ARXIU INFORMATITZAT DE TEXTOS CATALANS MEDIEVALS

Presentació

Al *Seminari de Filologia i Informàtica* de la Universitat Autònoma de Barcelona, s'està treballant en l'elaboració d'un *Arxiu Informatitzat de Textos Catalans Medievals*, per a la seva posterior conversió en un *Banc de Paraules* (Tresor de la Llengua Catalana Medieval), sota la direcció del Dr. José Manuel Blecua. El disseny i direcció del projecte és a càrrec de Joan Torruella i es compta amb la col·laboració d'especialistes en llengua i literatura medievals com la Dra. Lola Badia de la Universitat de Barcelona i el Dr. Jeremy N. H. Lawrance de la Universitat de Manchester. La finalitat del tractament informàtic dels documents catalans medievals és disposar d'una base de dades amb totes les paraules que hi apareguin, acompanyades d'altres informacions complementàries, tant històriques (primera documentació, zones lingüístiques dels seus usos, autors, gèneres on apareixen, etc.) com lexicogràfiques (grafia, categoria gramatical, accepció, etimologia, etc.).

El *Corpus*

Un dels primers problemes teòrics que hem hagut de resoldre és la definició del tipus de *corpus* que volíem, per tal que fos adequat als objectius que ens proposàvem. Pensant en la utilitat de l'*Arxiu* per a un bon nombre d'estudiosos, es va decidir elaborar un *Corpus Selectiu* de

tots els documents literaris, no literaris i notarials de la llengua catalana medieval. Això ens permet fer estudis sincrònics i sectorials de la llengua. Però dins d'aquest *Corpus Selectiu* hi englobem un *Corpus Exhaustiu* dels textos literaris.[1] La decisió d'incloure aquest *Corpus Exhaustiu* no és només pel gran valor lingüístic que pot tenir, sinó també per l'enorme interès que per als especialistes de la història de la literatura pot representar poder disposar dels textos de tots els escriptors medievals, en suport informàtic.

Cal puntualitzar, però, que es pretén que el *corpus* literari sigui exhaustiu pel que fa als *textos,* no pas pel que fa als *documents* que ens els fan arribar. Entenem per *text* la composició original de l'autor i per *document* el manuscrit o imprès que ens el conserva. En el cas dels autògrafs coincideix *text* i *document,* però en els altres casos el *document* sempre és posterior al *text* i, és clar, un text pot estar a diversos *documents*. Per això, en principi, de cada text s'escollirà el document o documents que es considerin més interessants (normalment, seguint un *stemma codicum*, el més proper de l'original).

Els punts que hem definit per dissenyar el *Corpus Selectiu* són els següents:

Recopilació:

— Extensió cronològica del Corpus.
— Extensió material del Corpus.
— Nombre de gèneres en què es dividirà el Corpus.
— Gèneres en què es dividirà el Corpus.
— Nombre de documents que s'elegiran per a cada gènere.
— Quins documents s'elegiran per a cada gènere.
— Part i extensió que s'elegirà de cada document.

1. *Corpus Selectiu* és aquell que es fa a partir de la representació estadística compensada de diversos grups, prèviament definits, de *documents*, ja sigui mitjançant el model proporcional o el model rectangular. *Corpus Exhaustiu* és aquell que, dels diversos grups prèviament definits, agafa la totalitat de *documents*.

Estadística:

— Tipus de mostra - Proporcional.
- Rectangular.

Lexicografia:

— Unitat lèxica.
— Criteris de lematització.
— Context.
— Referències.

L'ARXIU I ELS SEUS LÍMITS

Alguns aspectes que ha calgut decidir han estat el de la delimitació temàtica i el de la delimitació temporal dels *textos* que havien de formar el *corpus*. És clar que, almenys d'entrada, no podíem pretendre abastar tots els documents de qualsevol gènere on hi haguessin paraules catalanes. Per altra part, el terme medieval és un terme cronològicament ambigu i calia delimitar-lo. En l'aspecte temàtic s'ha decidit, amb el propòsit de formar el *Corpus Exhaustiu* de textos literaris, incloure al *corpus* tots els *textos* (no *documents*) escrits amb intencions literàries, mentre que dels no literaris se n'agafarà una mostra estadísticament representativa de cada gènere, de manera que tots els registres i nivells de la llengua hi estiguin representats. En el *Corpus Selectiu* es donarà preferència a aquelles obres que tot i no ésser estrictament literàries, o sigui, no haver estat escrites amb intencions literàries, avui, pel seu valor cultural, figuren en els manuals de la història de la literatura (per exemple, *Les Homilies d'Organyà* o les cròniques).

Tot i així, la inclusió d'algunes obres podria ser objecte de discussió, tant pel fet d'haver estat escrites per autors catalans però en llengua no catalana, com per ser *documents* amb barreja de paraules de diverses llengües, entre aquelles la catalana. En el primer cas, si bé en un principi s'havia pensat deixar de banda aquests *documents*, perquè no implicaven la llengua catalana, creiem que és important tenir-los al

corpus per a poder traçar el camí fet per algunes paraules arribades d'altres llengües, bàsicament del provençal. El segon cas, el de documents amb barreja de paraules de diverses llengües, pensem que aquests també poden ser força interessants per a la història de la llengua i, per tant, seran objecte igualment del nostre interès. Però en els casos de cançoners que també contenen poesies no catalanes agafarem només les que siguin catalanes. Els casos més difícils de decidir són els que impliquen el provençal, ja que moltes vegades la frontera entre les dues llengües no és gens clara.

Pel que fa al *Corpus Selectiu* dels textos no literaris, la representació estadística compensada dels diversos registres de la llengua implica una aportació més o menys proporcional dels diferents gèneres i nivells, o, el que és el mateix, implica la representació compensada dels *documents* dels diferents gèneres que componen aquest *corpus*.

Quant a l'aspecte cronològic, podem dir que, a grans trets, ens interessen des dels primers *textos* i *documents* on apareixen paraules catalanes fins als *textos* anteriors al segle XVI. Aquests *textos*, però, algunes vegades poden estar en *documents* del segle XVI o posteriors, no solament en manuscrits, sinó també en incunables o edicions.

Tot això ens ha portat a la necessitat d'elaborar un catàleg de tots els *textos* susceptibles de ser inclosos al *corpus*.

Les Edicions

Perquè el projecte sigui viable i àgil, és necessari que els *documents* siguin passats a suport magnètic i que tinguin una forma homogènia. Així facilitarem la posterior ordenació i catalogació de les paraules.

Ha estat necessari, doncs, decidir quin tipus d'edició havíem d'utilitzar a l'hora de transcriure els manuscrits, pensant, sobretot, que aquestes edicions havien de servir per a un corpus amb finalitats documentals. Ens vàrem decidir per l'*edició interpretativa*, entenent com a tal l'edició crítica que edita un sol manuscrit. Si bé per fer aquesta edició ens podem servir d'altres còdexs, caldrà marcar clarament les paraules que no estiguin documentades en el manuscrit base. Les edicions paleogràfiques, usades en altres projectes semblants, no ens inte-

ressen, ja que comporten diversos perills, sobretot els de mantenir els errors dels copistes i documentar *paraules fantasma*. Per altra part, l'edició crítica basada en la confrontació de diversos manuscrits, segons el sistema lachmanià, ens presenta un text ideal però que no es refereix a cap *document* en concret, de manera que resulta molt difícil atribuir a les unitats lèxiques referències geogràfiques i temporals concretes. Recalquem que estem inventariant les unitats lèxiques dels *documents* que ens han conservat els *textos* i que, per tant, atribuïm a cada unitat les dades espàcio-temporals dels *documents*, no les dels *textos* ni les dels autors.

També, per evitar interpretacions errònies, en les nostres edicions caldrà marcar totes aquelles paraules que no siguin catalanes. Això ho fem per estalviar-nos el perill de barrejar paraules de diverses llengües, en les quals podrien coincidir grafies de mots amb significats diferents (*casa* = llatí *cabana* / *casa* = català *casa*).

Però, de totes maneres, les normes d'edició que proposem —de les quals per problemes d'espai no parlarem aquí— respecten sempre els criteris filològics i són prou àmplies com per poder prendre com a punt de partida edicions que els filòlegs fan per als seus interessos acadèmics personals, de manera que amb molt poques modificacions siguin aptes per al nostre Arxiu.

El Text

D'aquesta manera, en una edició preparada per al tractament informàtic i destinada a extreure'n una concordança per a la posterior lematització i emmagatzemament de les seves paraules en una base de dades, hi podem trobar quatre tipus d'intervencions. Aquests quatre tipus caldrà tractar-los de manera diferent segons el paper que vulguem donar a les unitats lèxiques de cada un d'ells. Els quatre tipus són: a) el text de l'autor que constitueix l'obra, b) addicions posteriors degudes al mateix autor, c) addicions degudes al copista o a altres mans, i d) addicions degudes a l'editor. Així al text preparat per a l'ordinador pot haver-hi:

— Parts de text que volem que les seves paraules apareguin en la concordança com a lema i com a context d'altres lemes. Ex: el text del ms.
— Parts de text que volem que les seves paraules apareguin com a context dels altres lemes però que no apareguin com a lemes per elles mateixes. Ex: reconstruccions de l'editor, paraules estrangeres, etc., sense les quals no s'entendria el context d'un lema.
— Parts de text que volem que les seves paraules no apareguin en la concordança ni com a lema ni com a context d'altres lemes. Ex: glosses posteriors, cites afegides, etc.
— Parts del text que són només senyals que l'editor posa per facilitar les tasques informàtiques. Ex: inici de capítol, etc.

També, en una edició preparada per al tractament informàtic i destinada a extreure'n una concordança per a la posterior lematització de les seves paraules, podem trobar-hi cinc tipus de caràcters diferents segons la incidència o funció que vulguem que tinguin en el procés d'ordenació i lematització de les paraules:

— *Caràcters normals*: són aquells que volem que intervinguin en l'ordenació alfabètica de les paraules. Ex: a, b, c...
— *Caràcters diacrítics*: són aquells que volem que distingeixin entre homònims i, en conseqüència, consideren la paraula que porta el caràcter diacrític com un lema diferent a la mateixa paraula sense el caràcter diacrític i les ordenen una després de l'altra. Ex: accents, guions, etc.
— *Caràcters aleatoris*: són aquells que volem que formin part de les paraules però no volem que afectin per res la seva ordenació. Així, una paraula amb caràcters aleatoris quedarà ordenada dins el lema format per la mateixa paraula sense els esmentats caràcters. Ex: el grafema h quan no volem que sigui un element distintiu entre dues grafies.
— *Caràcters superflus*: són aquells que no volem que afectin l'ordenació alfabètica i que ni tant sols volem que apareguin escrits en les llistes i concordances. Ex: la barra transversal (/) que indica el canvi de foli.

— *Caràcters delimitadors*: són aquells que, vistos des de la perspectiva informàtica, volem que indiquin els límits de les paraules i que no afectin en res la seva ordenació. Ex: els signes de puntuació, l'espai, les cometes, etc.
(vegeu annex 2).

Funcionament

Per organitzar el treball, primer de tot, seleccionem, del *text* que volem editar, el *document* o *documents* més interessants que el contenen. Com ja s'ha dit, tot i que a vegades ens hem de servir d'altres manuscrits del mateix text per resoldre la lectura d'allò que en el manuscrit és il·legible o està mutilat, només documentem les paraules presents en el manuscrit base per tal de no posar en el mateix sac mots pertanyents a copistes, èpoques o zones geogràfiques diferents. Al manuscrit 9 de la Biblioteca de Catalunya, per exemple, hi ha un poema de R. de Cardona on la paraula *sen* no té sentit en aquell context; recorrem llavors al manuscrit Esp. 225 de la Bibliothèque Nationale de París que conté el mateix poema i, d'acord amb aquest, restituïm *sen* per *sens*, però no documentem aquest mot com a propi del manuscrit 9 ja que en realitat, tot i que és evident que aquest és el mot que hi ha d'anar, no hi consta. En la concordança, aquest mot apareixerà com a context de les altres paraules de la mateixa frase però no apareixerà com a lema.

Entrem, doncs, el *document* seleccionat dins l'ordinador; això ho fem amb un reconeixedor òptic de caràcters OCR (*scanner*) o bé manualment segons si disposem o no d'edicions fiables dels *documents* que volem entrar. Un cop tenim el text dins l'ordinador, un programa en fa la concordança (*Oxford Concordance Program*); és a dir, llista, com a lema, cada grafia diferent del *document* en ordre alfabètic i amb la seva freqüència d'aparició al costat. A sota de cada lema hi posa totes les frases del *document* (context) en què aquest lema apareix, amb, al principi de cada frase, la referència d'obra, pàgina i línia on el podem trobar (vegeu annex 1).

Després, un programa de lectura automàtica, anomenat *Bellaterra-90*, passa el resultat de la concordança a una base de dades, de manera

que es pot utilitzar la informació d'una forma molt més flexible i versàtil i se n'hi pot afegir automàticament. En aquesta base de dades hi tenim, d'una banda, els camps amb les informacions provinents de la concordança: *grafia de la paraula-lema, número de freqüència, frase de context i localització de la frase en el document*, a més dels camps amb les informacions històriques que es desprenen del document mateix comunes a totes les paraules: *autor, gènere, sigla del manuscrit, data i zona lingüística*. D'altra banda, tenim també a la base de dades uns camps lingüístics que s'ompliran, posteriorment, semiautomàticament: *lema, accepció, categoria gramatical, construcció, subcategoria, entrada, etimologia* i *traducció al català modern*. Així, per exemple, la grafia «puys», que trobem en el ms. 151 de la Biblioteca de la Universitat de Barcelona, que conté la prosa de Romeu Llull, una vegada passada a la base de dades portaria les següents informacions: *grafia*: *puys*; *freqüència al document*: 1; *frase de context: però, puys en lo portal portes no viu ne persona alguna; referència de la frase al document: pag. 3, lín. 38; autor: Romeu Llull; gènere: prosa; font: Barcelona, Bib. Univ. ms. 151; data: 1486; zona lingüística: central; lema: puys, accepció: 1; categoria: conjunció; entrada: puys; etimologia: *postius; català modern: puix.*

Per omplir els camps lingüístics, hem creat un programa, anomenat *TRANSCALC*, que compara les fitxes de la base de dades del document en què estem treballant (dB particular) amb les fitxes de la base de dades de l'*Arxiu* (dB general). La base de dades de l'*Arxiu* està composta per les fitxes de tots els documents que ja hi hem entrat fins aquell moment i que ja tenen, per tant, els camps lingüístics plens. Així, el programa compara les fitxes de la dB particular amb les de la dB general i, quan en troba una de la dB particular amb la mateixa grafia que una de la dB general, copia la informació dels camps lingüístics de la dB general a la fitxa de la dB particular o, en el cas que una determinada grafia de la dB particular no tingui cap parella igual en la dB general, el programa la marca amb un doble zero (00). D'aquesta manera, el filòleg només ha d'omplir els camps lingüístics de les fitxes marcades amb el doble zero i ratificar o rectificar les dades afegides automàticament de les fitxes de la dB particular. En el cas dels homònims, paraules amb la mateixa grafia però amb informa-

cions lingüístiques diferents, el programa copia totes les possibilitats, de manera que l'editor només haurà de triar la solució correcta.

Un cop fet tot aquest procés, les fitxes del nou document (dB particular) se sumen a les fitxes de la dB general (dB Arxiu) que conformen el *Banc de Paraules*. D'aquesta manera, com més documents passin a la dB Arxiu, menys possibilitats hi haurà de trobar fitxes doble zero quan utilitzem el programa amb nous documents.

Interessos del Banc

Hi ha dos aspectes bàsics en els quals el *corpus* d'un període determinat ens pot ajudar: (i) per proporcionar-nos exemples d'alguns tipus d'estructures o contexts característics de la llengua d'un determinat temps i lloc; (ii) per fer possible comparar dades que poden definir les relacions entre algunes variants i la norma.[2]

De la base de dades del *Banc de Paraules* se'n poden extreure moltes informacions, no solament per als filòlegs sinó també per als historiadors, juristes, historiadors de la literatura, etc., ja que es poden obtenir d'una manera fàcil i sistemàtica tot un seguit de dades lingüístiques, d'estil i històriques de cada mot. Aquestes informacions van des de les més simples, com saber quin és el primer document en el qual trobem escrita una determinada paraula, en quina zona lingüística s'usava, els autors que l'han utilitzat o les diferents formes gràfiques en què s'ha escrit, fins a combinacions força més complexes, com, per exemple, saber en quines regions s'utilitzava una determinada combinació de mots durant una època concreta, o bé, veure l'ús de la lletra 'h' entre vocals en els noms i adjectius de les obres en poesia dels autors valencians posteriors al 1450 i anteriors al 1500.

El servei que el *Banc* pot fer de cara a la confecció d'un diccionari històric també és considerable, donat que ens pot facilitar el coneixement de les diferents variants gràfiques que una paraula ha pres segons

2. L. T. Milic, *The Century of Prose Corpus*, «Literary & Linguistic Computing», Vol. 5, No. 3 (1990), pàgs. 203-208.

la zona lingüística en què s'ha escrit i ordenar-les cronològicament, o saber els diferents significats que un mot ha tingut segons els diversos autors i les distintes èpoques en què el trobem documentat. I sempre, i això és molt important, aquestes informacions s'acompanyen amb les cites textuals que les certifiquen.

Del *Banc* també se'n poden derivar sub-repertoris de temes específics o de llenguatges sectorials, com podria ser el cas d'un sub-repertori de textos notarials o de textos mèdics per a investigadors interessats només en aquests temes.

És important de mencionar el fet que el projecte és un treball obert i, per tant, sempre s'hi poden incorporar nous *documents* i noves infomacions sense que això impliqui fer cada vegada des del principi noves ordenacions i noves estadístiques, ja que el reciclatge i l'actualització de l'*Arxiu* són pràcticament automàtics.

JOAN TORRUELLA

ANNEX 1 - CONCORDANÇA

cel 1

1 2 P.Ma Yo ·m maravell com no ·s ve qui huylls ha, E cel qui -s ou perquè no vol entendre, E qui no sab perquè no

cell 2

5 5 B.Ce r No *mostrar may a -z altri s' amor pura Sinó ·s a cell qui porta benvolensa De cor e cors, e de tot son talan,
1 4 P.Ma vol entendre, E qui no sab perquè no vol apendre, E cell qui pot e sap, com bé no fa, E valent hom que fassa gran

com 3

1 4 P.Ma no sab perquè no vol apendre, E cell qui pot e sap, com bé no fa, E valent hom que fassa gran aulesa Per foll pl
1 1 P.Ma Yo ·m maravell com no ·s ve qui huylls ha, E cel qui -s ou perquè no vol ent
6 4 J.S. n fat pensamén Seguint Amor e son foll mandamén, Sí com *hom cech, volent ço que val poch; Mas ben és foll qui vo

cor 3

1 7 P.Ma ll plaser qui dura pauch momén, E de senyor qui pert cor de sa gen Per crueltat o per mal' averesa
5 6 B.Ce s'amor pura Sinó ·s a cell qui porta benvolensa De cor e cors, e de tot son talan, Car si altruy fa sguard ne be
6 2 J.S. En mal poders, enqueres en mal loch, Hay mis mon cor e mon fat pensamén Seguint Amor e son foll mandamén, Sí

cors 1

5 6 B.Ce r pura Sinó ·s a cell qui porta benvolensa De cor e cors, e de tot son talan, Car si altruy fa sguard ne bell sem

crueltat 1

8 P.Ma pauch momén, E de senyor qui pert cor de sa gen Per crueltat o per mal' averesa

ço 1

6 4 J.S. t Amor e son foll mandamén, Sí com *hom cech, volent ço que val poch; Mas ben és foll qui vol haver paria Ab dona

De 6

5 6 B.Ce ri s' amor pura Sinó ·s a cell qui porta benvolensa De cor e cors, e de tot son talan, Car si altruy fa sguard ne
5 1 B.Ce Dona gentils e d' excellén natura Que viu tots jorns en amorós voler, Deu p
1 7 P.Ma laser qui dura pauch momén, E de senyor qui pert cor de sa gen Per crueltat o per mal' averesa
1 7 P.Ma gran aulesa Per foll plaser qui dura pauch momén, E de senyor qui pert cor de sa gen Per crueltat o per mal' aver
5 6 B.Ce Sinó ·s a cell qui porta benvolensa De cor e cors, e de tot son talan, Car si altruy fa sguard ne bell semblan, A
6 6 J.S. ben és foll qui vol haver paria Ab dona vils, plena de tritxaria, E pus foll és qui vol amor servir Punt leyalme

esconexença 1

5 8 B.Ce fa sguard ne bell semblan, Al qu' és primer fa greu desconexença

Deu 1

5 3 B.Ce excellén natura Que viu tots jorns en amorós voler, Deu per tostemps metre tot son poder No *mostrar may a -z alt

Dona 2

5 1 B.Ce Dona gentils e d' excellén natura Que viu tots jorns en amoró
6 6 J.S. ue val poch; Mas ben és foll qui vol haver paria Ab dona vils, plena de tritxaria, E pus foll és qui vol amor ser

dura 1

1 6 P.Ma valent hom que fassa gran aulesa Per foll plaser qui dura pauch momén, E de senyor qui pert cor de sa gen Per cru

ANNEX 2 - EXEMPLES DE LLISTES DE FREQÜÈNCIES

... ...	n
aiçí	6
aymador	2
aimant	2
aymant	3
aymar	1
aimia	1
air	2
ahir	4
aire	2
aysí	4
... ...	n

... ...	n
aiçí	6
aymador	2
aimant	5
aymar	1
aimia	1
air	6
aire	2
aysí	4
... ...	n

En el tros de llista de freqüències alfabètica de l'esquerra la *y* ha estat considerada com un caràcter diacrític de *i*, i la *h* com un caràcter diacrític general, mentre que a la dreta hi ha la mateixa llista però el programa ha considerat la *y* igual que la *i*, i la *h* com un caràcter aleatori (v. 1.3).

... ...	n
deus	2
déus	4
dia	3
dient	2
dihent	1
dies	1
dyes	2
*diffam	1
diffamada	1
difamador	1
diffaman	1
difamar	3
diffamar	1
... ...	n

... ...	n
deus	4
dia	3
dient	3
dies	3
diffamada	1
difamador	1
diffaman	1
difamar	4
... ...	n
*diffam	1

Al tros de llista de freqüències alfabètica de l'esquerra el programa ha considerat la *e* accentuada com a caràcter diacrític de la *e* sense accentuar, mentre que a la llista de la dreta el programa ha igualat la *e* accentuada a la *e* sense accentuar; la *h* a la llista de l'esquerra és considerada com un caràcter diacrític en general, mentre que a la llista de la dreta està considerada com un caràcter aleatori; la *y* a la llista de l'esquerra és un caràcter diacrític de la *i*, mentre que a la dreta el programa ha considerat iguals la *y* i la *i*; l'*asterisc* a la llista de l'esquerra ha estat considerat o com un caràcter diacrític o com un caràcter aleatori; en canvi, a la llista de la dreta ha estat considerat el darrer caràcter de l'alfabet; la *dobla efa* a la llista de l'esquerra és un diacrític de la *efa simple*; per contra, a la de la dreta la *doble efa* és un caràcter igual a la *efa simple*.