   TOK  amb       amb\P
52   TOK  major          major\JQ--6S
53   TOK  o         o\C
54   TOK  menor          menor\JQ--6S
55   TOK  amplitud       amplitud\N5-FS
---  DLD  ,         =\DELIM
56   TOK  la        el\AFS
57   TOK  noci&oacute;         noci&oacute;\N5-FS
58   TOK  jur&iacute;dica          jur&iacute;dic\JQ--FS
59   TOK  de        de\P
60   PGR  l         el\AMS
61   TOK  vaixell        vaixell\N5-MS
---  DLD  .    EOS  =\DELS
##   TAG  </s>
##   TAG  </p>
##   TAG  <p>
```

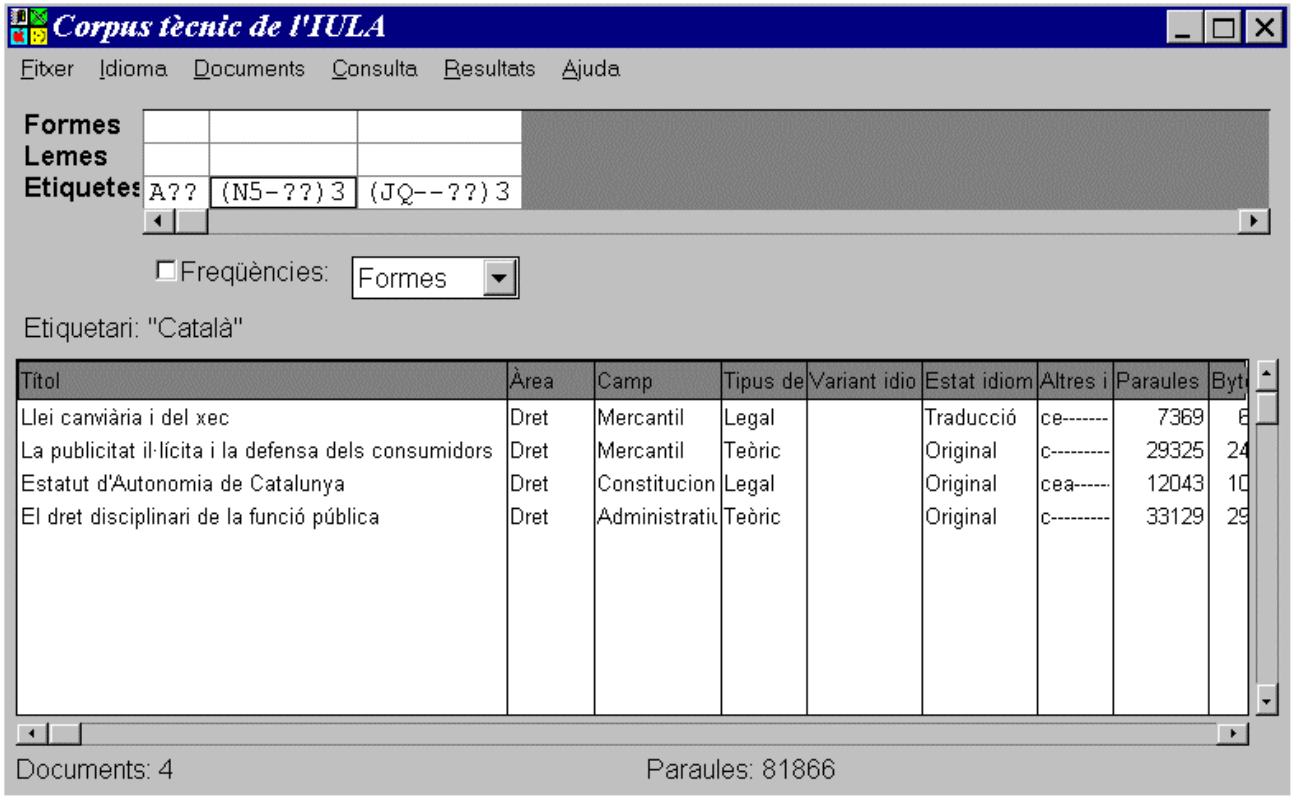
```
##   TAG  <s>
62   TOK  En   BOS  en\P
63   TOK  el        el\AMS
64   TOK  nostre         nostre\EP21MS
65   TOK  ordenament          ordenament\N5-MS
66   TOK  legal          legal\JQ--6S
---  DLD  ,         =\DELIM
67   TOK  l         el\A6S
##-  DLI  '         =\DELIM
68   PGR  article        article\N5-MS
##   TAG  <num>
69   TOK  146       num\X
##   TAG  </num>
70   TOK  de        de\P
71   PGR  l         el\AMS
##   TAG  <name>
72   TOK  Reglament del Registre Mercantil   reglament     del     registre
     mercantil\N4666
##   TAG  </name>
73   TOK  de        de\P
##   TAG  <num>
74   TOK  1956      num\X
##   TAG  </num>
---  DLS  ,         =\DELIM
75   TOK  transit&ograve;riament   transit&ograve;riament\D4
76   TOK  vigent         vigent\JQ--6S
---  DLS  (         =\DELIM
##   TAG  <abbr>
77   TOK  veg.      =\N5-66
##   TAG  </abbr>
78   TOK  disposici&oacute;        disposici&oacute;\N5-FS
79   TOK  transit&ograve;ria       transitori\JQ--FS
80   TOK  sisena         sis&egrave;\EO--FS
81   TOK  de        de\P
82   PGR  l         el\AMS
##   TAG  <name>
83   TOK  Reglament del Registre Mercantil   reglament     del     registre
     mercantil\N4666
##   TAG  </name>
84   TOK  de        de\P
##   TAG  <date ISO8601='12/29/1989'>
85   TOK  29 de desembre de 1989       =\T
##   TAG  </date>
---  DLS  ),        =\DELIM
86   TOK  suplint        suplir\VG----
87   TOK  la        el\AFS
88   TOK  llacuna        llacuna\N5-FS
89   TOK  de        de\P
```

```
90   PGR  l          el\AMS
##   TAG  <name>
91   TOK  Codi       codi\N4666
##   TAG  </name>
---  DLS  ,          =\DELIM
92   TOK  estableix       establir\V8R6S-
---  DLD  ,          =\DELIM
93   TOK  tamb&eacute;         tamb&eacute;\D4
94   TOK  en         en\P
95   TOK  aquest          aquest\ED--MS
96   TOK  sentit          sentir\HMS
---  DLD  ,          =\DELIM
97   TOK  que        que\RR---66
##-  DLE  "          =\DELIM
98   TOK  es         pr\R6-----
99   TOK  reputaran       reputar\VDU3P-
100  TOK  vaixells        vaixell\N5-MP
---  DLD  ,          =\DELIM
101  TOK  per        per\P
102  TOK  a          a\P
103  PGR  ls         el\AMP
104  TOK  efectes         efecte\N5-MP
105  TOK  de         de\P
106  PGR  l          el\AMS
##   TAG  <name>
107  TOK  Codi       codi\N4666
##   TAG  </name>
108  TOK  de         de\P
109  TOK  comer&ccedil;        comer&ccedil;\N5-MS
110  TOK  i          i\C
111  TOK  d          de\P
##-  DLI  '          =\DELIM
112  PGR  aquest          aquest\ED--MS
##   TAG  <name>
113  TOK  Reglament       reglament\N4666
##   TAG  </name>
---  DLS  ,          =\DELIM
114  TOK  no         no\D4
115  TOK  nom&eacute;s         nom&eacute;s\D4
116  TOK  les        el\AFP
117  TOK  embarcacions         embarcaci&oacute;\N5-FP
118  TOK  destinades           destinar\VC--PF
119  TOK  a          a\P
120  TOK  la         el\AFS
121  TOK  navegaci&oacute;          navegaci&oacute;\N5-FS
122  TOK  de         de\P
123  TOK  cabotatge       cabotatge\N5-MS
124  TOK  i          i\C
```

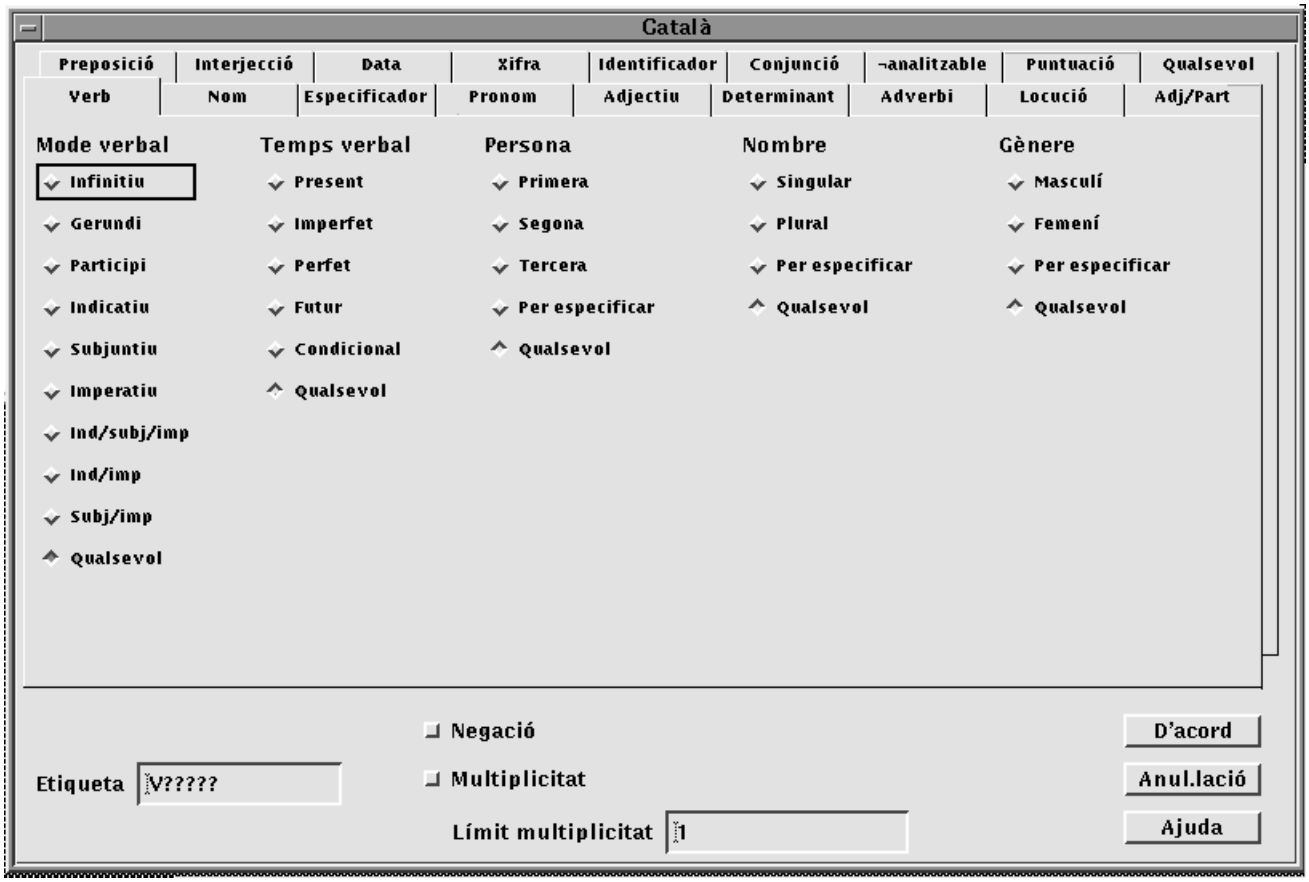
```
125  TOK  altura          altura\N5-FS
---  DLD  ,          =\DELIM
126  TOK  sin&oacute;          sin&oacute;\C
127  TOK  tamb&eacute;         tamb&eacute;\D4
128  TOK  els        el\AMP
129  TOK  dics       dic\N5-MP
130  TOK  flotants        flotant\JQ--6P
---  DLD  ,          =\DELIM
131  TOK  pontons         pont&oacute;\N5-MP
---  DLD  ,          =\DELIM
132  TOK  dragues         dragar\VDR2S-
---  DLD  ,          =\DELIM
133  TOK  g&agrave;nguils          g&agrave;nguil\N5-MP
134  TOK  i          i\C
135  TOK  qualsevol       qualsevol\EN--6S
136  TOK  altre           altre\EN--MS
137  TOK  aparell         aparell\N5-MS
138  TOK  flotant         flotar\VG----
139  TOK  destinat        destinar\VC--SM
140  TOK  o          o\C
141  TOK  que        que\RR---66
142  TOK  pugui           poder\V9R6S-
143  TOK  destinar        destinar\VI----
##-  DLI  -          =\DELIM
144  PGR  se         es\R6-----
145  TOK  a          a\P
146  TOK  serveis         servei\N5-MP
147  TOK  de         de\P
148  TOK  la         el\AFS
149  TOK  ind&uacute;stria         ind&uacute;stria\N5-FS
150  TOK  o          o\C
151  TOK  comer&ccedil;        comer&ccedil;\N5-MS
152  TOK  mar&iacute;tim       mar&iacute;tim\JQ--MS
153  TOK  o          o\C
154  TOK  fluvial         fluvial\JQ--6S
---  DLD  ".   EOS   =\DELS
##   TAG  </s>
```

Annex 9: Eina principal d'explotació del Corpus Tècnic

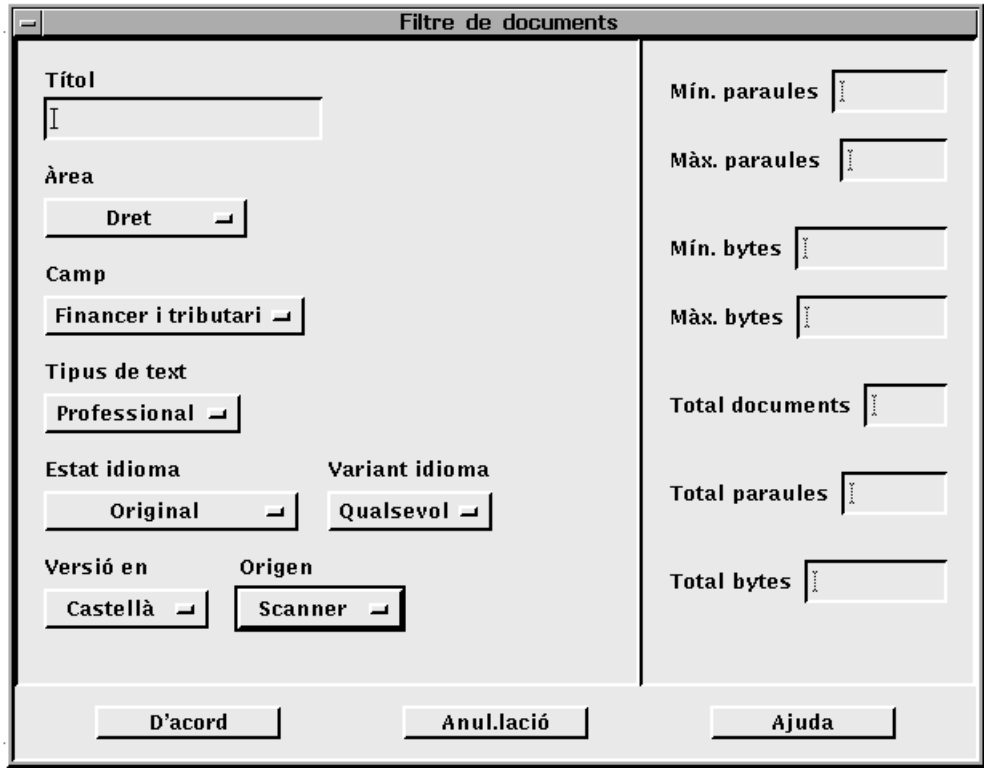
Extracció de concordances
a) Pantalla d'accés



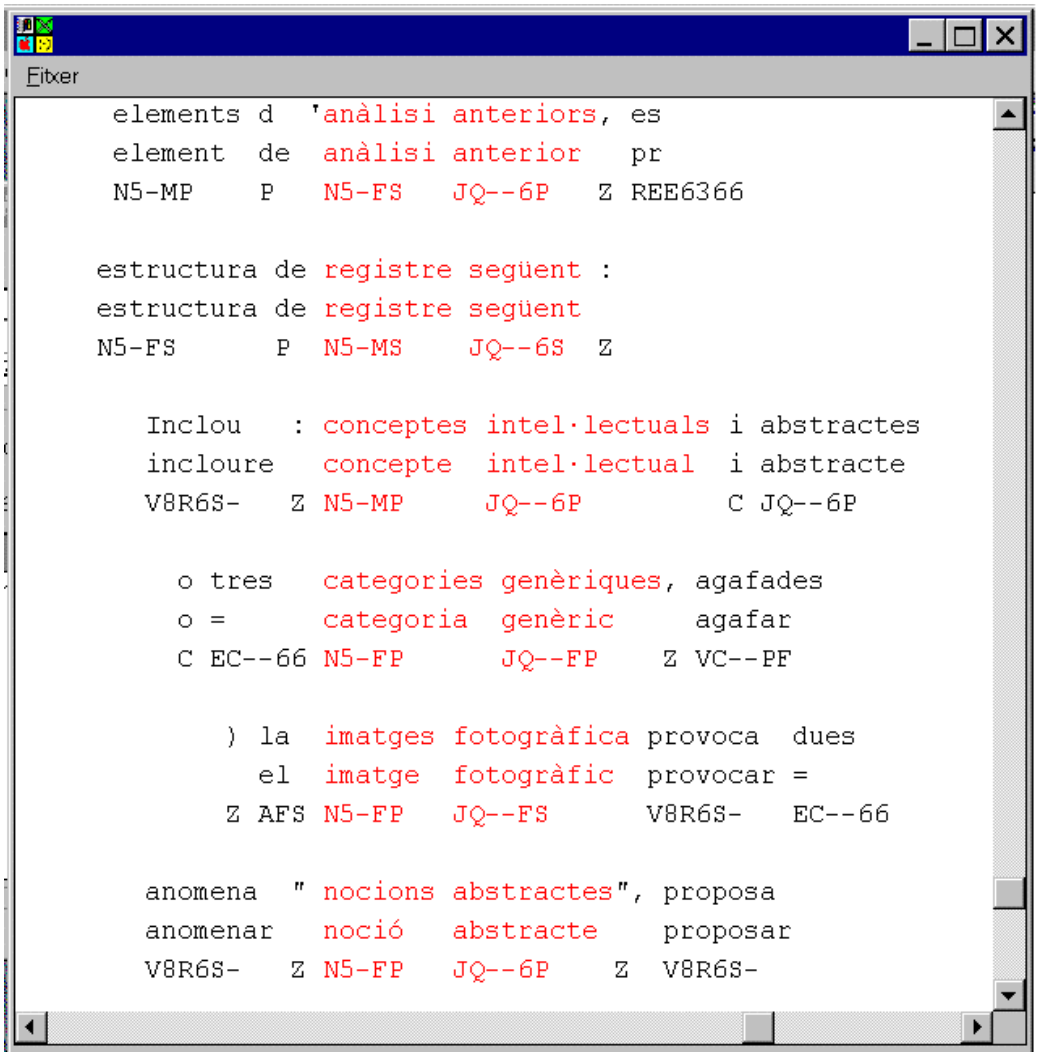
b) Pantalla de selecció de categories gramaticals



c) Pantalla de selecció de documents en funció de la informació de la capçalera



d) Pantalla de presentació de resultats



Annex 10: Estat actual (maig, 1997)

♦ En relació a la fase prèvia (vegeu apartat 3.3) es compta amb l'estructuració dels arbres de camp establerts pels especialistes per a cadascuna de les àrees temàtiques.

ARBRE DE CAMP DE L'ÀREA TEMÀTICA DE DRET:

- Dret privat
    - Dret civil
    - Dret mercantil
    - Dret laboral
    - Dret penal
    - Dret canònic
- Dret públic
    - Dret constitucional i estatutari
    - Dret administratiu
    - Dret financer i tributari
    - Dret internacional i públic
- Teoria del dret

ARBRE DE CAMP DE L'ÀREA TEMÀTICA D'ECONOMIA:

- Economia general i ensenyament
- Metodologia i història del pensament econòmic
- Mètodes matemàtics i quantitatius
- Microeconomia
- Macroeconomia i economia monetària
- Economia internacional
- Economia financera
- Economia pública
- Sanitat, educació i benestar
- Treball i economia demogràfica
- Dret i economia
- Organització industrial
- Administració d'empreses i economia d'empresa; màrqueting; comptabilitat
- Història econòmica
- Desenvolupament econòmic, canvi tecnològic i creixement
- Sistemes econòmics
- Economia agrícola i de recursos naturals
- Economia urbana, rural i regional
- Altres temes especials

ARBRE DE CAMP DE L'ÀREA TEMÀTICA DE MEDI AMBIENT:

- Medi natural
    - Ciències de la terra
    - Ciències de l'atmosfera
    - Ciències de l'aigua
    - Biologia
    - Ecologia
    - Recursos naturals
    - Energia
    - Desastres naturals
    - Geografia
- Assentaments humans
    - Ordenació del territori
    - Urbanisme
    - Edificació i vivendes
    - Transport
- Impacte ambiental
    - Degradació del medi
    - Contaminació
    - Agents contaminants
    - Efectes de contaminació
    - Residus
- Política i dret ambiental
    - Avaluació ambiental
    - Gestió ambiental
    - Dret ambiental
    - Planificació
    - Organitzacions governamentals
- Ciència i tecnologia ambiental i energètica
    - Tractament de l'aigua
    - Tractament de l'aire
    - Energia
    - Tècniques de laboratori
    - Equips
    - Ciència i tecnologia bàsica
- Medi social i organitzacions
    - Sociologia
    - Demografia
    - Política
    - Educació
    - Salut
    - Cultura
    - Art
    - Organitzacions no governamentals

ARBRE DE CAMP DE L'ÀREA TEMÀTICA D'INFORMÀTICA:

- Hardware
    - Arquitectura d'ordinadors
    - Microprogramació
    - Circuits lògics. Memòries. Comunicació de dades
- Organització dels ordinadors
    - Arquitectures de processadors
    - Xarxes de comunicació
    - Disseny d'ordinadors
    - Implementació d'ordinadors
    - Sistemes distribuïts. Xarxes d'ordinadors
    - Seguretat
- Software
    - Tècniques de programació
    - Mètodes de programació
    - Enginyeria del Software
    - Llenguatges de programació
    - Sistemes operatius
    - Compiladors
    - Paral·lelisme. Concurrència
    - Tècniques de validació i prova
- Estructures de dades
    - Estructures de dades
    - Representació de les dades
    - Codificació. Teoria de la informació
    - Fitxers
    - Encriptat d'informació
- Teoria de la computació
    - Algorítmica
    - Anàlisi d'algoritmes. Lògica de programes. Complexitat
    - Llenguatges formals. Autòmates
- Sistemes d'Informació
    - Models d'informació
    - Bases de dades
    - Emmagatzematge i accés a la informació
- Metodologia de la computació
    - Manipulació algebraica
    - Intel·ligència Artificial
        - Representació del Coneixement
        - Adquisició del Coneixement
        - Aprenentatge simbòlic i subsimbòlic
        - Resolució de Problemes
        - Planificació
        - Raonament

        - Sistemes Basats en el Coneixement. Sistemes Experts
    - Gràfics
    - Processament d'imatges
    - Simulació i modelat
    - Reconeixement de patrons
    - Procés de textos
- Aplicacions
    - Sistemes d'Informació
    - Ofimàtica
    - Aplicacions en Enginyeria
    - Aplicacions en Medicina
    - Aplicacions en Ciències socials
    - Disseny assistit per ordinador
    - Ensenyament assistit per ordinador
- Comunicació persona/màquina
    - Interfícies persona/màquina
    - Multimèdia
    - Visió
    - Modelització de l'usuari
- Entorn
    - Indústria de la Informàtica
    - Història de la Informàtica
    - Ordinadors i societat
    - Ordinadors i ensenyament
    - La professió de la Informàtica
    - Aspectes legals

ARBRE DE CAMP DE L'ÀREA TEMÀTICA DE MEDICINA:

- Anatomia
    - Regions del cos
    - Sistema musculoesquelètic
    - Sistema digestiu
    - Sistema respiratori
    - Sistema urogenital
    - Sistema endocrí
    - Sistema cardiovascular
    - Sistema nerviós
    - Òrgans dels sentits
    - Tipus de teixits
    - Cèl·lules
    - Fluids i secrecions
    - Terminologia animal
    - Sistema estomatognàtic
    - Sistema sanguini i immunitari
    - Estructures embrionàries
- Organismes
    - Invertebrats
    - Vertebrats
    - Bactèries
    - Virus
    - Algues i fongs
    - Plantes
- Malalties
    - Malalties bacterianes i fúngiques
    - Malalties víriques
    - Malalties parasitàries
    - Malalties neoplàsiques
    - Malalties musculoesquelètiques
    - Malalties del sistema digestiu
    - Malalties del sistema estomatognàtic
    - Malalties del sistema respiratori
    - Malalties otorinolaringològiques
    - Malalties del sistema nerviós
    - Malalties oftalmològiques
    - Malalties urològiques i genitals masculines
    - Malalties genitals femenines i complicacions obstètriques
    - Malalties cardiovasculars
    - Malalties dels sistemes sanguini i limfàtic
    - Malalties i transtorns neonatals
    - Malalties de la pell i del teixit connectiu
    - Malalties metabòliques i nutricionals

    - Malalties endocrinològiques
    - Malalties immunològiques
    - Traumatismes, malalties professionals, intoxicacions
    - Malalties animals
    - Simptomatologia i patologia general
- Productes químics i fàrmacs
    - Compostos inorgànics
    - Compostos orgànics
    - Compostos heterocíclics
    - Hidrocarburs policíclics
    - Contaminants ambientals i pesticides
    - Hormones, anàlegs hormonals i antagonistes
    - Agents controladors de la reproducció
    - Enzims, coenzims, inhibidors enzimàtics
    - Carbohidrats i agents hipoglicemiants
    - Lípids i agents antilipèmics
    - Factors de creixement, pigments i vitamines
    - Aminoàcids, vitamines, pèptids i proteïnes
    - Nucleòsids i nucleòtids
    - Depressors del sistema nerviós central
    - Agents del sistema nerviós central
    - Agents autonòmics
    - Agents neuromusculars
    - Agents cardiovasculars
    - Agents hematològics, gàstrics i renals
    - Agents antiinfecciosos
    - Agents antiparasitaris
    - Agents antineoplàsics i immunosupressors
    - Agents neuroreguladors
    - Factors immunològics i biològics
    - Materials biomèdics i dentals, biomaterials
    - Miscel·lània d'agents i fàrmacs
- Tècniques i equipaments analítics, diagnòstics i terapèutics
    - Diagnosi
    - Terapèutica
    - Anestèsia i analgèsia
    - Cirurgia i intervencions
    - Miscel·lània de tècniques
    - Ortodòncia
    - Equipaments i Subministres
- Psiquiatria i psicologia
    - Conducta i mecanismes de conducta
    - Principis i processos psicològics
    - Transtorns mentals i de la conducta
    - Disciplines, proves, teràpies i serveis

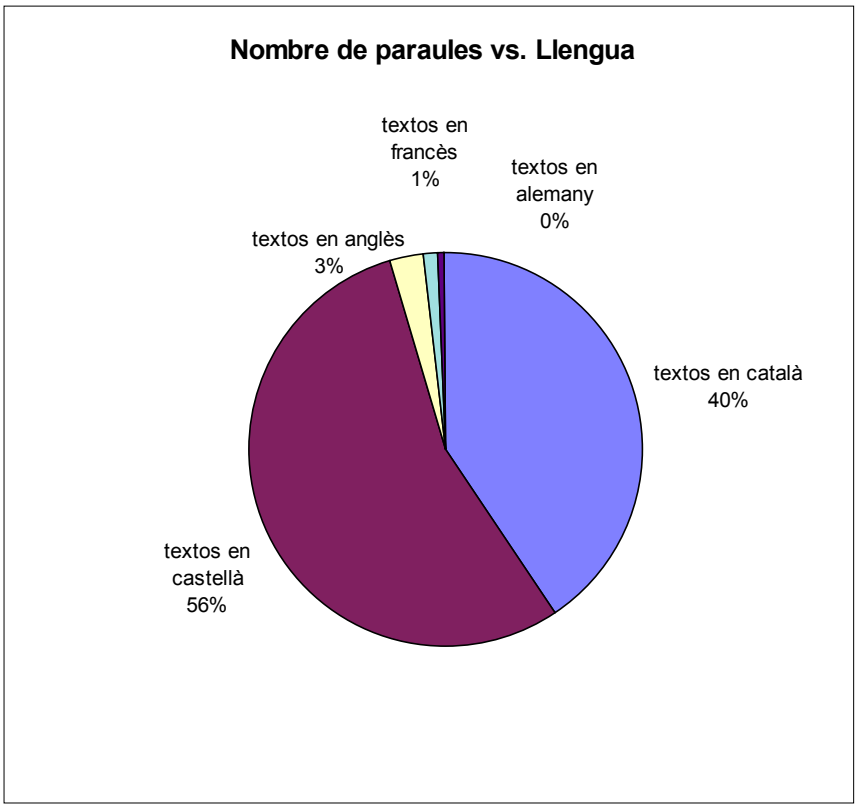
- Ciències biològiques
    - Ciències biològiques
    - Professions sanitàries
    - Salut pública i medi ambient
    - Fenòmens biològics, fisiologia cel·lular i immunitat
    - Genètica
    - Fenòmens bioquímics, metabolisme i nutrició
    - Fisiologia general
    - Fisiologia de la reproducció, fisiologia urogenital
    - Fisiologia de la circulació, fisiologia de la respiració
    - Fisiologia digestiva i oral, fisiologia de la pell
    - Fisiologia musculoesquelètica, del sistema nerviós i dels òrgans dels sentits
    - Fenòmens químics, fenòmens farmacològics
- Ciències físiques
- Antropologia, educació, sociologia i fenòmens socials
    - Ciències socials
    - Educació
    - Activitats humanes
- Tecnologia, Indústria, Agricultura
- Humanitats
- Informació Científica
- Grups Nominals
- Planificació i gestió sanitària
    - Característiques poblacionals
    - Facilitats, disponibilitats i serveis
    - Economia, organització, control
    - Administració de serveis de salut
    - Qualitat, accessibilitat i avaluació

- Pel que fa a les diferents etapes de processament del corpus, i concretament en la fase de tractament dels materials, s'està treballant en l'entrada de textos i en el seu marcatge estructural per tal d'augmentar el nombre de paraules de les diferents àrees temàtiques. En aquests moments es compta amb 6 milions de paraules i es preveu un ritme de creixement aproximat d'uns 3 milions de paraules per any.

  L'estat del marcatge estructural dels documents corresponents als cinc dominis d'especialitat és el següent:

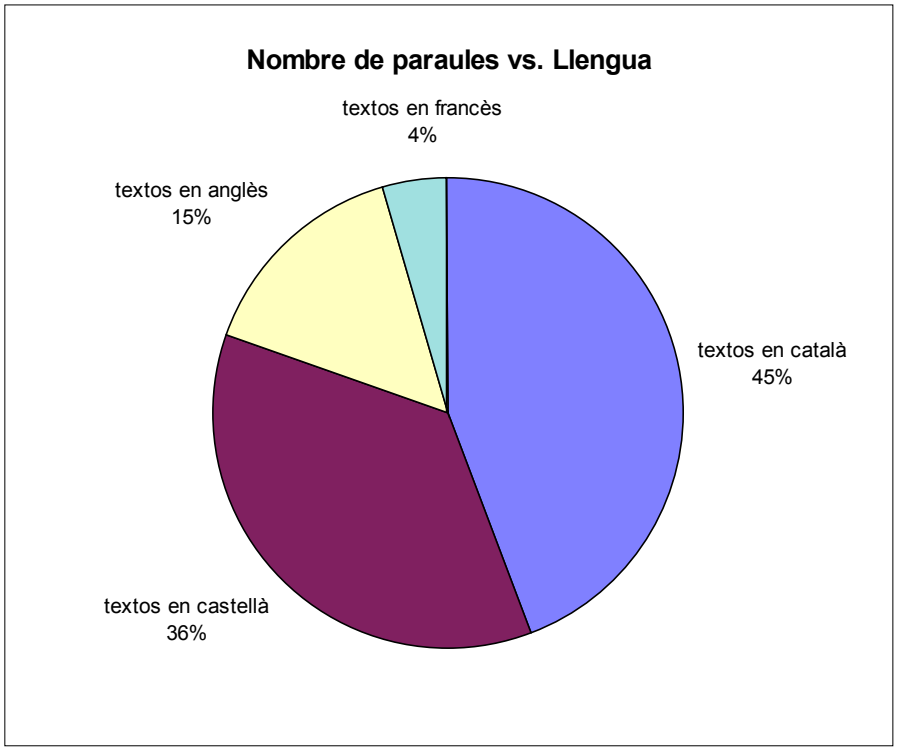
DRET:
Total de textos i de paraules marcades estructuralment per llengua:

	Mostres	Paraules
textos en català	387	1405542
textos en castellà	447	1916560
textos en anglès	15	98207
textos en francès	8	40277
textos en alemany	3	15617
T O T A L 166	860	**3476203**


Nombre de paraules vs. Llengua
textos en francès 1%
textos en alemany 0%
textos en anglès 3%
textos en català 40%
textos en castellà 56%

ECONOMIA:
Total de textos i de paraules marcades estructuralment per llengua:

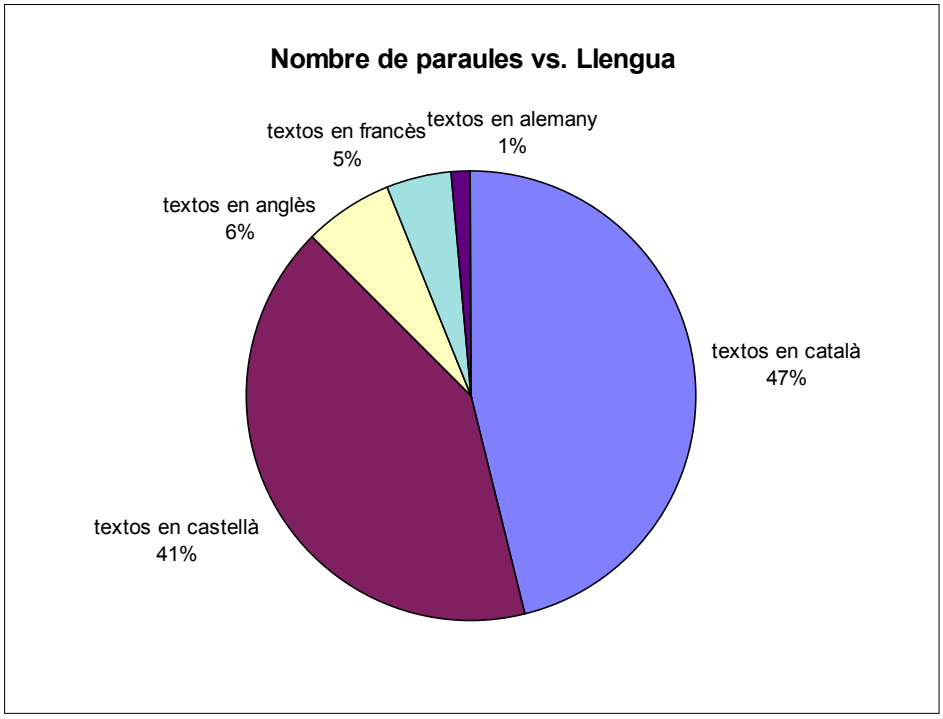
	Mostres	Paraules
textos en català	217	777846
textos en castellà	171	640160
textos en anglès	66	268050
textos en francès	33	78001
textos en alemany	0	0
T O T A L	487	**1764057**



MEDI AMBIENT:
Total de textos i de paraules marcades estructuralment per llengua:

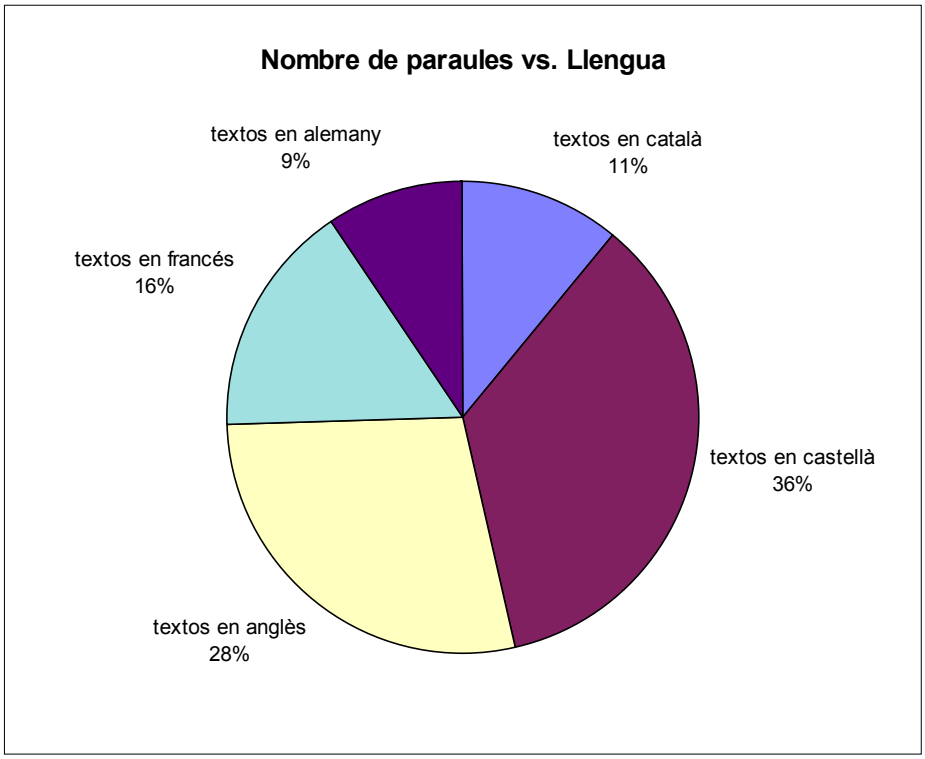
	Mostres	Paraules
textos en català	48	172581
textos en castellà	40	155224
textos en anglès	9	23481
textos en francès	5	17812
textos en alemany	2	5531
T O T A L	104	**374629**

.



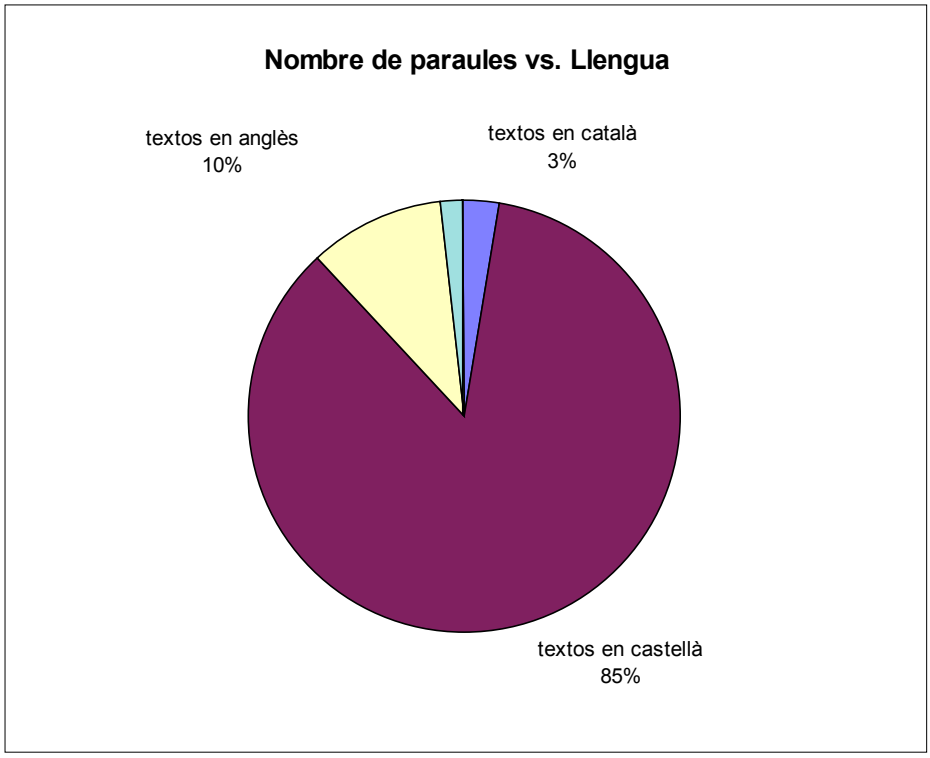
INFORMÀTICA:
Total de textos i de paraules marcades estructuralment per llengua:

	Mostres	Paraules
textos en català	25	98931
textos en castellà	76	312784
textos en anglès	65	248032
textos en francés	31	144748
textos en alemany	19	83228
T O T A L	216	**887723**



MEDICINA:
Total de textos i de paraules marcades estructuralment per llengua:

		Mostres	Paraules
textos en català		1	14583
textos en castellà		86	425801
textos en anglès		10	51389
textos en francés		0	8823
textos en alemany		0	0
T O T A L	8	97	**500596**



- En referència a l'etapa de processament lingüístic, i pel que fa als textos en llengua catalana, s'han desenvolupat les eines que permeten obtenir tots els textos del corpus completament desambiguats amb un marge d'encert del 95%. Pel que fa als textos en llengua castellana, actualment es treballa en la seva anàlisi morfològica i en la desambiguació. Vegeu els annexos 4, 5, 6 i 7.

- Finalment, en l'etapa d'explotació es treballa en el millorament del sistema d'extracció de concordances,
- i en el disseny d'altres aplicacions que permetin l'extracció de la informació que els usuaris del corpus demanen.

Annex 11: Membres del projecte

Direcció: Maria Teresa Cabré
Coordinació tècnica: Jordi Vivaldi

Consell Científic

Maria Teresa Cabré
M. Paz Battaner
Maite Turell
Toni Badia
Mercè Lorente
Lluís de Yzaguirre

Assessors d'àrea:

Dret	Carles Duarte Xavier Bernardí Esther Cabrera
Economia	Màxim Borrell Vicente Ortun Miquel Centelles
Medi Ambient	Pau Serra
Medicina	Toni Valero
Informàtica	Horacio Rodríguez Jordi Vivaldi

Marcatge estructural

Coordinació	Xavier Solé
Marcatge	Elisenda Bernal Cristina Corcoll Judit Feliu Sònia Jiménez Jordi Morel Mar Pongiluppi Elisabet Ricart Roser Saurí Teresa Vallès

Tractament Lingüístic

Coordinació	Carme Bach
Comprovació	Judit Aumatell Núria Castillo Helena Pàmpols Sergi Torner

Etiquetaris lingüístics

Català	Jordi Morel Carme Bach
Castellà	Sergi Torner Roser Martinez

Eines informàtiques

Coordinació	Jordi Vivaldi
Lematitzador	Lluís de Yzaguirre
Analitzador morfològic	Toni Badia Toni Tuells
Preprocés	Manel Pujol
Desambiguació lingüística	Lluís de Yzaguirre Carme Bach
Desambiguació estadística	Jordi Vivaldi Jordi Morel
Explotació del Corpus	Jordi Vivaldi

Documentació: Mireia Ribera

Infraestructura informàtica: Jesús Carrasco

Altres membres del projecte Corpus (grup LATRAL)

Lluís Codina, Jaume Martí, Victòria Alsina, Enric Vallduví, Jenny Brumme, Janet deCesaris, Josep M. Fontana, Louise McNally, Guilhem Naro, Esteve Clua, Carme Colominas, Judit Freixa, Cristina Gelpí, Montserrat Forcadell, Montserrat Gonzàlez, Montserrat Ribas, Cristòfor Rovira, Elisabet Solé, Carles Tebé, Rosa Bayà, Laura Borràs, Rosa Estopà, Eduardo Sosa, Joan A. Mesquida, Marta Carulla.