Teoria i pràctica de la lematització en el Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis-IVITRA[117]

Jordi M. Antolí Martínez

1. Els corpus textuals informatitzats com a eina del filòleg

Les fonts per a reconstruir l'evolució d'una llengua són, desgraciadament per a la filologia, més aviat escasses. Bàsicament tenim la documentació escrita, que per bé que pot ser ben diversa (documents jurídics i administratius, religiosos, literaris, documentació privada i també científica, etc.), no és del tot fiable si es pretén usar-la amb valor de transcripció per a reconstruir la realitat lingüística pretèrita. I llevat d'aquestes fonts escrites, l'investigador té ben poca cosa més on agafar-se, tot i que —és cert— no hem d'oblidar algunes altres fonts orals que, inserides en el nostre discurs contemporani, ens remeten a estadis anteriors de la llengua: la fraseologia i l'onomàstica, a més de l'estudi comparatiu de les diferents variants dialectals d'una llengua. Tot plegat, però, faves comptades.

Amb una limitació tan evident de fonts a les quals recórrer, al filòleg interessat en l'estudi diacrònic o sincrònic de la llengua del passat solament li resta aprofitar tan bé com puga les mostres que té a l'abast per a formular les hipòtesis de treball. I en l'actualitat, atès el desenvolupament que han viscut les tecnologies de la informació i la comunicació,

[117] Aquest article és un dels resultats de la tasca desenvolupada arran de la concessió de la beca d'iniciació a la investigació del Vicerectorat d'Investigació, Desenvolupament i Innovació de la UA (de l'1 d'octubre del 2011 al 31 de desembre del mateix any) i la consegüent incorporació de qui el signa com a becari del Departament de Filologia Catalana de la UA al Grup d'Investigació «Traducció de clàssics valencians a llengües europees: estudis lingüístics, literaris i traductològics comparats» (VIGROB 125), a l'Institut Superior d'Investigació Cooperativa «Institut Virtual Internacional de Traducció» (ISIC/2012/22) i, al seu si i com a col·laborador, a la *Multilingual Digital Library of the Mediterranean Neighbourhood-IVITRA* (MICINN FFI2010-09064-E) i al Digicotracam (Programa Prometeo de la Generalitat Valenciana per a Grups d'Investigació en I+D d'Excel·lència, projecte cofinançat pel FEDER de la UE, ref. Prometeo-2009-042).

desaprofitar les aplicacions que aquestes ofereixen és, si més no, una irresponsabilitat. Fa poc més d'una dècada, l'única manera de garbellar un mostrari suficient d'ocurrències a partir de les quals formular les hipòtesis era espigolant, llapis en mà, en un o més textos les mostres necessàries. Avui, l'aparició de *corpora* textuals informatitzats com el Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis-IVITRA o, associat a aquest, el Corpus Informatitzat del Català Antic (CICA) per a la llengua catalana medieval permeten aconseguir en pocs segons un mostrari d'ocurrències suficientment fiables extretes de centenars d'obres[118] per a la llengua catalana medieval. Pensem el temps que hauria d'esmerçar l'investigador a aconseguir aquestes mostres a la manera tradicional.

La dels corpus lingüístics informatitzats és una àrea d'investigació puixant en els darrers vint anys. L'avantatge més evident d'aquests productes és el de permetre processar grans recopilacions de textos fent ús de mitjans computacionals per a processar-los (enriquir-los amb informació diversa: categorització, lematització, assignació d'idioma...) de manera que, una vegada sistematitzats i ordenats, aquests textos enriquits puguen aprofitar, fent ús d'unes eines apropiades de recuperació de la informació, com a base empírica per a assolir unes conclusions tan objectives com siga possible.

Ara bé, el desenvolupament de corpus com el que ens ocupa, el Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis, té una dificultat afegida, i és que mentre que el gruix dels corpus construïts fins al moment se centren en la llengua actual,[119] especialment en textos literaris —amb un model de llengua ortogràficament regularitzat—, aquests altres corpus pretenen servir com a base per a l'estudi de la

[118] Val a dir, amb tot, que els resultats que es poden obtenir amb un corpus o amb l'altre no són els mateixos, atès que les eines de gestió de corpus i de recuperació de la informació desenvolupades al si de l'IVITRA tenen un potencial major que les del CICA. S'ha de tenir present que el processament que es fa dels textos, que és del que parlarem en el present article, és fonamental, però és evident que tant o més importants són les aplicacions creades per a la recuperació automàtica de la informació, ja que aquestes determinen decisivament el tipus de cerca que podrà fer l'investigador.

[119] Valga com a exemple el CORDE de la RAE.

llengua antiga en diacronia.[120] Aquesta particularitat comporta uns reptes evidents, i és en aquest sentit que les aportacions que s'hi han fet poden merèixer una certa reflexió, que és la que dóna lloc al present article.

2. El Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis i les eines que el fan possible[121]

El projecte institucional de recerca IVITRA (Institut Virtual Internacional de Traducció) de la Universitat d'Alacant ha estat pioner en l'aplicació de les TIC en el processament de textos medievals, concretament de llengua catalana. Per a fer-ho possible, s'ha desenvolupat al si del projecte una nova generació d'eines informàtiques amb l'objectiu de crear i gestionar *corpora* textuals. Només per esmentar-ne alguns dels ja efectius podríem citar els programes Introcorpus®, Metatagging®, Ivitratech® i el Metaconcor®. A grans trets, aquesta tecnologia permet introduir, processar, emmagatzemar i recuperar de manera selectiva la informació d'un corpus de textos. L'Introcorpus® és el programa destinat a l'edició i introducció de textos en el sistema. Realitza una primera edició del text, prèvia a la seua incorporació en el corpus, un procés que implica la revisió del text, la selecció de l'idioma, la fixació d'un format (amb salts de pàgina, tabulacions, capítols...). Una de les funcions d'aquesta aplicació és portar a terme automàticament una primera anotació morfosintàctica i lematitzar les paraules del text que s'introdueix. Aquesta assignació automàtica de la categoria gramatical (nom, quantificador, adjectiu...) i del lema («deure», «per», «déu»...) s'aconsegueix mitjançant diccionaris generats prèviament a partir de la informació afegida a les formes lèxiques dels textos ja introduïts. En definitiva, el procés dut a terme mitjançant aquesta eina conclou amb el text incorporat a la biblioteca digital.

Una vegada introduït, el text s'incorpora a l'eina Ivitratech®, el dipòsit que conté el corpus textual (en el nostre cas, el Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis); en altres paraules,

[120] La funció primera de tots dos corpus és la d'aprofitar com a base empírica per a l'elaboració de la Gramàtica del Català Antic.

[121] Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis, dirigit per Vicent Martines & Josep Martines, IVITRA, Universitat d'Alacant.

és una base de dades on s'emmagatzemen totes les obres incorporades; el corpus elaborat per l'IVITRA conté en aquest moment un total de 192 obres literàries procedents de l'àmbit cultural de l'antiga Corona d'Aragó, moltes d'aquestes valencianes. És sobre aquesta biblioteca digital que operen la resta de programes. El Metatagging®, per la seua banda, és el programa que permet resoldre les deficiències de l'etiquetatge dut a terme per l'Introcorpus®. La possibilitat de trobar formes que no hagen estat introduïdes prèviament en el corpus i, per això, que el programa no reconega, o el fet que hi haja paraules homògrafes justifica la necessitat de corregir o completar el treball fet automàticament. Ara s'introduirà la informació (categoria, lema, idioma) de forma manual i es podrà decidir si el canvi mereix o no ser extrapolat dintre del text o del corpus, o no.

És en aquest punt que ens aturarem per analitzar amb més deteniment una de les fases de l'etiquetatge, la de la lematització, que es duu a terme per construir aquest Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis al si de l'IVITRA.

3. La lematització en el nostre corpus: sentit i problemàtica

Per a entendre la problemàtica de la lematització s'ha de partir del sentit que aquesta té, un sentit que potser no resulta tan evident pel fet que el corpus de què partim té com a objectiu no l'elaboració d'un diccionari, sinó el fet de servir de matèria digerida per a eines de recuperació selectiva de la informació. Doncs bé, la importància d'assignar un lema a les diferents formes del paradigma (gràfic, flexiu) d'un mot ve donat pel fet que aquest lema, siga quin siga —tant fa que siga la forma canònica o simplement una xifra— actua de nexe que enxarxa totes les paraules que el comparteixen. Això comporta que, d'aquesta manera, poden ésser recuperades alhora totes les formes que integren el paradigma flexiu o gràfic d'un mot amb la introducció en la eina de recuperació (en aquest cas el Metaconcor) d'una forma qualsevol (i no necessàriament el lema). Hom pot demanar, per exemple, que a partir de la forma «caval» es mostren totes les ocurrències de totes les possibilitats gràfiques i flexives d'aquest substantiu: «cavals», «cavall», «cavalls», «cavayl»... I açò, que pot semblar una beneitura (hom cercaria, si es donés el cas,

«cavall» i no «cavalls» o «caval» per recuperar tot el paradigma), no ho és tant si tenim en compte que l'usuari de les eines no necessàriament sap exactament quin és el lema amb què pot trobar el mot: «malestruc» o «malastruc»? Totes dues formes, introduïdes al Metaconcor, oferirien totes les formes amalgamades pel lema (en aquest cas, «malastruc»). A partir d'açò que hem dit, podríem coincidir amb Elena Sánchez (2010: 273) que el lema, en el nostre corpus informatitzat, és la:

> Representació formal del conjunt de formes flexionades i de variants gràfiques o històriques d'un mot, que serveix com a clau abstracta interna per a relacionar-les entre si i en possibilita la recuperació conjunta.

Quedem-nos, doncs, amb la idea que el lema no és visible a l'usuari de les eines de recuperació, però és fonamental.

Consegüentment, si això és el lema, la lematització serà el procés pel qual s'assigna un lema a aquelles formes que es decideix, per factors filològics i també de recuperació de la informació, que formen una unitat. D'entrada podríem considerar que la tasca de la lematització no és gaire complicada. Més ben dit, no és una mampresa complicada si ens cenyim a la llengua estàndard; de fet, llevat d'algunes desambiguacions inevitables, la lematització d'uns textos amb un model de llengua estàndard es pot dur a terme de manera automàtica introduint, simplement, la informació d'un diccionari normatiu, els paradigmes de flexió de la llengua i extrapolant l'etiquetatge a totes les formes del corpus. Ara bé, aquesta tasca, senzilla si prenem uns textos d'una llengua ortogràficament regularitzada, esdevé bastant més complicada si els mots del corpus textual no acompleixen aquest requisit. Els problemes, de manera més concreta, que afecten la lematització dels mots de textos antics i que impedeixen l'automatització total d'aquesta tasca són:

a) La presència d'unitats lèxiques absents dels diccionaris normatius i descriptius existents pel fet que són arcaics, dialectals, col·loquials, etc.

b) La variació formal (morfològica, fonètica) dins del corpus i respecte de la llengua actual fruit de l'evolució de la llengua i de l'absència d'una normativa gramatical establerta.

c) La variació gràfica pròpia d'un estadi no normativitzat de la llengua.

d) La multiplicació de les formes lèxiques —atesa la variació que es dóna en la llengua antiga— augmenta la possibilitat d'existència d'homògrafs i, per tant, dificulta la possibilitat de fer extrapolacions.

Tots aquests factors alenteixen, doncs, la tasca de l'etiquetatge, ja que per cada forma que puga prendre un mot s'ha d'assignar el lema. Per exemple, l'etiquetatge del quantificador «poc» comporta no solament extrapolar el lema i la categoria a totes les formes que coincidisquen amb aquesta, sinó també a les variants gràfiques que pot haver-hi, com «poch». Més encara, l'extrapolació indiscriminada a tot el corpus comportaria ignorar la possible homografia que hi pot haver entre el quantificador i la tercera persona del singular del pretèrit perfet simple del verb «poder» en la conjugació antiga: ell «poc». No és, doncs, una faena ni fàcil ni ràpida.

Del que acabem de dir es desprèn que en l'elaboració d'un corpus textual informatitzat amb vocació diacrònica —que comporta, i és ací on rau el problema, el treball amb textos antics— l'automatització de la lematització (i de l'assignació de categories) pot ser molt limitada, mentre que guanya pes el treball «manual» que desenvolupen els membres i col·laboradors del projecte i que té com a objecte la desambiguació de les formes homògrafes o l'etiquetatge de les formes que no reconeix el diccionari que fa la primera batuda al text. La construcció d'aquest corpus, doncs, ha de situar-se a mitjan camí entre l'automatització total (ràpida, però indiscriminada) i l'etiquetatge exclusivament manual (exacte, però extremadament costós i lent). La faena al detall queda reservada, com hem dit adés, per als casos que no han estat lematitzats automàticament (bé perquè són nous en el corpus, bé perquè no se n'ha fet l'extrapolació perquè hi ha homografia). Aquesta tasca de l'etiquetatge manual (i, en concret, de la lematització) que es duu a terme en la construcció del nostre corpus comporta dos moments:

1) La identificació del mot: quina és la categoria i amb quin lema el trobaríem en les fonts de consulta. Aquest moment és el que diferencia el treball al detall de l'automàtic —que no permet aquesta minuciositat—,

ja que l'anàlisi del context permet l'assignació exacta de la categoria i, per tant, del lema.

2) A partir de la informació que ja tenim, hem de ser capaços d'atribuir a la forma un lema que responga als criteris adoptats per a la confecció del corpus.

En les pàgines que resten ens centrarem a analitzar la problemàtica que comporta el primer pas, és a dir, la identificació del mot. El segon moment mereix un esment en un article a banda.

4. Les categories del corpus

A diferència de l'etiquetatge que es fa servir per a la constitució del diccionari de partida del corpus (la informació extrapolada en una primera batuda als textos)[122], el treball posterior de correcció i compleció que es duu a terme amb el Metatagging implica fer l'operació inversa, ja que es parteix d'una forma, se li assigna una categoria i, després, se li assigna també el lema. Per tant, la selecció del lema és posterior a la determinació de la categoria i depèn d'aquesta.[123] Abans de continuar, doncs, és necessari exposar quines són les categories que reconeix el nostre corpus i quina és la forma escollida en cada cas per a la lematització:

Adjectiu	Possessiu
Adverbi	Preposició
Article	Pronom

[122] En la construcció d'aquest primer diccionari, que s'enriquirà posteriorment amb la informació introduïda manualment, es parteix d'un lema, se li associa el paradigma corresponent i se li assigna una categoria. Posteriorment això s'extrapola als textos a mesura que aquests són introduïts.

[123] Per exemple, mentre que el procés dut a terme en la creació del diccionari implicava partir del lema «deure», associar-hi el paradigma i assignar-hi la categoria de verb, en la revisió feta amb el Metatagging partim de la forma «deu» i comprovem en cada cas la categoria que és: verb? Substantiu? Quantificador? El lema, doncs, variarà segons la categoria de què es tracte: si és un verb serà «deure», si és un substantiu, «déu» o «deu», segons el cas; si és un quantificador, «deu».

Coordinant	Subordinant
Demostratiu	Substantiu
Interjecció	Quantificador
Nom propi	Verb
Altres	

Algunes remarques que s'han de fer de l'assignació del lema en cada cas, són:

a) El lema de les categories article (personal o determinat), adjectiu i substantiu serà sempre el masculí singular (o femení singular en els casos que l'adjectiu o substantiu no tinga forma de gènere masculí).

b) Els pronoms (forts i febles) es lematitzaran a partir de les formes *jo, tu, ell, nosaltres, vosaltres // em, et, el, ens, us // li //en, es, hi, ho, si.*

c) En els quantificadors, el lema és —si n'hi ha— la forma del masculí singular; un cas particular és el dels numerals ordinals, que es lematitzen a partir del cardinal.

d) El lema dels possessius (febles o forts) serà, depenent del posseïdor: *meu, teu, seu, nostre, vostre, seu, llur.*

e) En el cas dels adverbis, preposicions, coordinants i subordinants, en general sol passar que el lema és igual a la forma; en els casos en què hi ha variació es recorre al lema que ofereix el diccionari.

f) Els noms propis es lematitzen de manera diferent si es tracta d'un antropònim o d'un topònim. Els topònims s'actualitzen, mentre que els antropònims bé s'actualitzen (en el cas de noms de reis, noms de personatges bíblics i, en general, de personatges històrics destacats), bé es mantenen amb la forma que prenen com a lema (en la resta de casos).

g) Els demostratius prenen com a lema les formes: *aquest, aqueix, aquell / això, açò, allò.*

5. Els treball filològic en la lematització d'un corpus lingüístic informatitzat[124]

A l'hora, però, de reconèixer la forma en qüestió, pot sorgir el dubte i no saber què és el que es té al davant. Els coneixements filològics i humanístics en general i el treball en equip contribueixen notablement a la resolució d'aquests problemes d'identificació, que vénen donats per factors diversos, semblants als que adés apuntàvem com a obstacles per a l'automatització de l'etiquetatge en un corpus textual informatitzat que continga textos antics: la variació diacrònica (sobretot morfològica i fonètica) o gràfica de les formes presents en els textos respecte de les formes que són estàndards en l'actualitat o el fet que hagen caigut en desús. Tot plegat disfressa les formes i ens les fa llunyanes, difícils d'identificar. Tot seguit farem una ullada general a aquests obstacles a partir de les dificultats concretes que hem anat trobant en la tasca de l'etiquetatge; no hi farem aportacions a l'estudi de la gramàtica històrica, sinó que explicarem, senzillament, en què consisteix la faena que fem.

En alguns casos aquesta variació gràfica és fruit de la manca d'unes convencions ortogràfiques fermes, i és sovint un dels aspectes que obstaculitza més el reconeixement de les formes amb què es treballa. Sense pretendre entrar massa al detall en la qüestió, s'ha de tenir present la problemàtica que comporta l'escriptura de les llengües romàniques i, en concret, del català en època medieval. I és que, per a la representació d'aquestes, es parteix de les grafies de representació del llatí, amb el buit que açò comporta a l'hora de representar els nous fonemes apareguts fruit de l'evolució del llatí vulgar i de la variació fònica en les llengües romàniques. Atenent a aquesta explicació s'entén el ball de grafies que es dóna a l'hora de representar el so /ʎ/, amb els dígrafs «ll», «ly», «yl», «lh» o senzillament una «l». Un cas semblant és el del fonema /ɲ/, representat «ny», «yn» o «nn» en paral·lel a l'anterior, però també apareix fins i tot com a «y». El so /ʤ/, per la seua banda, es representa amb les grafies «y», «j» o «g», mentre que la variació que es dóna en les grafies que representen i diferencien la /s/ i la /z/ també s'ha de tenir present.

[124] Tots els exemples que segueixen provenen del CIMTAC.

A tot açò s'hi ha d'afegir que el prestigi del llatí, juntament amb el fet que sovint els escrivents en llengua catalana ho eren també en llatí, també comportà l'assignació d'unes grafies d'acord amb l'etimologia i no amb la realitat fonètica, com l'ús del dígraf «ch» que representa el fonema /k/ a final de mot o en paraules d'origen grec, com «Christ»; podem trobar el dígraf «ph» representant el fonema /f/ atenent a factors etimològics (llatinismes que alhora vénen del grec) com en «triumph» (*EPV-II*, 85-90[125]), fins i tot una «p», com en «pantacia» ('fantasia' *DP*, 233, 13 i 233, 18). És també per factors etimològics o per donar llustre erudit que trobarem el dígraf «th» en mots com «thesauritzar» (*DdC*, 413, 17) ('atresorar'), «prothonotari» (*DdC*, 322, 19; *EVM-II*, 25, 7 etc.), «apothecari», etc. I bé, podríem continuar enumerant casos i casos en què es dóna una oscil·lació de grafies per a representar un mateix so: la grafia «y» com a iod, l'alternança *u*/*v* per a representar la consonant, l'ús no ben delimitat d'una o dues *r* per a remarcar si es tracta de la vibrant simple o múltiple...

Un bon exemple de com la variació gràfica pot disfressar l'aparença d'un mot és «alaugades» (104, 8), extret de la *Doctrina pueril* de Ramon Llull, on la grafia «l» representa el so /ʎ/ i la «g» és /ʤ/; si a açò afegim el ball de grafies que es dóna en els textos procedents del sector oriental del domini lingüístic fruit de la presència de la vocal neutra en contextos àtons, s'entén que la forma que s'amaga al darrere d'aquesta representació és el participi femení plural del verb «alleujar»; d'acord amb les convencions ortogràfiques actuals el representaríem com a «alleujades».

Aquest exemple ens obliga a parlar d'una altra mena de variació que pot dificultar la identificació del mot, i és la fonètica, és a dir, aquells

[125] D'ara en avant farem referència a les obres del Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis (CIMTAC) a partir de les abreviatures següents: *CD*: *Crònica Desclot*; *CeG*: *Curial e Güelfa*; *CiC-I*: *Clams i Crims a la València Medieval I*; *CM*: *Crònica Muntaner*; *DdC*: *Dotzè del Crestià*; *DP*: *Doctrina Pueril*; *EEJ*: *Edats i Epístola de Jesucrist*; *EVM-II*: *Epistolari de la València Medieval II*; *EPV-II*: *Epistolari de la València Medieval II*; *LlC*: *Llibre de totes maneres de confit*; *PC*: *Un matrimoni desavingut i un gat metzinat. Procés criminal barceloní del segle XIV*; *PM*: *Poemes* d'Ausiàs March; *PR*: *Memorial del Pecador Remut*; *SS*: *Sent Soví*; *TC*: *Tractat de Confessió* d'Antoni Canals.

casos en què es fa una representació fonètica del mot en qüestió i d'això resulta una variant gràfica d'aquest divergent del que convé la norma actual i divergent també respecte a d'altres possibles representacions per al mateix cas presents en el corpus. Dins d'aquest cabàs, la variació en la plasmació gràfica de la vocal neutra dels parlars orientals és una constant; no és estrany, doncs, trobar vacil·lacions com *adalita/adelita*, *àar/àer*, *epelen/apellen*, *deveylar/davallar*, *felir/fallir*, *aleta/eleta*, *alig/elig*, *alataren/alletaren*, *varmal/vermell*, *netures/natures*, *fedat/fadat*... En la identificació dels mots sovint tenim com a ajuda el fet que apareguen bimembracions sinonímiques del tipus «fedat o astrat» (*DP*, 226, 17). D'una altra manera, «fedat» podria interpretar-se, en el mateix context, com a «fedar» ('tacar'). Ara bé, «astrat», derivat d'«astre» ('sort o destinació prefixada a la vida humana, segons el cos celeste que se considera que la regeix', *DCVB* s. v. «astre») ens dóna la solució.

La variació gràfica que trobem en el consonantisme —en els casos que palesen la realitat fonètica de la llengua de l'època— és donada per factors més diversos que els del vocalisme, entre els quals destaquen les metàtesis, les dissimilacions i assimilacions a la sonoritat, punt i mode d'articulació o les elisions.

Així doncs, és freqüent, per exemple, l'alternança entre sonorització i manteniment d'oclusives sordes intervocàliques o en posició inicial; no és estrany trobar formes del tipus «bestonagues» amb la [p] sonoritzada (hem de tenir present que, a més, s'hi dóna un ball de grafies per a representar la neutra, el lema que atribuiríem en l'actualitat seria «pastanaga»)[126]. Les altres dues oclusives sordes les podem trobar també sonoritzades, com ocorre en «qüenditat» (*Ll*C: 285, 9; 285, 12; 286, 23; 290, 2; 291, 14; 291, 16) amb la [t] i en «reguonèxer» (*Ll*C: 275, 20; 277, 8; 282, 16)[127] amb la [k]. Pel que fa a aquest darrer fonema, apareix

[126] Aquesta forma, «bestonagues», la trobem en el *Llibre de totes maneres de confits* (280, 10; 180, 11; 280, 15; 280, 19; 280, 21; 280, 23; 281, 6); sobre el ball entre el fonema sord i l'equivalent sonor *p/b*, ja n'ha parlat J. Coromines en el DECat (*s. v.* «pastenaga»), que apuntava que la *b* era una grafia «arabitzada».

[127] J. Coromines ja en parla en el DECat (*s. v. conèixer*): «Reconèixer [fi s. XIII]: ja apareix tres o quatre vegades en Desclot (*reco-* almenys en part del mss.); en tota l'Edat

sonoritzat també en posició inicial en noms propis com «Gostantina» (Cocentaina) o «Gostança» (Costança)[128]. Hi ha canvis, però, que afecten també al mode d'articulació; per exemple, s'hi troben casos de /l/ en què, mantenint el punt d'articulació i la sonoritat, es canvia el mode d'articulació i esdevenen /ɾ/; en són exemples «closta» (*Ll*C: 286, 11) per «crosta» o «perar» per «pelar»[129]. De dissimilacions, en trobem del tipus «arble»[130] per «arbre», una solució, siga dit de passada, semblant a la d'«albre», comuna a diversos punts del domini lingüístic català.

Un altre fenomen constatable en els textos i que dificulta la identificació de les formes és el de la ultracorrecció; en són exemples la lateralització de la [w] en posició implosiva en casos com «alciure»[131] i tot

---

Mj., i fins molt més tard, lluiten en *rec-* amb les formes *regonèixer*: *regonèixer les mesures* a. 1358; Alart, *Inv*LC; que aplega encara una dotzena de formes en *-g-* en documents rossellonesos del s. XIV, des de *regonegueres* a. 1311, fins a *regonegren* 1379, amb algunes formes de conjugació tan arcaïtzant com *eu regonesc(h)* (Pres. 1), aa. 1375, 1382, 1383, 1411, junt, però, amb *regonech* Pres. 1, 1323, i *regonec* pf. 3, 375; *regonegut* també en les *Lleg. Rim.* De Sevilla, fi s. XIII (878)»; acaba constatant que la forma amb *-g-* encara es pot documentar entre els renaixencistes.

[128] Els exemples estan presos del *Sumari d'Espanya* de Berenguer de Puigpardines («Gostantina»: 119, 12 i 119, 16; «Gostança»: 106, 26; 116, 33 etc.). Aquesta relaxació de la /k/ inicial no és, però, estranya a la nostra llengua; per a veure'n més i més recents, vegeu els exemples que troba Josep Martines en el valencià del XIX (MARTINES: 2000, 174-181).

[129] Al DCVB s'hi comet, segons el nostre parer, l'error de remetre del lema «perar» a «parar» en lloc de a «pelar»; solament cal veure l'exemple que s'hi posa: «Pendreu los présechs e perar-los-heu ab un guauinet». Més encara, si es consulta el lema «pelar», al segon punt se'ns posa un exemple idèntic al nostre: «E lauors ab un drap pelar-les-heu [les ametles]». Per recolzar més encara aquesta proposta, vegem algun dels exemples següents (tots procedeixen del *Llibre de totes maneres de confits*, de la segona meitat del XV):

«Pendreu los codonys e trematreu-los al forn, e, aprés que seran ben cuyts, perar-los-heu»

«Pendreu les nous e perrar-les-heu, e, com sien perrades, forredar-les-heu»

«tu les pendras he perar-les-às ab un guanivet» [les peres]

[130] Aquest exemple concret està pres dels *Començaments de medicina* de Ramon Llull (primera meitat del segle XIV).

[131] Aquest exemple es troba en diverses obres: EEJ 1, 16; 144, 9; PR 153, 20; CiC-I 118, 21; CeG 238, 9. El procés pel qual es confongué la /l/ i la /w/ en posició implosiva

el paradigma verbal que se'n deriva: «alcís» (*CD* 169, 5; *CM* 164, 1), «alciuré» (*EEJ* 123, 4), «alciuria» (*CiC* 141, 16; *CeG* 152, 2), «alcia» (*EEJ* 288, 2; *CM* 592, 4), «alcieren» (*CD*152, 22), «alciam» (*PR* 188, 25), «òlsia» (*PC* 14, 20), «ouciés» (*PC* 33, 4) o «olciure» (*PC* 33, 10).[132]

A banda d'aquests fenòmens més o menys habituals en els textos catalans medievals, val a dir que s'hi troben també alguns mots que, pel fet que són arcaismes, impliquen també algun problema a l'hora de reconèixer-los. Per exemple, trobem els adjectius «entregue» (*PM* 38, 9; *SS*, 88, 7 etc.)[133] (encara viu a l'extrem meridional del domini lingüístic), o el substantiu «túixec». Pel que fa al segon, que trobem amb la forma «túxech», és a dir, l'evolució patrimonial del llatí TŎXĬCŬM,[134] el trobem

(on s'accentua la velaritat de la /l/) degué ser marcat socialment de manera negativa, de manera que els esforços per a corregir en l'escriptura la pronunciació «aut» per «alt», «autar» per «altar», «auba» per «alba» o, fins i tot, «augutzir» per «algutzir» comportaren en alguns casos la ultracorrecció, com passa en «alciure» per «aucire», on la [w] no és el resultat de la vocalització de la [l], sinó de la diftongació de la [o] àtona inicial en [aw], atès que la forma llatina de què procedeix és «occīd re».

[132] Aquests darrers tres exemples presenten la *o* i alhora la *l*<*w*. Procedeixen tots tres, dins del corpus informatitzat, d'*Un matrimoni desavingut i un gat metzinat. Procés criminal barceloní del segle XIV* editat i estudiat per Joan Anton Rabella i Ribas (1998). Val a dir que la primera forma, «òlsia», en realitat era analitzada com a «mólsia» per l'editor del text en la nota 765 de la pàgina 192, una forma del paradigma del verb *molsir*. Aquesta segona proposta —la lectura a partir del verb *occir*— és la que ja apuntà Josep Moran (2000) en la ressenya que féu de l'edició de Rabellas i és més coherent donat el context.

[133] INT GRUM> int gru (elisió de la nasal, s. I regla 40) > in't gru (accentuació del llatí vulgar, regla I; l'accent cau sobre la penúltima pel fet que en llatí tardà la seqüència d'obstruent+líquida es considera que travava la vocal precedent i, doncs, tot i que fos breu possibilitava que rebés l'accent) > en'tEgro (abaixament de les vocals laxes, regla III, i pèrdua posterior de la tensió, regla IV)> en'tEgr (regla XXXIV, reducció vocàlica 3)> en'tEgre (regla d'avançament de la [ ], la XL)> en'tegre (regla XIVb d'ascensió de la [E])> en'trege (metàtesi). Les referències corresponen a l'obra de Carles Duarte i Àlex Alsina (1984).

[134] T X C M> t ks k (elisió de la nasal, s. I, regla 40)> 't ks k (accentuació del llatí vulgar, regla I)> 't kseko (abaixament de les vocals laxes, regla III, i pèrdua posterior de la tensió, regla IV)> 't seko (espirantització de c [-coronal], s. I, regla 5)> 't jseko (vocalització d'espirant+obstruent, s. IV-V, regla 7)> 'tw jseko (diftongació de , s. IV-V, regla VII)> 'twujseko (ascensió de , s. IV-V, regla VIII)> 'twujsek (reducció vocàlica,

documentat en el nostre corpus en cinc ocasions, i és present en tres obres: la *Doctrina pueril* de Ramon Llull (252, 6; segona meitat del segle XIII), *Tractat de confessió* d'Antoni Canals (22, 7 i 210, 13; primera meitat del segle XV) i al *Memorial del pecador remut*, de Felip de Malla (163, 6; segona meitat del s. XV). Un altre d'aquests mots patrimonials caiguts en desús en favor de semicultismes és el de «fabregar»,[135] que ni tan sols apareix al DCVB i que és present al corpus sota la forma «fabregan» (a les *Vides de sants rosselloneses* 164, 20), «fabreguen» (*EVM-I* 12, 48) o «fabregada», «fabregarà», «fabregaran» i «fabregar» (als *Furs de València*, respectivament, 0, 1733; 0, 7188; 0, 6521 i 0, 6521). També és una forma sorgida per la via popular «feel» (enfront a *fidel*, la forma culta presa del llatí); aquesta és una forma bastant transparent i, a més, amb entrada al DCVB; no passa el mateix amb la variant «fezel»[136], sorgida fruit de l'estridentització de la [ð], procedent alhora d'una [d] que ha patit una espirantització. Un darrer exemple d'aquestes evolucions patrimonials és el de «corobsió»,[137] amb l'ascens a /o/ de la *u* breu llatina que s'esdevingué en el llatí tardà (en paral·lel, doncs, a «corrompre»), una forma desplaçada pel semicultisme «corrupció». Al capdavall, es tracta de casos interessants per a l'estudi de l'evolució fonètica del català (i de les llengües romàniques en general) i, per això mateix, és necessari tenir nocions de gramàtica històrica per a lematitzar

---

regla XXXIV, s. IV-V)> ʻtwuj ek (palatalització de *js*, s. VI-VII, regla 20)> ʻtuj ek (absorció de semiconsonant s. IX, regla IX)> ʻtuj ek (apòcope, s. VIII-IX, regla XXXV)> ʻtu ek (absorció de la semiconsonant). Les referències corresponen a l'obra de Carles Duarte i Àlex Alsina (1984).

[135] La forma que ji ha acabat imposant-se és «fabricar», reintroduïda a partir del llatí «fabr care» sense lenició de l'oclusiva ni obertura de la vocal breu.

[136] Les formes «fasel», «fesel», «fezel», «fesels», «fezels» les constatem, dins el nostre corpus, en els *Diàlegs de sant Gregori* i en les *Vides de Sants Rosselloneses*, dues obres que lingüísticament corresponen al català septentrional, enfront de les formes «feel», «fel» i variants gràfiques i flexives que són generals en la resta d'obres; aquestes formes, les darreres, són el resultat de l'elisió de la [ ].

[137] L'exemple procedeix dels *Començaments de medicina* de Ramon Llull (47, 32); es testimonia en huit ocasions amb aquesta forma i també com a «corobció».

correctament la forma i permetre així que l'investigador interessat puga recuperar-la amb facilitat.[138]

La derivació genera sovint formes difícils de reconèixer i, més encara, de lematitzar. Per exemple, l'adjectiu «plantivoses»[139], que potser hauríem d'interpretar a partir del substantiu *plantiu>plantivós* ('lloc poblat de plantes'), en paral·lel a altres adjectius formats a partir d'aquest sufix com *caliu>calivós* o *saliva>salivós*. No existeix, però, cap adjectiu «plantivós» en el DCVB ni en el DIEC, com tampoc en el DECat. Un altre exemple és el de «abatayons»[140] (*Un llibre reial mallorquí del segle XIV*: 234, 20); hi trobem el sufix *-ó* afegit al verb «abatollar» (colpejar amb batolles; hem d'entendre que un «abatolló» és un 'colp de batolla'. Un tercer exemple és «refertegava»[141] (*Sumari d'Espanya*, B. De Puigpardines: 109, 3), una forma que potser derive del participi passat de «referir» substantivat: «refert» o «referta» (allò que s'ha contat, la cosa referida) amb l'afegitó del sufix *-ejar*. Un darrer cas és el d'«axalenar», que J. Rabella i Ribas (1998, 330) ha considerat semànticament vinculat a «eixelebrat», però amb «un creuament amb alguna forma verbal derivada de *exhalar*, o d'algun verb semblant gràficament». En realitat sí que hi ha un vincle amb *eixelebrat*, però no semàntic, sinó formal: el prefix *ex-*: «axalenat» és «e(i)xalenat»: mancat d'alé, fet que casa amb el context:

[138] El lema escollit per als arcaismes que no apareixen en les fonts de referència és la forma actual: per a «fabregar», «fabricar».

[139] Aquesta forma apareix en la *Flor de les Històries d'Orient* (108, 17): «e soferrí gran fretura e desayra, ell e sa gent, entró tant que foren venguts en una bona terra on trobaren terres *plantivoses* e abundants de tots béns. En aquella terra stigueren molts dies a gran repòs.»

[140] «eyl sí estania ordi per batre en la sua era ab la sua companya, entre els quals hi era en Mateu Aymaric, fiyl d'en Arnau Aymarich, e stanent lo blat en la era sí y vénch en Berenguer Parató, dient al dit Mateu: "Digues, tu, per què m'as feta aquesta enpara que m'as feta?", e ·yl dit Mateu li respòs e dix: "Que no le t'e feta, mas fer-la t'é!", e aquel dix: "Yo no ·t són tangut, que tu és tangut a mi!" "Ans", dix lo dit Mateu: "és tu tangut a mi!", e salavòs lo dit Berenguer sí s'acostà al dit Mateu, dient-li: "Fiyl del barba merdose!", o axí li ·n dóna de parer, "que yo ·t pagaré!", e salavòs lo dit Berenguer Parató mantinent donà *abatayons* ab dos darts que tania per l'escana al dit Mateu irosement».

[141] «E aquí los reys li digueren lur contesa, en especial lo rey En Pere d'Aragó, qui més *refertegava* que sua devia ésser».

«Con aq*ue*st *testis* se·n anàs al Born ves casa sua, encon[32*v*]trà lo dit Arnau Marq*uè*s, lo qual venia fort axalenat» (RABELLA: 1998, 59-60).

Finalment hi ha els casos que no podem resoldre de cap manera perquè superen els nostres coneixements. Alguns d'aquests no tenen, però, solució possible. Per exemple, el cas d'«adoure» en el *Sumari d'Espanya* de Berenguer de Puigpardines (122, 19)[142], que només pot tenir sentit si suposem un error de transcripció: «adoure» és, entenem, «acloure» (forma documentada en el DCVB, sinònim de «cloure», que lliga perfectament amb el context).

No té cap trellat eternitzar l'enumeració de casos que tenen una certa complicació a l'hora de identificar-los i lematitzar-los. La idea que pretenïem transmetre és, senzillament, que el treball que es fa amb la lematització —i amb l'etiquetatge en general— no és pur automatisme, sinó que és també filologia i exigeix uns certs coneixements de gramàtica històrica.

6. A TALL DE CONCLUSIÓ

Només pretenïem, amb aquestes pàgines, explicar quin és el sentit i quines les dificultats que té la lematització en un corpus informatitzat com el desenvolupat al si de l'IVITRA. De fet, podríem haver-nos estalviat temps i resumir-ho tot plegat: com sembres, colliràs. Al capdavall, la tasca de lematitzar —i també la de categoritzar— és això: preparar les unitats lèxiques perquè l'investigador puga recuperar un mostrari tan ajustat com siga possible als seus interessos. Procurem, al capdavall, que la palla que es barrege amb el gra no siga tanta perquè no pague la pena revisar-la; però que tampoc siga tan poca perquè perdem mostres d'interès i la panoràmica quede esbiaixada. I aquesta preparació del material lingüístic que fem ha d'ésser minuciosa i, sense perdre de vista que la tasca és gegantina i les nostres capacitats limitades, tan exacta com siga possible, perquè de la mateixa manera que el nostre treball

[142] «E com les noves vingueren al rey En Pere, qui ja era en Catalunya, féu-ne moltes gràcies a Déu, e féu-ne fer grans alegries en Barcelona. E de aquí envià misatgers al rey Carles, qui era en França, dient-li que ell era vengut en Catalunya per donar fi e conclusió a la batalla, e axí, que ves en quina manera se havia *adoure* ne la forma com havia venir a fi.»

facilitarà la investigació en filologia, els nostres errors poden falsejar la base empírica de què parteixen els estudis.

BIBLIOGRAFIA

DUARTE, Carles i Àlex Alsina (1984): *Gramàtica històrica del català*, v. 1, Barcelona, Curial.

COROMINES, Joan (1983-1991): *Diccionari Etimològic Complementari de la Llengua Catalana* (DECat), Barcelona, Curial.

MARTINES, Josep (2000): *El valencià del segle XIX. Estudi lingüístic del «Diccionario Valenciano» de Josep Pla i Costa*, Alacant-Barcelona, Institut Interuniversitari de Filologia Valenciana - Publicacions de l'Abadia de Montserrat.

MORAN I OCERINJAUREGUI, Josep (2000): ressenya d'«*Un matrimoni desavingut i un gat metzinat». Procés criminal barceloní del segle XIV*, *Llengua i literatura*, 11, p. 549-551.

RABELLA I RIBAS, Joan Anton (1998): «*Un matrimoni desavingut i un gat metzinat». Procés criminal barceloní del segle XIV*, Barcelona, Publicacions de l'Abadia de Montserrat - Institut d'Estudis Catalans.

SÁNCHEZ, Elena (2010): «Lingüística de *corpus* i els clàssics literaris: el repte d'etiquetar la llengua antiga», en Nancy DE BENEDETTO i INES RAVASINI, *Da Papa Borgia a* Borgia Papa. *Letteratura, lingua e traduzione a Valencia*, Lecce, Pensa MultiMedia Editore, p. 265-282.