JOAQUIM RAFEL

UN CORPUS GENERAL DE REFERÈNCIA DE LA LLENGUA CATALANA

INTRODUCCIÓ

En altres llocs i ocasions hem donat informació sobre l'origen i el desenvolupament del projecte de l'Institut d'Estudis Catalans anomenat Diccionari del Català Contemporani (DCC), dins el qual es realitza el Corpus Textual Informatitzat de la Llengua Catalana (CTILC).[1] No repetiré ací qüestions relacionades amb la justificació, l'inici, l'execució i el desenvolupament d'aquest projecte, que hom pot trobar en algunes d'aquestes publicacions, sinó que en aquest treball em referiré exclusivament a alguns aspectes de la caracterització del CTILC, i, sobretot, donaré compte del seu estat actual i de les possibilitats d'explotació, que, malgrat no haver-lo pogut completar encara, són ja una realitat.

Només recordaré breument en aquests paràgrafs introductoris, per a destacar-ne la motivació bàsicament lexicogràfica, que el projecte DCC té com a objectiu principal la futura elaboració d'un diccionari descriptiu del català d'acord amb els principis metodològics i tecnològics que la lexicografia ha assumit en els darrers anys com a conseqüència dels avenços de la ciència lingüística i de les possibilitats que la tècnica moderna posa a la nostra disposició. Un dels principis metodològics fonamentals d'aquest projecte és que l'elaboració d'un diccionari d'aquestes característiques ha

(1) «Cap a un diccionari del català contemporani», *Segon Congrés Internacional de la Llengua Catalana, IV, Àrea 3: Lingüística Social* (celebrat a Palma del 30 d'abril al 4 de maig de 1986), Palma, 1992, pàgs. 589-595. «El "Corpus textual automatitzat de la llengua catalana"» (en col·laboració amb J. M. Solanellas). *Actas de las II Jornadas Españolas de Documentación Automatizada. 20-22 de Noviembre de 1986. Ponencias y comunicaciones* (Torremolinos-Málaga, 1986), pàgs. 147-161; «El "Diccionari del català contemporani"». *Serra d'Or*, 1987, pàgs. 428-431; «El "Corpus textual informatitzat de la llengua catalana" i el "Diccionari del català contemporani". Un proyecto del Institut d'Estudis Catalans», *Anthropos*, 1988, Documentación cultural e información bibliográfica, pàgs. V-VII; «El "Diccionari

d'utilitzar com a font principal un corpus textual suficientment representatiu de la llengua que vol reflectir, en comptes de fonamentar-se, com s'ha fet tradicionalment en lexicografia, en els diccionaris preexistents i en un mètode de treball basat en la intuïció o en la mera competència dels lexicògrafs. Entre altres molts avantatges que ara no comentaré, això permet d'establir el significat dels mots a partir de l'ús que se n'ha fet en la llengua, en comptes d'utilitzar mètodes apriorístics o intuïtius; permet també d'exemplificar amb frases preses dels textos reals els diferents valors i usos de cada mot i les diferents estructures sintàctiques de què pot formar part, i permet, encara, la utilització d'informació de caràcter estadístic, que pot arribar a tenir una gran importància en l'estudi científic d'una llengua.

Cal dir, però, que, per bé que la idea d'aquest corpus sorgí d'una necessitat concreta i amb una finalitat específica, des del primer moment de la seva elaboració s'ha tingut en compte d'estructurar-lo de tal manera que no en resultin restringides les possibilitats d'utilització; s'han pres les precaucions metodològiques i tècniques necessàries perquè el corpus pugui ésser emprat amb profit per a qualsevol tipus d'investigació lingüística que requereixi la utilització de dades procedents de la llengua escrita. L'esforç humà i els recursos necessaris per a portar a terme un projecte d'aquesta naturalesa són molt importants; això ens féu pensar des del primer moment que el resultat d'aquests esforços havia de poder ésser utilitzat en múltiples aplicacions d'una manera indefinida.

CARACTERÍSTIQUES DEL CORPUS TEXTUAL INFORMATITZAT DE LA LLENGUA CATALANA (CTILC)

Característiques generals

La caracterització general de qualsevol corpus té en compte factors de naturalesa temporal, de naturalesa qualitativa i de naturalesa quantitativa; és a dir, el període de temps que abraça, el tipus de text que conté i l'extensió total de text que l'integra.

Des del punt de vista temporal, el CTILC abraça des de 1833,[2] com a data simbòlica de la represa de l'ús literari de la llengua catalana en l'època moderna, fins al moment actual.[3] Per una altra banda, els textos tinguts en compte per a formar part del corpus són textos publicats, sigui en forma de llibre o d'opuscle, full solt, diari, revista, etc.[4] Els textos seleccionats han estat introduïts d'una manera íntegra, amb les poques excepcions d'algunes obres que, per la seva extensió, no era possible introduir totalment; és important també tenir en compte que el text font, que s'utilitza com a referència definitiva, és sempre la primera edició de l'obra, i que el seu contingut és introduït al corpus sense fer-hi cap correcció ni modificació; això significa que el corpus incorpora tota la diversitat de formes i grafies que trobem, sobretot, en textos no normalitzats.

del català contemporani": treballs realitzats i previsions de futur». *Llengua & Literatura*, Revista anual de la Societat Catalana de Llengua i Literatura, 5, 1992-93 [1994], pàgs. 733-737; «Le "Diccionari del català contemporani" et le "Corpus textual informatitzat de la llengua catalana". Brève description du projet et état des travaux», *Actes du XX Congrès International de Linguistique et Philologie Romanes, Université de Zurich (6-11 avril 1992)*, IV, Tübingen-Basel, 1993, pàgs. 816-821. D'altra banda, en les diferents memòries d'activitats de l'Institut hom pot trobar informació sobre l'estudi previ, l'aprovació, l'inici i el desenvolupament progressiu d'aquest projecte: *Memòria d'activitats. (octubre 1982-desembre 1983)*, Barcelona, 1984, pàgs. 131 i 135-137; *Memòria d'activitats 1984*, Barcelona, 1986, pàgs. 136 i 137; *Memòria d'activitats. Curs 1988-89*, Barcelona, 1990, pàgs. 25-28; *Memòria d'activitats. Curs 1989-90*, Barcelona, 1991, pàgs. 137-139; *Memòria d'activitats. Curs 1990-91*, Barcelona, 1992, pàgs. 147-159 i *Memòria d'activitats. Curs 1991-92*, Barcelona, 1993, pàgs. 165-177. Un altre títol que forma part de la història del projecte és la publicació interna *Investigación lingüística 1989-92*, Generalitat de Catalunya, Departament de Presidència. Comissió Interdepartamental de Recerca i Innovació Tecnològica (CIRIT). «Programes de recerca i desenvolupament de la Generalitat de Catalunya», 3, (s. d.) [Anexo nº 3. Diccionario del catalán contemporáneo. Corpus textual informatizado de la lengua catalana. Desarrollo general y estado actual del proyecto (43 pàgs.). Anexo nº 4. Diccionari del català contemporani. Informe sobre la fase de lematización. Descripción y métodos (34 pàgs)].

(2) Hi ha alguna obra anterior a aquesta data, però a efectes del còmput de grups cronològics prenem l'any 1833 com a data inicial.

Quant a la naturalesa de la llengua o a la temàtica dels textos, he de dir que el CTILC inclou textos de caràcter literari, que és el més habitual en corpus que existeixen per a d'altres llengües, però també textos de caràcter no literari, és a dir, textos que tracten de qualsevol dels temes possibles i des de qualsevol perspectiva, sense una finalitat artística o una intenció estètica; la proporció entre els textos literaris i els no literaris és del 40% per als primers i el 60% per als segons; aquests percentatges han estat establerts després d'estudiar a fons el conjunt de la producció escrita en català i tenint com a principal objectiu el de mantenir una representativitat adequada, proporcional, en la mesura possible, al volum de textos publicats i a la varietat tipològica.

Pel que fa a l'extensió del corpus, per bé que no està del tot acabat, com ja he dit, i com precisaré encara més endavant, podem donar una estimació molt acostada a la realitat, pel fet que hi ha introduïts a l'ordinador (i, per tant, comptabilitzats) pràcticament el 90% dels textos. El total de l'extensió prevista és d'entre cinquanta-dos i cinquanta-tres milions de mots. La repartició d'aquesta xifra, segons la tipologia dels textos i les seccions cronològiques que hem establert, és la següent:[5]

LLENGUA NO LITERÀRIA

	Secció I (1833-1973)	Secció II (1874-1913)	Secció III (1914-1988)	TOTAL
0. Correspondència	2.597	40.252	77.750	**120.599**
1. Filosofia	6.886	210.337	1.532.615	**1.749.838**
2. Religió i teologia	235.218	533.478	2.203.350	**2.972.046**
3. Ciències socials	202.995	909.723	4.494.161	**5.606.879**
4. Premsa	109.686	462.242	2.998.680	**3.570.608**
5. Ciències pures i naturals	23.715	372.237	1.828.993	**2.224.945**
6. Ciències aplicades	225.303	573.767	3.680.495	**4.479.565**
7. Belles arts. Divertiment. Jocs. Esports	5.027	416.406	2.361.817	**2.783.250**
8. Llengua i literatura	136.310	333.840	1.759.025	**2.229.175**
9. Història i geografia	107.855	426.895	2.962.889	**3.497.639**
Total	**1.055.592**	**4.279.177**	**23.899.775**	**29.234.544**

LLENGUA LITERÀRIA

	Secció I (1833-1973)	Secció II (1874-1913)	Secció III (1914-1988)	TOTAL
1. Assaig	0	381.164	2.670.096	**3.051.260**
2. Narrativa	536.748	2.499.426	10.885.237	**13.921.411**
3. Poesia	245.472	604.528	1.836.732	**2.686.732**
4. Teatre	378.981	836.246	2.471.084	**3.686.311**
Total	**1.161.201**	**4.321.364**	**17.863.149**	**23.345.714**

(3) En realitat les darreres obres seleccionades corresponen a 1988, any en què es tancà le selecció quan es realitzà aquesta tasca, que prenia com a marc períodes de cinc anys. Això no vol dir que, una vegada completat, el corpus no pugui ésser actualitzat progressivament a partir d'un pla adequat i dels recursos necessaris per a portar endavant aquesta tasca.

(4) Només s'han introduït alguns textos no publicats, com cartes o escriptures notarials; en aquests casos, els textos font són manuscrits o mecanoscrits. Aquests han estat tinguts en compte per a completar el panorama de la llengua escrita amb uns textos que, si bé no han estat publicats, tenen la característica d'haver estat donats per definitius; la carta, perquè en el moment d'enviar-la és un text que s'ha donat per acabat, i les escriptures notarials, perquè la seva pròpia naturalesa els dóna la característica de textos perfets.

(5) Aquesta repartició s'ha d'entendre com a definitiva pel que fa a la llengua no literària, perquè aquesta part del corpus està definitivament acabada. En canvi, les dades corresponents a la llengua literària representen l'estimació que es pot fer en aquest moment, perquè encara no s'han acabat els treballs de constitució d'aquesta altra part del corpus; tanmateix, aquesta estimació és molt aproximada a la realitat.

Els textos seleccionats

A l'hora de seleccionar els textos per a formar part del corpus, l'objectiu principal ha estat d'assegurar el major grau de representativitat possible de tots els textos publicats dins els límits temporals establerts, i al mateix temps un equilibri raonable entre les diferents classes. Per aconseguir aquest objectiu s'ha dividit el total de l'extensió temporal del corpus (1833-1988) en vint-i-tres grups cronològics (vuit grups de deu anys cada un entre 1833 i 1913, i quinze grups de cinc anys cada un entre 1914 i 1988); a l'hora de fer la selecció, s'ha procurat que, no solament d'una manera general, sinó també dins cada un d'aquests grups, hi hagi una representació adequada de cada tipus de text, entenent per tipus cada un dels gèneres (narrativa, teatre, poesia i assaig) pel que fa a la llengua literària, i cada una de les deu àrees temàtiques establertes, subdividides, la major part, en diverses subàrees, també fins a deu, pel que fa a la llengua no literària (aquestes àrees temàtiques, que, per qüestions pràctiques, segueixen de prop els criteris de classificació decimal de les biblioteques, són: 0. Correspondència, 1. Filosofia, 2. Religió i teologia, 3. Ciències socials, 4. Premsa, 5. Ciències pures i naturals, 6. Ciències aplicades, 7. Belles arts, divertiments, jocs i esports, 8. Llengua i literatura, 9. Història i geografia). Entenem que la representació és adequada des del punt de vista cronològic i temàtic quan podem garantir fins a un cert punt una relació de proporcionalitat entre la quantitat de text seleccionat per a cada tipus i per a cada grup cronològic i el volum de text de cada tipus publicat dins cada un d'aquests grups. A aquests criteris generals se n'hi afegeixen uns altres de complementaris, com són la consideració de les diferències dialectals, els subgèneres, o bé la temàtica o l'estil diferents dins els textos corresponents a un mateix gènere literari, per exemple. La minuciositat d'aquesta operació ha tingut com a objectiu d'aconseguir el màxim grau de representativitat i d'equilibri des de diversos punts de vista, a fi que el CTILC pugui tenir la funció que hem apuntat de corpus general de referència de la llengua catalana; sembla obvi que les conclusions a què podem arribar partint d'un corpus equilibrat i representatiu tenen un valor més alt que aquelles que podríem deduir d'un corpus que no presentés aquestes característiques.

El total d'obres seleccionades és de 3.302, que es reparteixen en 2.296 corresponents a la llengua no literària i 1.006 a la llengua literària.

Les dades del corpus

En una part dels corpus, més o menys extensos, que existeixen per a diferents llengües, la unitat de treball a través de la qual es pot accedir a la informació que contenen és el mot entès com a unitat gràfica, tal com apareix en el text font; això té l'inconvenient de l'alt grau d'ambigüitat que poden arribar a presentar aquestes unitats, que

varia segons les llengües (en català, per exemple, la forma *fins* pot correspondre a una preposició, a un substantiu masculí, a un substantiu femení o a un adjectiu; *creuen* pot ésser una forma del verb *creuar* o una forma del verb *creure*; *causa* pot ésser un substantiu o una forma del verb *causar*, o bé pot fer part d'una locució prepositiva; *cap* pot ésser un substantiu, una preposició, un adjectiu o bé una forma del verb *cabre*; *deu* pot correspondre a un adjectiu numeral, a un substantiu femení, a un substantiu masculí, o, encara, a dues formes diferents del verb *dar* o a dues formes diferents del verb *deure*, etc.). Aquest fet pot esdevenir un inconvenient a l'hora de consultar les dades del corpus, perquè l'usuari té accés a una informació massa diversa a partir d'una mateixa unitat d'interrogació.

En altres corpus textuals, en canvi, s'ha dut a terme una primera anàlisi de caràcter lingüístic de les dades, que té com a un dels objectius de desfer el tipus d'ambigüitat a què m'acabo de referir (ambigüitat gramatical), i, a la vegada, relaciona entre elles les diferents formes pertanyents a una mateixa sèrie flexional; es tracta de l'operació anomenada *lematització*. D'acord amb això, doncs, des del punt de vista del grau d'interpretació de les dades que contenen, els corpus es poden classificar en no lematitzats i lematitzats. El CTILC és un corpus del segon tipus; el seu procés d'elaboració inclou una primera fase d'introducció i validació de les dades textuals i una segona fase en la qual es realitza la lematització de totes les ocurrències del corpus.[6]

A través de la lematització, s'aconsegueixen, doncs, dos objectius particulars, que són, de fet, un conseqüència de l'altre: cada una de les ocurrències de cada mot gràfic (és a dir, cada una de les seves aparicions al llarg del text) és categoritzada gramaticalment i associada a una forma de referència anomenada *lema* (que es correspon aproximadament amb allò que podem considerar una entrada de diccionari); per exemple, davant cada una de les ocurrències de la forma gràfica *nous*, a la vista de l'entorn contextual, es determina a quina de les possibilitats de categorització gramatical correspon de les quatre que pot tenir aquesta seqüència de caràcters: NOU *adj.*, NOU *subst. masc.*, NOU *subst. fem.*, i NOURE *verb* (en el primer cas és la forma de masculí plural; en el segon i en el tercer, la forma de plural; i en el quart, la forma de segona persona del singular del present d'indicatiu); o bé, davant les diferents ocurrències de la forma gràfica *cap* en el text, es determina si corresponen a CAP *subst. masc.*, a CABRE *verb*, a CAP *prep.* o a CAP *adj.* (en el primer cas és la forma de singular, en el segon la de tercera persona del singular del present d'indicatiu, i en el tercer i en el quart, no té categorització morfològica, perquè correspon al que correntment anomenem mots invariables. Així, per una banda es desambigüen gramaticalment les formes homògrafes, i per una altra s'agrupen sota un mateix lema els diferents components d'una mateixa sèrie inflectiva; és a dir, algunes de les ocurrències de *cap* s'agrupen amb *caps* sota el lema CAP *subst. masc.*, i certes altres ocurrències de la mateixa grafia es relacionen amb *caben, cabem, cabran, cabrien, càpiga,* etc., sota el lema CABRE *verb*; de la mateixa manera, algunes de les ocurrències de la grafia *nous* s'agrupen

(6) La lematització es realitza en el CTILC a través d'un procediment semiautomatitzat; l'ordinador elabora una proposta de lematització, que després ha d'ésser examinada i simplement donada per bona o bé completada manualment. El sistema (programes i fitxers de lemes i formes necessaris), que no puc comentar aquí amb detall, ha estat elaborat completament dins els treballs del CTILC.

amb *noves, nova, nou,* sota el lema NOU, *adj.*, altres s'agrupen amb *nou* sota el lema NOU *subst. masc.*, o bé sota el lema NOU *subst. fem.*, segons els seus contextos, i altres, juntament amb *noïa, nourà, nogui,* etc., sota el lema NOURE *verb.* En un corpus sense lematitzar un usuari podria demanar només informació a través de la grafia *cap*, o de la grafia *nous*, i la informació que obtindria sobre les diferents ocurrències d'aquestes grafies mitjançant aquesta forma d'interrogació seria indiscriminada des del punt de vista lingüístic; és a dir, rebria, sense diferenciar, les formes corresponents al substantiu masculí, les corresponents al verb, a la preposició i a l'adjectiu, en el cas de *cap*, o a l'adjectiu, al substantiu masculí, al substantiu femení i al verb, en el cas de *nous;* en canvi, en un corpus lematitzat d'acord amb aquests principis, l'usuari té accés a les dades agrupades adequadament a partir del substantiu masculí CAP, del verb CAURE, de la preposició CAP, o de l'adjectiu CAP, o bé de l'adjectiu NOU, del substantiu masculí NOU, del substantiu femení NOU o del verb NOURE, perquè la base de dades on es troben emmagatzemades les diferents ocurrències del corpus conté aquesta informació.[7]

Si reprenem l'exemple de *cap*, podem dir que, en principi, el lema CAP *subst. masc.* tindrà relacionades dues formes, la de singular *(cap)* i la de plural *(caps)*, i que les diferents ocurrències d'aquest lema que hi ha en el corpus estaran associades a l'una o a l'altra d'aquestes dues formes. Afegiré, però, alguns comentaris que ajudaran el lector a fer-se una idea de la manera com s'ha tractat l'extraordinària varietat formal que apareix en el corpus i alguna qüestió de caràcter morfològic. En primer lloc cal tenir en compte que el mateix procés de lematització, a més de relacionar les diferents formes d'una mateixa sèrie flexional, agrupa també les diferents variants gràfiques que poden aparèixer en els textos (així, en el cas del lema CAP *subst. masc.*, trobem formes com *cab, cáp, câp, cabs, cáps,* a part de les normatives; o, en el cas del lema FILL, *subst. masc.*, trobem *figl, fii, fill, fiy, fiís, fiys,* a més de *fill* i *fills*).[8] Per una altra banda, a l'hora de definir els criteris per a la lematització, decidírem classificar com a formes d'un mateix lema –el corresponent al mot primitiu– tots els derivats apreciatius (diminutius, augmentatius, pejoratius, intensius),[9] per als quals, doncs, no es creen nous lemes; així, en l'estat actual del corpus, apareixen com a formes del lema CAP *subst. masc.* els mots *cabet, caparró, caparronet, caparrot, caperrot, capet, cabets, caparrins, caparronets, caparrons, caparrots, caperrots, capets, cabot*; i, en el cas de GRAN *adj.*, trobem *regran, regrán, regrans, grandaçot, grandàs, grandet, grandissim, grandíssim, grandot, grandeta, grandísima, grandissima, grandíssima, grandota, grandassassos, grandassots, grandets, grandissims, grandíssims, grandolassos, grandots, grandòts, grandetes, grandísimas, grandíssimas, grandíssimes*; totes aquestes formes van codificades amb una marca que permet tractar-les conjuntament, a part de la resta, si convé, per a un tipus d'estudi específic. Cal també tenir en compte que, quan un mot apareix usat metalingüísticament en un text, es codificat com a tal, i, en conseqüència, es classifica com una forma diferenciada de la resta, encara que tingui la mateixa grafia i el mateix codi morfològic que alguna altra. Totes aquestes qüestions es poden veure en els exemples de les figures 3 i 4.

(7) No puc estendre'm en aquest treball en tota una sèrie d'aspectes relacionats amb la lematització. Cal dir, però, que hi ha hagut una certa confrontació de parers sobre si és preferible lematitzar els corpus o no lematitzar-los: l'actitud de no lematitzar es basa en l'opinió que qualsevol tipus d'intervenció sobre el text verge pot condicionar les conclusions ulteriors. Per a aquesta qüestió ara em limitaré a remetre el lector a dos treballs que resumeixen aquestes posicions contraposades: MAURICE TOURNIER; «Sur quoi pouvons-nous compter? Réponse à Ch. Muller», in *De la plume d'oie à l'ordinateur. Mélanges en l'honneur de Gérald Antoine*, [Responsables de la publication: GILBERT BOISSIER: DANIELLE BOUVEROT], Nancy, 1984, pàgs. 131-136; i CHARLES MULLER; «Lemmatisation et information», *Hommage à Pierre Guiraud* [Comité de réd.: CHARLES P. BOUTON: ÉTIENNE BRUNET, LOUIS-JEAN CALVET] *Annales de la Faculté des Lettres et Sciences Humaines de Nice*. 52, 1985, pàgs. 285-291. Al marge, però, de debats ideològics, els enormes avantatges pràctics d'un corpus lematitzat són actualment reconeguts d'una manera generalitzada: «Although there is nothing new in the idea of assigning individual words to a particular word-class, it is no trivial task to do this with every word in a body of text. But there are considerable benefits to having a corpus in which every word is accompanied by a label which indicates what type of word it is.» (GEOFFREY LEECH, STEVEN FLIGELSTONE; «Computers and Corpus Analysis», in CHRISTOPHER S. BUTLER (ed.); *Computers and Written Texts*, 1992, pàgs. 124-125).

D'altra banda, admetent la lematització com una manera adequada de categoritzar les unitats d'un corpus a fi de facilitar-ne una explotació més efectiva, es presenten una sèrie de problemes metodològics de no fàcil solució: cal prendre una sèrie de decisions, entre d'altres, sobre la manera de

categoritzar casos que permeten més d'una interpretació (conjunció o *adverbi*, substantiu o adjectiu, verb o substantiu –en el cas de certes ocurrències de l'infinitiu–, verb o adjectiu –en el cas de certs participis–, etc.); cal tenir en compte que els textos utilitzen de vegades la llengua fins als seus límits, i que aquest fet planteja una sèrie de problemes als quals cal trobar una solució, sovint qüestionable. Els autors que acabo de citar es refereixen d'una manera tan breu com contundent a aquest problema: «One problem is that what class a word belongs to turns out in many cases to be very difficult to decide.» (*Ibid.*, pàg. 125) A part d'això, cal decidir també el grau d'intervenció que hom vol practicar al llarg d'aquesta operació (per exemple, classificant per separat els mots que formen part de locucions o perífrasis, o bé agrupant-los en un sol lema complex), que pot ésser més suau o més severa. Per a totes aquestes qüestions és convenient d'establir criteris tan clars com sigui possible, perquè, al llarg de la seva aplicació, la lematització del corpus es faci d'una manera coherent i amb el màxim grau d'uniformitat: en el cas del CTILC, la resolució d'aquesta problemàtica ha donat lloc a un voluminós *Manual de lematització*.

(8) Hi ha casos bastant complexos. El verb veure, per exemple, en l'estat actual del corpus, té 639 formes diferents; de les quals 162 corresponen a usos metalingüístics.

(9) Aquestes formacions amb afixos no lexicals no solen ésser tingudes en compte en els diccionaris. Evidentment, no hi poden donar lloc a una entrada nova, però des del punt de vista de la descripció de l'ús lingüístic, és molt important de conèixer quines de les possibilitats virtuals es materialitzen realment, en quines circumstàncies o situacions ho fan, quina freqüència tenen en la realitat de la llengua, o amb quines bases lexicals concretes es combinen preferentment aquest tipus d'afixos.

Hi ha, però, casos en què la variació va més enllà de la purament gràfica (i no té res a veure amb els derivats apreciatius ni amb els usos metalingüístics a què m'acabo de referir); es tracta dels casos en què aquesta no és merament gràfica, sinó que comporta una diferència en l'estructura fonològica o morfològica del mot, que pot semblar suficient per a determinar lemes diferents, però amb una similitud prou forta com per a mantenir algun tipus de relació lògica que permeti tractar-los, si convé, d'una manera agrupada (ens referim a casos com *almorzar* i *armosar*, al costat d'*esmorzar*; o bé *abercoc, abricoc, albaricoc, bercoc* o *obrecoc*, al costat d'*albercoc*; *juriol*, al costat de *juliol; radere*, al costat de *darrere; rellonge, relonge* o *relotge*, al costat de *rellotge; esgolfa* o *gorfa*, al costat de *golfa*, etc.). Per a resoldre aquests casos prenguérem la decisió de crear dues categories de lemes, que resten relacionats jeràrquicament dins el sistema informàtic: els lemes principals i els lemes secundaris; això permet de tractar-los independentment, cada un amb les seves formes associades, o bé acumular les dades dels secundaris als principals, quan la naturalesa de la informació que cerquem ho requereix (a les figures 1 i 2 podem veure uns exemples d'aquest tipus d'estructura).

LEMA	RANG	FORMA
ESMORZAR M	principal	esmorçar S
		esmorsà S
		esmorsar S
		esmorzar S
		esmorçars P
		esmorsars P
		esmorzars P
		esmorzar MET
ALMORZAR M	secundari	esmorsar S
		almorzar S
		esmorzar MET
ARMOSAR M	secundari	armosar S
		armosarot DS

FIGURA 1. En aquesta taula podem observar la relació jeràrquica establerta entre el lema ESMORZAR, com a lema principal, i els lemes ALMORZAR i ARMOSAR, etiquetats com a secundaris; cada un està relacionat amb les formes corresponents. Això permet, amb la contribució dels programes adequats, tractar-los de manera separada, com a lemes independents, o bé de manera conjunta, com si es tractés d'un sol lema, segons l'objectiu de la recerca projectada, o, simplement, de la naturalesa de la informació que es vol obtenir.

LEMA	RANG	FORMA
DONCS	principal	donç
		donchs
		dónchs
		dònchs
		doncs
		dóncs
		dòncs
		dons
		donchs MET
		doncs MET
		dòngks MET
		dòngs MET
		dons MET
		dòns MET
DONC	secundari	donc
		donch
DONCES	secundari	donças
		donsas
DONQUES	secundari	doncas
		donques
DÒS	secundari	dos
		dòs
		dôs
		dòs MET

FIGURA 2. Veiem en aquesta figura un altre exemple de lemes relacionats amb el criteri de principal i secundaris: principal DONCS; secundaris, DONC, DONCES, DONQUES, DÒS. Cada un agrupa les seves formes. Les formes corresponents al lema DONCS podrien haver estat agrupades d'una altra manera; a l'hora de prendre les decisions, però, es considerà que la diferència entre formes com *donç* o *dons*, d'una banda, i *doncs* o *donchs*, d'una altra banda, no era suficient per a donar d'alta lemes diferents (DONCS i DONS).

LA BASE DE DADES TEXTUAL DE LA LLENGUA CATALANA (BDTLC)

Una vegada acabat el procés de lematització de cada una de les obres introduïdes prèviament en el sistema informàtic, i després d'haver estat donats per vàlids, els resultats d'aquesta operació passen a formar part d'una única base de dades, que conté tota la informació necessària per a una adequada explotació del corpus: és la Base de Dades Textual de la Llengua Catalana.

Aquesta base de dades conté, per a cada una de les ocurrències del corpus, les informacions següents: la forma gràfica, la localització de l'ocurrència (codi de l'obra, pàgina, línia i número d'ordre del mot dins la línia), el codi morfològic i el número de referència del lema que li ha estat atribuït. D'altra banda, la base de dades també conté una sèrie d'elements que permeten reconstruir el context corresponent a cada ocurrència; aquestes dades són, bàsicament, tots aquells codis que corresponen al que podríem considerar *grosso modo* signes de puntuació (punts, comes, punts i comes, punts suspensius, signes d'interrogació i d'exclamació, etc.), i també les parts del text originari que no han estat tingudes en compte durant el procés de lematització, ja que en la fase d'introducció havien estat codificades com a no analitzables (nombres expressats en xifres, citacions d'altres autors o en d'altres llengües, etc.), o bé com a noms propis, que no són tinguts en compte en el procés de lematització, però que cal recuperar a l'hora de reconstruir el text. Cal remarcar, per a acostar una mica el lector a l'estructura dels fitxers informàtics que contenen tota la informació del corpus i dels programes que la tracten, que el text en la seva forma seqüencial (tal com es troba en les obres de referència) no és emmagatzemat en cap lloc de la memòria de l'ordinador; fins i tot en els casos en què la màquina ens forneix un context, aquest és construït cada vegada mitjançant l'execució dels programes de consulta, que en cerquen els diferents components entre les unitats d'informació que conté la base de dades, a les quals acabo de fer referència.

Per a l'explotació de la base de dades han estat elaborats una sèrie de programes que permeten, d'una banda, l'obtenció de llistats diversos, i, d'una altra banda, la consulta interactiva.[10]

Pel que fa als llistats que hom pot obtenir en aquest moment, fonamentalment poden ésser de lemes, de lemes i formes o de formes; les classificacions poden fer-se per ordre alfabètic directe, per ordre alfabètic invers o per ordre decreixent de freqüències. Aquests llistats es poden efectuar de tot el corpus, o de parts del corpus definides prèviament d'acord amb criteris cronològics o tipològics; poden realitzar-se també d'una obra o d'un conjunt d'obres determinades; d'altra banda, s'ha dissenyat també un tipus de llistat es què es pot apreciar la repartició de la freqüència dels diferents lemes amb criteris cronològics o tipològics. En l'apèndix d'aquest treball (figures I-V) en podem veure algunes mostres.

(10) Aquests programes foren elaborats inicialment per a la consulta interna i el manteniment de la base de dades, però, demostrada la seva utilitat general, i després de completar-los en alguns aspectes, avui permeten que el corpus pugui ésser consultat per qualsevol usuari per a obtenir-ne la informació necessària.

El sistema de consulta interactiva permet accedir a la base de dades a través del *lema*, a través de la *forma* o a través de la *localització*; però aquesta darrera modalitat, i, en part, la segona, tenen un interès exclusivament intern per a la validació i el manteniment de la base de dades. Per una altra banda, el sistema permet accedir, o bé al conjunt de tot el corpus, o bé a una part de les dades, que poden ésser seleccionades, com un subcorpus, per l'usuari; aquest subcorpus pot ésser definit també a partir de criteris cronològics, a partir de criteris tipològics, o d'ambdós alhora, o bé a partir d'un autor o d'un grup d'autors, o d'una obra o un grup d'obres; a més, si la consulta afecta un lema o una forma que té una freqüència molt elevada en el corpus, i, per tant, fóra molt feixuc de consultar-ne tots els contextos, el programa permet de ferne una selecció prèvia, a través de l'aplicació d'unes taules de nombres aleatoris, prenent com a referència un tant per cent del total d'ocurrències, o un nombre concret, fixats per l'usuari.

A través del sistema de consulta interactiva, podem obtenir, per exemple, si entrem per un lema determinat, la freqüència total del lema, les formes que té associades i la freqüència de cada una (vegeu-ne uns exemples, tal com ens apareixen a la pantalla de l'ordinador, a les figures 3 i 4), i, si accedim a continuació a una d'aquestes formes, obtenim les diferents ocurrències concretes que presenta en el corpus (vegeu la figura 5); podem accedir també al context que correspon a cada una d'aquestes ocurrències, i a una sèrie de dades relacionades, com són, l'autor i el títol de l'obra on es troben, l'any de publicació, el tipus de text (literari/no literari i les seves subdivisions), la localització específica (pàgina, línia, número d'ordre dins la línia), i les dades morfosintàctiques de la forma i del lema; si el context ofert en primera instància (tres línies físiques de l'edició de referència) fos insuficient per a la finalitat de la consulta, l'usuari pot accedir de manera immediata al paràgraf sencer; a les figures 6, 7 i 8 podem veure tres exemples d'obtenció de context, amb el context breu que apareix en primer lloc, i el context més extens que l'usuari pot demanar a continuació, si li cal; la figura 9 ens mostra un exemple de context en el cas d'un derivat apreciatiu associat al lema CAP *subst. masc.* i, encara, la figura 10, un exemple de context en un ús metalingüístic.

Per a poder proporcionar tota aquesta informació, el programa de consulta accedeix inicialment a la BDTLC, que conté només la informació més indispensable, sense repeticions innecessàries, amb la finalitat de fer les consultes el més àgils i ràpides possible. A partir d'aquestes dades, el programa recorre a uns altres fitxers: els del Diccionari Bàsic Informatitzat (DBI)[11] i els del Repertori d'Autors i Obres (RAO).[12] El DBI inclou les dades corresponents a la grafia dels lemes i de les formes i la seva categoria gramatical, i el codi de procedència (que ens indica si el lema o la forma es troben en alguna de les fonts lexicogràfiques o gramaticals utilitzades per a construir el DBI, o bé si no són recollits en aquestes fonts i, per tant, procedeixen únicament dels textos introduïts); d'altra banda, el programa obté del RAO el nom de l'autor, el títol de l'obra, l'any de publicació, i altres dades relacionades amb l'obra (tipus de llengua, etc.).

(11) El DBI és un element clau en l'operació de lematització, que es realitza per un procediment semiautomatitzat; consta d'un inventari de lemes i d'un inventari de formes (les formes flexionals que corresponen als dits lemes), cada una de les quals està relacionada lògicament amb el lema o els lemes a què pot estar associada. Inicialment el DBI contenia 88.067 lemes i 631.182 formes; en el moment actual, després d'haver-hi incorporat els lemes i les formes nous que han sorgit dins els textos en el curs de l'elaboració del corpus, consta de 160.952 lemes i 883.083 formes.

(12) El RAO fou constituït amb la finalitat de facilitar la tasca de selecció de les obres que havien de formar part del corpus. Aquest repertori, que consta de 6.273 autors i 26.121 obres, permeté creuar dades, com les dates de publicació de les obres, els tipus de text i els noms dels autors, a fi de poder prendre les decisions de la manera més objectiva possible. En el procés de consulta s'utilitza per a recuperar *in extenso* les dades referents a les publicacions, que en la BDTLC apareixen codificades.

BDTLC RECONSTRUCCIÓ SEGONS L'ORIGINAL A PARTIR D'UN LEMA

Lema: cap
Codi: 37.921

Cat. gram.: M
Freq. lema: 18.032

Selecció de la forma

	Forma	C.M.	Freq. abs.
	cap	FS	1
	cab	S	5
	cap	S	15.864
	cáp	S	6
	câp	S	2
	cabs	P	1
X	caps	P	2.006
	cáps	P	4
	cabet	DS	3
	caparró	DS	40
	caparronet	DS	3
	caparrot	DS	12
	caperrot	DS	1
	capet	DS	17
	cabets	DP	1
	caparrins	DP	1
	caparronets	DP	1
	caparrons	DP	7
	caparrots	DP	2
	caperrots	DP	1
	capets	DP	8
	cab	MET	1
	cabet	MET	1
	cabot	MET	1
	cap	MET	37
	caparró	MET	2
	caps	MET	4

FIGURA 3. Relació de formes corresponents al lema CAP M (substantiu masculí) tal com ens les mostra el programa de consulta interactiva de la base de dades. Hi podem apreciar, degudament codificades, les formes de singular (S) i de plural (P). També hi figuren els derivats apreciatius (diminutius, augmentatius, pejoratius o intensius), diferenciats amb el codi D, i les formes que apareixen en contextos metalingüístics (MET); per a cada una de les formes gramaticals podem observar també les variants gràfiques que han aparegut en els textos. Al costat de cada una de les formes diferenciades tenim la freqüència, és a dir, el nombre de vegades que apareix en el corpus. A la capçalera de la pantalla podem observar la suma d'aquestes freqüències, és a dir, la freqüència del lema. En aquesta imatge podem observar que hem seleccionat (X) la forma **caps** P, que té una freqüència igual a 2.006, a fi d'obtenir-ne més informació.

BDTLC RECONSTRUCCIÓ SEGONS L'ORIGINAL A PARTIR D'UN LEMA

Lema: gran
Codi: 42.814

Cat. gram.: AI
Freq. lema: 56.772

Selecció de la forma

Forma	C.M.	Freq. abs.	Forma	C.M.	Freq. abs.
G.	S	2	grandassos	DMP	1
g.	S	1	grandassots	DMP	1
gran	S	40.585	grandets	DMP	19
grán	S	33	grandissims	DMP	2
grân	S	2	grandíssims	DMP	7
grant	S	24	grandolassos	DMP	1
gran	P	3	grandots	DMP	1
grands	P	11	grandòts	DMP	1
grans	P	15.732	grandetes	DFP	5
gráns	P	1	grandísimas	DFP	1
grants	P	11	grandíssimas	DFP	3
regran	DS	3	grandíssimes	DFP	12
regrán	DS	2	gran	MET	30
regrans	DP	1	grand	MET	4
grandaçot	DMS	1	grandàs	MET	5
grandàs	DMS	1	grandassa	MET	2
grandet	DMS	30	grandassàs	MET	1
grandissim	DMS	8	grandassot	MET	1
grandíssim	DMS	52	grandet	MET	1
grandot	DMS	6	grandísim	MET	1
grandeta	DFS	9	grandissim	MET	1
grandísima	DFS	2	grandíssim	MET	3
grandissima	DFS	5	grandot	MET	2
grandíssima	DFS	134	grandota	MET	1
grandota	DFS	2	grans	MET	3
grandassassos	DMP	1	grant	MET	1

FIGURA 4. Aquesta imatge ens mostra el mateix pas del procés de consulta, però per a un altre lema, el lema GRAN AI (adjectiu invariable). La relació de formes, en aquest cas, és més llarga i complexa que la del lema CAP, com es pot apreciar a simple cop d'ull. Les formes G. i g. corresponen a casos en què el mot apareix abreujat: *G. A. U.* (Gran Arquitecte de l'Univers) en una obra de Josep Torras i Bages (*¿Què és la masoneria?*, 1884) i *g. fol.* (gran foli) en una obra de Jordi Rubió i Balaguer (*Com s'ordena i cataloga una biblioteca*, 1917), respectivament.

BDTLC RECONSTRUCCIÓ SEGONS L'ORIGINAL A PARTIR D'UN LEMA

Selecció de les localitzacions de la forma

Forma: caps Codi morf.: P Freq. abs.: 2.006
Lema: cap Cat. gram.: M Codi lema: 37.921

	Obra	Niv1	Niv2	Localitzacions	Obra	Niv1	Niv2	Localitzacions
	1.110	2		15,14,9	1.127			34,8,6
				15,18,11				34,24,10
X				17,7,3				34,28,1
	1.110	7		1,24,7				34,30,8
	1.111	4		1,25,10	1.128			117,16,4
				3,19,5	1.131			51,28,11
				4,13,3				80,9,3
	1.112	1		3,18,3	1.134			22,23,7
	1.112	2		20,36,7				29,12,7
	1.114	3		5,1,5				30,27,8
	1.116			33,17,2				83,28,2

FIGURA 5. Aquesta imatge ens mostra el pas següent del programa de consulta, que consisteix en la relació de totes les localitzacions corresponents a les ocurrències de la forma consultada. De fet aquí veiem només un petit fragment (les vint-i-dues localitzacions que caben en una pantalla) de la relació de les 2.006 ocurrències de la forma *caps*. Per a cada ocurrència tenim la localització específica expressada de la manera següent: codi de l'obra, número de la pàgina, número de la línia i número d'ordre del mot dins la línia; en algunes obres hi figura també una referència de nivell, que és una especificació intermèdia aplicada en els casos que dins una mateixa obra recomença la numeració de les pàgines (el cas més freqüent correspon a la premsa, en què no s'ha conservat la referència a la disposició física de la informació en l'original, sinó que s'ha sotmès a una reestructuració). En aquest cas hem seleccionat (X), a fi de demanar-ne la informació contextual, l'ocurrència que es troba a l'obra 110, codi 2 de nivell 1, pàg. 15, lín. 18, mot núm. 11.

BDTLC RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor: Publicacions Periòdiques
Títol: Tele-Estel, $$179, Barcelona, 1970.
Obra: 1.110 N.1: 2 N.2: Localització : 17,7,3

Any d'edició:	1970	Forma.	: caps
Tipus public.:	Premsa	C. morf.	: P Plural
Total d'ocurrències	: 39.705	Lema . .	: cap
Total de lemes usats	: 5.366	C. Gram.	: M Nom masculí
Total de formes usades	: 8.916	Codi lema:	37.921

CONTEXT

«d'aquella carrera històrica era la ciutat de Berlín. En aquell moment, però, els *caps* de les democràcies occidentals, amb l'exclusió de Churchill, no s'adonaren de l'envergadura del dilema. Roosevelt, molt»

Autor: Publicacions Periòdiques
Títol: Tele-Estel, $$179, Barcelona, 1970.
Obra: 1.110 N.1: 2 N.2: Localització : 17,7,3

CONTEXT

«La qüestió alemanya
Estats Units: En el moment de començar l'ofensiva dels aliats contra l'Alemanya de Hitler es produí una autèntica carrera entre els exèrcits anglosaxons que atacaven des de l'Oest i l'Exèrcit roig que venia de l'Est. El premi del que arribava primer a la meta podia molt bé ésser el predomini polític a l'Europa Central. La meta d'aquella carrera històrica era la ciutat de Berlín. En aquell moment, però, els *caps* de les democràcies occidentals, amb l'exclusió de Churchill, no s'adonaren de l'envergadura del dilema. Roosevelt, molt segur d'ell mateix, i convençut de la seva força d'atracció personal, (anomenava Stalin "oncle Joe", l'oncle Pep) estava convençut que el cap soviètic no tenia ambicions europees i que només volia ésser el primer a entrar a Berlín per qüestions de prestigi nacional. La història encara no ha posat en clar si la decisió de parar l'avanç de les forces nord-americanes per a permetre l'entrada primer que ningú de l'Exèrcit roig a la capital alemanya, fou d'Eisenhower o de Roosevelt. El fet, però, és que foren els soldats russos els que entraren victoriosament al Berlín hitlerià.»

FIGURA 6. En aquesta figura podem veure dues noves etapes del procés de consulta. En primer lloc, el programa ens mostra un context breu, que correspon sempre a tres línies físiques de l'edició de referència; aquest context va acompanyat, a la capçalera de la pantalla, per una informació detallada sobre l'obra a què correspon l'ocurrència seleccionada (en aquest cas, el núm. 179 de la revista *Tele-Estel*, publicat a Barcelona l'any 1970), el grup tipològic (Premsa), l'extensió de l'obra (39.705 ocurrències), el nombre de lemes que conté (5.366) i el nombre de formes (8.916). També hi veiem les especificacions gramaticals del lema i de la forma. En molts casos el context que se'ns mostra serà suficient per a l'objectiu que motiva la consulta; si no fos així, podem avançar un pas més per a obtenir el paràgraf sencer, que veiem en la segona pantalla reproduïda: aquesta darrera forma de representació, a diferència de l'anterior, manté la reproducció exacta del format original pel que fa a la divisió de línies.

BDTLC RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor: Francesc Curet i Payrot
Títol: Visions barcelonines 1760-1860 VIII. Muralles enllà
Obra: 485 N.1: N.2: Localització : 207,20,12

Any d'edició: 1956
Tipus public.: Ciències Socials
Total d'ocurrències : 99.013
Total de lemes usats : 8.016
Total de formes usades : 14.055

Forma. . : caps
C. morf. : P Plural
Lema . . : cap
C. Gram. : M Nom masculí
Codi lema: 37.921

CONTEXT

«Allò que no diuen els documents fidedignes, ho supleixen la tradició i la llegenda que omplen els buits, lliguen tots els *caps* i, prenent peu d'un petit indici, munten sovint un castell imaginari que»

Autor: Francesc Curet i Payrot
Títol: Visions barcelonines 1760-1860, VIII. Muralles enllà
Obra: 485 N.1: N.2: Localització : 207,20,12

CONTEXT

«Allò que no diuen els documents fidedignes, ho supleixen la tradició i la llegenda que omplen els buits, lliguen tots els *caps* i, prenent peu d'un petit indici, munten sovint un castell imaginari que hom arriba a prendre per real. En moltes tradicions hi ha, però, quelcom de veritat, inflada i deformada en trametre's de generació en generació. De les tradicions autèntiques hom pot fer-ne cas, si les estudia amb les degudes precaucions, perquè, almenys, ens expliquen la interpretació que donaven els avantpassats als fets ignots o poc coneguts. Quan la tradició és convertida en llegenda, esdevé ja una ficció rondallística o poètica que embelleix amb l'art de la imaginació, els motius en què és inspirada. Ben al contrari de les tradicions "a posteriori", o sigui les apòcrifes que hom inventa amb exuberància alarmant, i atribueix falsament als avantpassats, les quals, a més de llur absurditat, desorienten els qui de bona fe els donen crèdit.»

FIGURA 7. Una altra mostra del mateix estadi ens ofereix una nova ocurrència de la mateixa forma *caps:* en aquest cas de l'obra *Visions barcelonines 1760-1860,* de Francesc Curet i Payrot, corresponent al volum VIII, titulat *Muralles enllà,* publicat l'any 1956, i classificat com a Ciències socials. El context ens mostra un altre possible ús d'aquest lema.

BDTLC RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor: Emerencià Roig i Raventós
Títol: La pesca a Catalunya
Obra: 530 N.1: N.2: Localització : 116,2,2

Any d'edició: 1927
Tipus public.: Ciències Aplicad.
Total d'ocurrències : 25.038
Total de lemes usats : 2.843
Total de formes usades : 4.325

Forma. . : caps
C. morf. : P Plural
Lema . . : cap
C. Gram. : M Nom masculí
Codi lema: 37.921

CONTEXT

«de llargada, proveïdes de surada i plom, uns petits *caps* de corda als extrems per a llegir unes xarxes a les altres. Aquestes peces són quatre; el»

Autor: Emerencià Roig i Raventós
Títol: La pesca a Catalunya
Obra: 530 N.1: N.2: Localització : 116,2,2

CONTEXT

«Quan les barques surten a pescar a les quatre de la tarda, hom diu que van de prima; quan surten a les dues de la matinada, van de matinada. Tant en un cas, com en l'altre, les embarcacions van a unes dues o tres milles mar endins, on calen. Les peces de sardinals són unes xarxes de cinc canes de llargada, proveïdes de surada i plom, uns petits *caps* de corda als extrems per a lligar unes xarxes a les altres. Aquestes peces són quatre; el llur conjunt, és anomenat calada de sardinals.»

FIGURA 8. Encara un tercer exemple de context corresponent a la mateixa forma, en aquest cas de l'obra *La pesca a Catalunya* d'Emerencià Roig i Raventós, publicada l'any 1927, i classificada dins el grup Ciències aplicades. Amb aquests tres casos posats com a mostra el lector pot fer-se una idea dels resultats que pot proporcionar una consulta sistemàtica del corpus.

BDTLC RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor: Francesc Martínez i Martínez
Títol: Còses de la meua tèrra. (La Marina). Terça tanda i darrera
Obra: 441 N.1: N.2: Localització : 3,28,4

Any d'edició:	1947	Forma. .	: cabet
Tipus public.:	Ciències Socials	C. morf.	: DS Aspectual singular
Total d'ocurrències	: 57.133	Lema . .	: cap
Total de lemes usats	: 6.646	C. Gram.	: M Nom masculí
Total de formes usades	: 10.735	Codi lema:	37.921

CONTEXT

«be de jòcs, ara per a dormir a aquéll; per açò coloquen el *cabet* del nin al muscle i alvançant un pèu queden les cames obertes»

Autor: Francesc Martínez i Martínez
Títol: Còses de la meua tèrra. (La Marina). Terça tanda i darrera
Obra: 441 N.1: N.2: Localització : 3,28,4

CONTEXT

«La passejadora es una gicona que no te
més faena que la de aguantar el criançó, i
llevant del rato que aquèst te que mamar,
sempre están a la briva per els carrers juant
i fent mal de cap al vehinat ab ses cantoríes
be de jòcs, ara per a dormir a aquéll; per
açò coloquen el *cabet* del nin al muscle i
alvançant un pèu queden les cames obertes
a estil de compás, i començen a engrun-
sarse carregant el còs en un i atre peu alter-
nativament, cantant el consabut noni, noni...,
pegantli al mateix temps suaus tronets en la
esqueneta.»

FIGURA 9. Aquesta imatge ens mostra un exemple de forma amb codi D, és a dir, un derivat apreciatiu; en aquest cas la forma *cabet*, associada al mateix lema CAP. Correspon a l'obra *Còses de la meua tèrra*, de Francesc Martínez i Martínez, publicada l'any 1947 i classificada com a Ciències socials.

BDTLC RECONSTRUCCIÓ DEL CONTEXT SEGONS L'ORIGINAL

Autor: Manuel Sanchis Guarner
Títol: Gramàtica valenciana
Obra: 270 N.1: N.2: Localització : 113,32,2

Any d'edició:	1950	Forma. .	: cap
Tipus public.:	Llengua i Liter.	C. morf.	: MET Ús metalingüístic
Total d'ocurrències	: 66.777	Lema . .	: cap
Total de lemes usats	: 6.167	C. Gram.	: M Nom masculí
Total de formes usades	: 12.648	Codi lema	: 37.921

CONTEXT

«114. En posició final de paraula s'escriu generalment p; exemples: *cap*, cep, macip, glop, cup "recipient", camp, serp, colp, Calp, Asp, etc.»

FIGURA 10. Una darrera mostra, encara dins el procés de consulta de contextos, ens ofereix un exemple d'ús metalingüístic de la forma *cap*, corresponent al mateix lema CAP (substantiu masculí): es tracta de l'obra *Gramàtica catalana* de Manuel Sanchis Guarner, publicada l'any 1950, i classificada dins el grup Llengua i literatura.

SITUACIÓ ACTUAL I OBJECTIUS IMMEDIATS

Com es desprèn del que ja he dit al llarg d'aquest article, els treballs de constitució del CTILC són molt avançats en aquest moment.[13] La part del corpus que es refereix als textos de caràcter no literari és totalment acabada. Els 29.234.544 mots de què consta han estat lematitzats i la informació que resulta d'aquesta operació ha estat incorporada a la base de dades, de tal manera que pot ésser consultada i se'n poden extreure els llistats que preveuen els programes d'explotació. El seu caràcter representatiu i equilibrat, tant pel que fa a la dimensió temporal, com pel que fa a la tipologia dels textos que configuren aquesta part del corpus, fa que puguem considerar-la un corpus acabat, que pot servir de punt de referència per a l'estudi de la llengua, en la seva vessant no literària.

La situació de la resta del corpus (la part corresponent a la llengua literària) en aquest moment és la que segueix: hi ha 18.110.247 mots introduïts a l'ordinador i revisats; d'aquests, però, només 4.245.262 han estat lematitzats. Per a acomplir les previsions del projecte manca introduir, doncs, poc més de 5.000.000 de mots i lematitzar-ne uns 19.000.000. Tanmateix, el conjunt de mots lematitzats de text literari forma part també de la base de dades i ofereix les mateixes possibilitats de consulta i explo-

(13) Per bé que les tasques preparatòries i l'inici del projecte es remunten a l'any 1985, el gruix de la tasca de constitució i el treball regular consolidat a gran escala es produí durant els anys 1989 a 1992; si s'hagués pogut continuar el treball al mateix ritme que aquests anys, el corpus s'hauria acabat a finals de 1994, però un replantejament en els fons de finançament produí una disminució primer i després una aturada del procés de constitució (introducció de nous textos i lematització). Durant aquests dos darrers anys s'ha treballat sobretot en la revisió de la base de dades, en el millorament dels programes d'explotació, a fi de poder-ne oferir la consulta externa, i en la realització de càlculs relatius a la freqüència dels elements lèxics i a la seva distribució.

tació que la resta. Sobre les perspectives d'acabament d'aquesta part del corpus, només puc dir que depenen de les contingències pressupostàries; en tot cas, però, la finalització del corpus és un objectiu prioritari.

Quant a les possibilitats d'explotació i consulta del corpus, en parlar del contingut de la BDTLC ja he fet referència a la naturalesa dels programes elaborats amb aquest objectiu i a la naturalesa de la informació que podem obtenir a través de la seva aplicació. En aquest moment, però, l'usuari del sistema de consulta té també accés a dades específiques sobre la distribució de la freqüència de cada lema en els diferents grups tipològics; aquests càlculs formen part del projecte de publicació d'uns índexs de lemes del corpus amb informació detallada sobre les freqüències i llur distribució. La informació sobre la distribució de la freqüència dels lemes en els diferents grups en què es pot dividir el corpus amb criteris de caràcter tipològic afegeix una informació qualitativament molt important a la simple freqüència entesa com el nombre absolut d'ocurrències en el corpus, o, fins i tot, a la freqüència relativa (percentatge que representa la freqüència d'un lema en relació amb l'extensió del corpus). És generalment admès que, entre dos mots de freqüència igual o similar, el que reparteix les seves ocurrències d'una manera més uniforme entre els diferents tipus de text representa millor el vocabulari fonamental d'una llengua que no pas aquell que ofereix una repartició més desequilibrada, el qual, contràriament, representa el lèxic usat amb preferència en un tipus o uns tipus de text determinats.[14] Actualment el programa de consulta ofereix, doncs, d'una manera provisional, per a cada un dels lemes del subcorpus no literari, la distribució de la seva freqüència entre els deu grups tipològics que distingim (això és el que anomenem la freqüència real); també ens ofereix altres dades, com el que anomenem freqüència previsible o teòrica, que representa el valor hipotètic de la freqüència d'un lema en cada grup tipològic en el supòsit que es repartís d'una manera completament uniforme (es tracta de la repartició de la freqüència total d'un lema d'una manera proporcional al nombre total d'ocurrències que correspon a cada grup tipològic);[15] la diferència entre la freqüència real i la teòrica és la desviació absoluta respecte de les previsions teòriques, i també apareix a la pantalla quan consultem les freqüències dels lemes del corpus (a la figura VI de l'apèndix podem observar les dades de repartició de la freqüència d'un lema i la desviació tal com apareixen en la presentació que s'ofereix actualment).

Els valors d'aquesta desviació són il·lustratius, però cal tenir en compte que no són directament comparables entre els diferents lemes, perquè es refereixen a valors absoluts de la freqüència, la qual és diferent per a cada lema (només serien comparables en lemes de freqüència igual); d'altra banda, tampoc són directament comparables els corresponents a diferents grups tipològics dins un mateix lema, perquè representen valors absoluts relacionats amb una frequència teòrica diferent per a cada grup. A fi de poder comparar les diferències en la repartició dels diferents elements lexicals podem recórrer a l'expressió de la freqüència real en termes proporcionals a

(14) Hi ha diferents propostes per a substituir la freqüència per algun altre paràmetre, o per a modificar el valor de la freqüència en funció de la distribució, a fi de ponderar millor el valor dels elements lexicals en la llengua; en aquests moments, dins els treballs del CTILC estem estudiant la manera més adequada d'avaluar aquests conceptes i de classificar la informació amb el criteri més ponderat possible.

(15) Aquest concepte substitueix el de freqüència mitjana, que utilitzaríem si el corpus tingués el mateix nombre d'ocurrències en cada grup tipològic.

la freqüència teòrica (raó entre la freqüència real i la freqüència previsible), que representa un valor proporcional referit a la unitat; tenint en compte, doncs, que, si un lema presentés una repartició uniforme entre els diferents grups, aquest valor seria 1, les variacions respecte de la unitat representen una desviació proporcional en més o en menys: els valors inferiors a 1 representen una freqüència defectiva i els superiors a 1 una freqüència sobreabundant. En l'apèndix d'aquest treball, a manera d'exemple que permet fer-se càrrec de les possibilitats d'aquestes informacions, podeu veure el valor de la freqüència proporcional d'alguns lemes del corpus per a cada grup tipològic (figura VII), i a continuació (figura VIII), una representació gràfica d'aquests valors, que permet observar d'una manera fàcilment perceptible i en termes comparables, les diferències de repartició de la freqüència dels lemes en els diferents tipus de text.

Deixant de banda les moltes i variades aplicacions de la informació de caràcter quantitatiu i les conseqüències que se'n poden derivar, faré, per a acabar, només un breu esment de les perspectives futures i de les necessitats immediates en la tasca d'anàlisi i d'explotació del corpus. Al començ d'aquest treball he mencionat l'objectiu primordialment lexicogràfic que motivà aquest projecte, i he dit aleshores que un dels principis metodològics que el justifiquen és la possibilitat d'establir les significacions dels mots a partir de l'ús, i de poder determinar les combinacions reals d'elements lèxics més freqüents en el text, i els tipus d'estructures sintàctiques en les quals cada un d'aquests es pot trobar usat. Hem vist que en la fase de lematització s'atribueix a cada una de les ocurrències que constitueixen el text unes categories (morfològiques i sintàctiques) que permeten classificar la informació des del punt de vista gramatical; d'una manera anàloga, i en una fase ja pròpiament prelexicogràfica s'ha de començar a classificar les diferents ocurrències de cada lema d'acord amb una tipologia d'usos que permeti arribar a treure'n conseqüències relatives a l'aspecte semàntic i a l'aspecte sintàctic, així com també a propòsit de les associacions recurrents d'unitats lexicals en el text; a partir d'aquí podem prefigurar l'estructura dels articles d'un futur diccionari basat realment en les dades empíriques que l'ús ens forneix. Des del punt de vista de l'organització del treball, això representa l'elaboració d'una sèrie d'eines informàtiques que permetin prosseguir l'anàlisi del corpus en les línies indicades, i emmagatzemar adequadament els resultats d'aquesta recerca per a utilitzar-los en la tasca lexicogràfica pròpiament dita, que ha de donar lloc, d'una banda, a un o a diversos diccionaris en format convencional i en forma de llibre, però també, d'una altra banda, a un gran diccionari electrònic susceptible d'ésser consultat interactivament a través d'una xarxa entrellaçada de possibilitats.

JOAQUIM RAFEL I FONTANALS
Universitat de Barcelona
Institut d'Estudis Catalans

APÈNDIX

	Lema	c.g.	proc.	freqüència	lema principal
	relliscant	AI	DCC	1	
	relliscar	VI	DFA	147	
	relliscós	A	DFA	26	
	relló	M	DCC	2	
	rellogament	M	DCC	1	
	rellogar	VTP	DFE	3	
	rellogat	A	DCC	6	
	rellogat	M	DFA	8	
*	rellonge	M	DCC	14	rellotge
*	rellonger	M	DCC	1	rellotger
*	rellongeria	F	DCC	3	rellotgeria
	rellotge	M	DFA	969	
	rellotger	A	DEC	3	
	rellotger	M	DFA	53	
	rellotgera	F	DFA	1	
	rellotgeria	F	DFA	77	
	rellotges	MP	DFA	1	
	relluent	AI	DFA	5	
	relluir	VI	DFA	52	
*	relonge	M	DCC	1	rellotge
*	relotge	M	DCC	20	rellotge
*	relotger	M	DCC	2	rellotger
	reluctància	F	DFA	3	
	relum	M	DCC	1	
	relumbró	M	DCC	3	
	rem	M	DFA	336	
*	rém	M	DCC	10	raïm
	remà	L	DFA	1	
	remador	M	DFA	4	
	remagencar	VT	DFA	1	
	remaire	M	DEC	1	
*	remament	M	DCC	1	romanent
	remanament	M	DCC	1	
*	remandre	V	DCC	3	romandre
	remanegar	V	DCC	1	
	remanència	F	DCC	1	
*	remanent	AI	DCC	1	romanent
*	remanent	M	DCC	41	romanent
*	remangar	V	DCC	1	arremangar
	remar	VI	DFA	37	
	remarca	F	DFA	156	
	remarcable	AI	DFA	1.007	
	remarcablement	AV	DFA	51	
	remarcadament	AV	DCC	1	
	remarcar	VT	DFA	2.587	
	remarcat	A	DCC	34	
	remat	M	DCC	76	
	rematada	F	DFA	22	
	rematadament	AV	DFA	3	
	rematador	A	DCC	7	
	rematament	M	DCC	5	
	rematant	M	DCC	10	
	rematar	VTP	DFA	196	
	rematat	A	DEC	19	
*	remate	M	DCC	1	remat

FIGURA 1.

Índex general de lemes del subcorpus no literari. Ordenació alfabètica. En aquesta mostra podem observar la informació que ens dóna aquest tipus de llistat. En la columna de l'esquerra apareixen els lemes classificats alfabèticament; els que porten un asterisc al davant corresponen a la categoria de secundaris, i cada un d'aquests té a l'extrem dret del llistat la indicació del lema principal a què està associat. A continuació del lema apareix la categoria gramatical, i al costat, el que anomenem codi de procedència (és a dir la informació sobre si el lema en qüestió és recollit en les fonts lexicogràfiques bàsiques tingudes en compte (DFA = *Diccionari Fabra*; DEC = *Diccionari Enciclopèdia Catalana*), o bé no hi és enregistrat; en aquest cas es tracta d'un lema que ha estat donat d'alta en el DBI com a conseqüència d'haver aparegut en els textos del corpus (té el codi DCC) [el codi DFE s'utilitza en els casos en què el mot es troba a tots dos diccionaris citats, però amb codificació gramatical no coincident]. La columna numèrica ens dóna la freqüència absoluta de cada lema, és a dir, el nombre total d'ocurrències que presenta. El nombre total de lemes que apareixen en aquests índexs generals del corpus no literari és de 119.000.

FIGURA II.1

freqüència Absoluta	freqüència relativa	lema	c.g.	proc.
3.069.217	10,764607	el	AR	DFA
2.023.094	7,095559	de	PO	DFA
918.865	3,222718	i	C	DFA
713.677	2,503066	a	PO	DFA
697.763	2,447251	ell	P	DFA
617.377	2,165314	un	AR	DFA
578.536	2,029088	en	PO	DFA
565.692	1,984041	ésser	VI	DFA
502.994	1,764141	del	CT	DFA
486.397	1,705931	que	P	DFA
410.940	1,441282	per	PO	DFA
364.405	1,278070	que	C	DFA
318.723	1,117851	haver	VIA	DFA
282.654	0,991347	no	AV	DFA
238.759	0,837394	amb	PO	DFA
222.826	0,781513	al	CT	DFA
202.156	0,709017	aquest	A	DFA
163.743	0,574292	com	AV	DFA
155.990	0,547100	seu	A	DFA
149.752	0,525222	més	AV	DFA
148.471	0,520729	fer	VVP	DFA
147.592	0,517646	o	C	DFA
127.452	0,447010	tot	A	DFA
126.302	0,442976	poder	VT	DFA
119.389	0,418730	jo	P	DFA
116.943	0,410152	tenir	VTP	DFA
107.359	0,376538	hi	P	DFA
99.632	0,349437	altre	A	DFA
84.314	0,295712	dir	VVP	DFA
78.855	0,276566	si	C	DFA
76.208	0,267282	però	C	DFA
66.269	0,232424	anar	VA	DFA
61.548	0,215866	qual	AI	DFA
60.404	0,211853	nostre	A	DFA
58.546	0,205337	en	P	DFA
58.360	0,204684	mateix	A	DFA
52.215	0,183132	pel	CT	DFA
51.862	0,181894	any	M	DFA
51.523	0,180705	molt	AV	DFA
51.419	0,180340	donar	VTP	DFA
50.911	0,178559	estar	VVP	DFA
50.324	0,176500	gran	AI	DFA
50.114	0,175763	també	AV	DFA
48.647	0,170618	aquell	A	DFA
46.640	0,163579	dos	AN	DFA
46.164	0,161910	veure	VTP	DFA
46.143	0,161836	entre	PO	DFA
44.939	0,157613	quan	AV	DFA
44.680	0,156705	ja	AV	DFA
43.499	0,152563	primer	A	DFA
41.871	0,146853	això	P	DFA
41.292	0,144822	trobar	VVP	DFA
40.718	0,142809	perquè	C	DFA
40.408	0,141722	ho	P	DFA

FIGURA II.

Índex general de lemes del subcorpus no literari. Ordenació per freqüència (decreixent). Oferim al lector les tres primeres pàgines d'aquest índex, en les quals pot observar els 162 lemes més freqüents en el subcorpus no literari; a la primera columna apareix la freqüència absoluta; a la segona columna, la freqüència relativa o percentatge que representa la freqüència absoluta en relació al total de mots del corpus (en aquest cas del subcorpus no literari). Les altres dades ja són conegudes (grafia del lema, codi gramatical i codi de procedència).

FIGURA II.2

freqüència Absoluta	relativa	lema	c.g.	proc.
40.095	0,140624	anar	VI	DFA
39.065	0,137011	algun	A	DFA
38.426	0,134770	part	F	DFA
38.409	0,134711	son	A	DFA
37.801	0,132578	què	P	DFA
37.165	0,130348	home	M	DFA
37.156	0,130316	sobre	PO	DFA
36.602	0,128373	encara	AV	DFA
36.480	0,127945	dia	M	DFA
36.548	0,121169	cosa	F	DFA
34.341	0,120443	voler	VT	DFA
33.767	0,118430	ni	C	DFA
33.549	0,117665	sense	PO	DFA
32.590	0,114302	caldre	VI	DFA
32.263	0,113155	temps	M	DFA
32.215	0,112987	fins	PO	DFA
31.034	0,108844	manera	F	DFA
30.966	0,108606	tan	AV	DFA
30.889	0,108336	molt	A	DFA
30.029	0,105320	així	AV	DFA
29.833	0,104632	qui	P	DFA
29.004	0,101725	vida	F	DFA
28.158	0,098758	cada	AII	DFA
28.023	0,098284	tu	P	DFA
27.933	0,097968	bo	A	DFA
27.669	0,097042	nou	A	DFA
27.307	0,095773	on	AV	DFA
27.255	0,095590	sinó	C	DFA
27.150	0,095222	cas	M	DFA
26.668	0,093532	tant	AV	DFA
26.573	0,093198	sempre	AV	DFA
26.543	0,093093	saber	V	DFA
26.197	0,091880	lloc	M	DFA
26.194	0,091869	vegada	F	DFA
25.713	0,090182	forma	F	DFA
25.684	0,090081	posar	VVP	DFA
25.439	0,089221	llur	AI	DFA
25.205	0,088401	tot	P	DFA
24.919	0,087397	passar	VVP	DFA
24.813	0,087026	arribar	VIP	DFE
24.064	0,084399	déu	M	DFA
23.865	0,083701	des	PO	DFA
23.757	0,083322	obra	F	DFA
23.576	0,082687	poble	M	DFA
23.445	0,082228	català	A	DFA
23.147	0,081183	ben	AV	DFA
22.806	0,079987	deixar	VTP	DFA
22.527	0,079008	terra	F	DFA
22.425	0,078650	fet	M	DFA
21.511	0,075445	cap	AII	DFA
20.971	0,073551	ara	AV	DFA
20.571	0,072148	punt	M	DFA
20.357	0,071397	deure	VT	DFA
20.158	0,070699	segons	PO	DFA

FIGURA II.3

freqüència Absoluta	relativa	lema	c.g.	proc.
19.992	0,070117	general	AI	DFA
19.823	0,069524	estat	M	DFA
19.758	0,069296	aigua	F	DFA
19.345	0,067848	parlar	VVP	DFE
19.341	0,067834	hom	P	DFA
19.156	0,067185	casa	F	DFA
18.764	0,065810	senyor	M	DFA
18.580	0,065165	cert	A	DFA
18.312	0,064225	portar	VTP	DFA
18.207	0,063857	semblar	VI	DFA
18.182	0,063769	segle	M	DFA
17.954	0,062969	nom	M	DFA
17.935	0,062903	món	M	DFA
17.866	0,062661	tres	AN	DFA
17.828	0,062527	venir	VI	DFA
17.817	0,062489	treball	M	DFA
17.679	0,062005	doncs	C	DFA
17.670	0,061973	creure	VVP	DFA
17.612	0,061770	en	AR	DFA
17.344	0,060830	només	AV	DFA
17.254	0,060514	bé	C	DFA
17.199	0,060321	etcètera	L	DFA
17.129	0,060076	propi	A	DFA
16.910	0,059308	formar	VVP	DFE
16.794	0,058901	exemple	M	DFA
16.746	0,058732	país	M	DFA
16.408	0,057547	després	PO	DCC
16.384	0,057463	petit	A	DFA
16.110	0,056502	avui	AV	DFA
16.021	0,056190	prendre	VVP	DFA
15.868	0,055653	seguir	VTP	DFA
15.851	0,055593	durant	PO	DFA
15.827	0,055509	ciutat	F	DFA
15.815	0,055467	moment	M	DFA
15.684	0,055008	començar	V	DFA
15.469	0,054254	quedar	VIP	DFA
15.320	0,053731	conèixer	VVP	DFA
14.901	0,052262	presentar	VTP	DFA
14.864	0,052132	social	AI	DFA
14.829	0,052009	segon	A	DFA
14.786	0,051858	produir	VTP	DFA
14.720	0,051627	contra	PO	DFA
14.459	0,050711	sols	AV	DFA
14.320	0,050224	tal	AV	DFA
14.310	0,050189	paraula	F	DFA
14.243	0,049954	dins	PO	DFA
14.122	0,049529	diferent	AI	DFA
13.967	0,048986	sentit	M	DFA
13.942	0,048898	dret	M	DFA
13.915	0,048803	tractar	VVP	DFA
13.887	0,048705	quin	A	DFA
13.859	0,048607	relació	F	DFA
13.779	0,048326	acabar	VVP	DFE
13.738	0,048183	força	F	DFA

lema	c.g.	proc.	freqüència absoluta	freqüència relativa.
regularitzar	VT	DFA	134	0,000469
irregularitzar	V	DCC	2	0,000007
singularitzar	VTP	DFE	76	0,000266
popularitzar	VTP	DFE	132	0,000462
proletaritzar	VT	DEC	6	0,000021
desproletaritzar	V	DCC	5	0,000017
militaritzar	VT	DFA	8	0,000028
remilitaritzar	V	DCC	1	0,000003
sedentaritzar	V	DCC	2	0,000007
tartaritzar	VT	DFA	2	0,000007
iberitzar	V	DCC	2	0,000007
suberitzar	VTP	DFE	1	0,000003
canceritzar	V	DCC	1	0,000003
atmosferitzar	V	DCC	1	0,000003
estrangeritzar	VTP	DFA	1	0,000003
encoleritzar	V	DCC	1	0,000003
polimeritzar	VT	DFA	1	0,000003
caracteritzar	VTP	DFA	2.262	0,007933
descaracteritzar	V	DCC	5	0,000017
eteritzar	VT	DFA	1	0,000003
cateteritzar	V	DCC	1	0,000003
cauteritzar	VT	DFA	11	0,000038
parqueritzar	V	DCC	1	0,000003
pulveritzar	V	DCC	49	0,000171
vampiritzar	V	DCC	1	0,000003
satiritzar	VT	DFA	25	0,000087
martiritzar	VT	DFA	61	0,000213
arboritzar	V	DCC	1	0,000003
herboritzar	VI	DFA	19	0,000066
ruboritzar	VP	DFA	10	0,000035
teoritzar	VI	DFA	67	0,000234
meteoritzar	VTP	DFA	8	0,000028
metaforitzar	VI	DFA	1	0,000003
euforitzar	V	DCC	2	0,000007
categoritzar	V	DCC	6	0,000021
rigoritzar	V	DCC	1	0,000003
vigoritzar	VT	DFA	58	0,000203
revigoritzar	VT	DEC	2	0,000007
inferioritzar	V	DCC	4	0,000014
superioritzar	V	DCC	1	0,000003
interioritzar	V	DCC	45	0,000157
exterioritzar	VTP	DFE	186	0,000652
prioritzar	V	DCC	9	0,000031
majoritzar	V	DCC	2	0,000007
valoritzar	V	DCC	62	0,000217
avaloritzar	V	DCC	1	0,000003
revaloritzar	V	DCC	53	0,000185
supervaloritzar	V	DCC	1	0,000003
desvaloritzar	V	DCC	19	0,000066
cloritzar	V	DCC	1	0,000003
memoritzar	VT	DEC	17	0,000059
atemoritzar	V	DCC	21	0,000073
minoritzar	V	DCC	1	0,000003
sonoritzar	VT	DEC	4	0,000014

FIGURA III.

Índex general de lemes del subcorpus no literari. Ordenació alfabètica inversa. Aquest llistat ens presenta els lemes classificats alfabèticament per la dreta. Aquest tipus de classificació, anomenat correntment *invers* o *a tergo*, té com a objectiu de poder localitzar els mots per la seva terminació; és útil per a diversos tipus d'estudi, sobretot de morfologia (flexiva o lèxica). Les altres dades que hi apareixen ja han estat comentades.

fibrós		A	DFA		1	0,000353	
	fibrós		MS	DFA		1	0,000353
ficar		VTP	DFE		32	0,011316	
	ficar		IF	DFA		12	0,004243
	ficat		RMS	DFA		5	0,001768
	fico		1PI	DFA		1	0,000353
	fica		3PI	DFA		1	0,000353
	fiquen		6PI	DFA		1	0,000353
	ficava		1II	DFA		2	0,000707
	ficava		3II	DFA		5	0,001768
	ficà		3PT	DFA		2	0,000707
	ficaràs		2FU	DFA		1	0,000353
	fiquin		6PS	DFA		1	0,000353
	fica		2IM	DFA		1	0,000353
ficat		A	DCC		2	0,000707	
	ficat		MS	DCC		1	0,000353
	ficats		MP	DCC		1	0,000353
ficció		F	DFA		2	0,000707	
	ficció		S	DFA		1	0,000353
	ficcions		P	DFA		1	0,000353
fictici		A	DFA		2	0,000707	
	fictici		MS	DFA		1	0,000353
	fictícia		FS	DFA		1	0,000353
fidel		AI	DFA		10	0,003536	
	fidel		S	DFA		9	0,003182
	fidels		P	DFA		1	0,000353
fidelitat		F	DFA		3	0,001060	
	fidelitat		S	DFA		3	0,001060
figa		F	DFA		21	0,007426	
	figa		S	DFA		3	0,001060
	figues		P	DFA		18	0,006365
figuera		F	DFA		14	0,004950	
	figuera		S	DFA		9	0,003182
	figueres		P	DFA		5	0,001768
figura		F	DFA		33	0,011670	
	figura		S	DFA		26	0,009194
	figures		P	DFA		7	0,002475
figurar		VVP	DFA		3	0,001060	
	figurar		3PI	DFA		1	0,000353
	figurava		3II	DFA		2	0,000707
fil		M	DFA		13	0,004597	
	fil		S	DFA		8	0,002829
	fils		P	DFA		3	0,001060
	filet		DS	DCC		1	0,000353
	filets		DP	DCC		1	0,000353

FIGURA IV.

Índex de lemes i formes d'una obra. Aquest llistat ens mostra un índex de lemes amb les seves formes, però, en comptes de tractar-se d'un índex general de tot el corpus (que tindria un aspecte molt més complex), presenta la informació corresponent a una sola obra del corpus: *Camins de França* (1934), de Joan Puig i Ferreter. En el llistat apareixen els diferents lemes classificats alfabèticament (seguits del codi gramatical, el codi de procedència, la freqüència absoluta i la freqüència relativa); a continuació de cada lema hi ha les formes que li han estat associades, seguides del codi morfològic, el de procedència, la freqüència absoluta i la freqüència relativa. La classificació de les formes es fa d'acord amb una taula dels codis morfològics (el singular abans que el plural, el masculí abans que el femení, etc.).

		No Lit.	1	2	3	4	5	6	7	8	9	0
		45		18	8	5	2	1	7	2	2	
		0,000158		0,000616	0,000145	0,000146	0,000091	0,000022	0,000255	0,000091	0,000059	
1	agonitzant	52										
		VI	3	8	12	4		1	6	16	9	
2	agonitzar	80	0,000243	0,000214	0,000177	0,000274	0,000218	0,000117	0,000022	0,000219	0,000825	0,000269
		F	4			3		2	6		24	
3	agora	50	0,000152	0,000136	0,000237		0,000054		0,000045	0,000219		0,000716
		AV	11		5	1	2	1	1	1		
4	agosaradament	11	0,000033	0,000038		0,000171	0,000018	0,000058	0,000022	0,000036	0,000045	
		A	118	5	5	27	19	6	10	9	22	15
5	agosarat	170	0,000517	0,000414	0,000296	0,000171	0,000492	0,000556	0,000275	0,000226	0,000329	0,001008	0,000448	
		W	2474	9	94	414	466	211	395	211	40	614	20
6	agost	2.698	0,008176	0,008688	0,000533	0,003219	0,007546	0,013642	0,009704	0,008942	0,007714	0,001834	0,018375	0,017048
		M	25			4	2	16		2	1	
7	agotament	26	0,000079	0,000087		0,000072	0,000058	0,000362		0,000091	0,000029	
		V	73	6	3	9	11	3	22	2	3	14	
8	agotar	73	0,000222	0,000256	0,000355	0,000102	0,000164	0,000322	0,000137	0,000498	0,000073	0,000137	0,000418	
		A	35	2	3	4	6	2	9	3	3	2	1
9	agotat	36	0,000109	0,000122	0,000118	0,000102	0,000072	0,000175	0,000091	0,000203	0,000109	0,000137	0,000059	0,000852
		A	43		10	5	9	7	3		2	7	
10	agraciat	47	0,000142	0,000151	0,000342	0,000091	0,000263	0,000321	0,000067		0,000091	0,000209	
		AI	46	140	111	180	81	175	98	67	102	10	
11	agradable	1.400	0,004258	0,003547	0,002727	0,004795	0,002023	0,005269	0,003725	0,003961	0,003583	0,003072	0,003052	0,008524
		AV	39	2	2	2	8	2	11	3	2	6	
12	agradablement	47	0,000142	0,000136	0,000118	0,000068	0,000036	0,000234	0,000091	0,000249	0,000109	0,000091	0,000179	0,000852
		VIP	1789	77	180	187	650	90	124	205	109	126	41	
13	agradar	3.470	0,010555	0,006283	0,004565	0,006165	0,003408	0,019029	0,004139	0,002807	0,007495	0,004998	0,003770	0,034950
		A	12			2	4	1	1		2	1	1	
14	agradat	20	0,000060	0,000042		0,000068	0,000072	0,000029	0,000045		0,000073	0,000045	0,000029	
		A	35	2	4	5	1	12	2	2	3	4		
15	agradivol	59	0,000179	0,000122	0,000118	0,000137	0,000091	0,000029	0,000551	0,000045	0,000073	0,000137	0,000119	
		M	61		28	7	10	1	3	1	7	3	1	
16	agrado	61	0,000185	0,000214		0,000959	0,000127	0,000292	0,000045	0,000067	0,000036	0,000321	0,000089	0,000852
		A	222	5	16	61	19	18	16	36	22	25	4	
17	agrados	254	0,000772	0,000779	0,000296	0,000548	0,001111	0,000556	0,000827	0,000362	0,001316	0,001008	0,000748	0,003409
		AV	30	1	1	8	6	5		2	2	5		
18	agradosament	34	0,000103	0,000105	0,000059	0,000034	0,000145	0,000175	0,000229		0,000073	0,000091	0,000149	
		M	506	14	123	72	107	18	32	33	29	73	5	
19	agrairement	610	0,001865	0,001777	0,000830	0,004213	0,001312	0,003132	0,000827	0,000724	0,001206	0,001329	0,002184	0,004262
		VT	702	17	64	78	298	19	36	31	43	66	50	
20	agrair	920	0,002798	0,002465	0,001007	0,002192	0,001421	0,008724	0,000873	0,000815	0,001133	0,001971	0,001975	0,042622

FIGURA V.

Índex de lemes del subcorpus no literari. Ordenació alfabètica amb distribució tipològica de la freqüència. Aquest tipus de llistat ens mostra la distribució de la freqüència de cada lema entre els diferents grups tipològics en termes de freqüència absoluta i de freqüència relativa; aquest darrer valor permet comparar la freqüència d'un lema en els diferents grups, o la de diferents lemes en un grup. Els criteris de selecció permeten limitar l'aplicació d'aquest format als lemes que apareixen en *n* grups com a mínim i/o tenen una freqüència igual o superior a *x*. La mostra que oferim aquí pertany a un llistat restringit en el qual només figuren els lemes que apareixen almenys en 5 grups tipològics i presenten una freqüència total superior a 9 (10 o més).

Lema: 14394 F cresta

Grup tipològic	Freq. prev.	Freq. real	Desviació
1 Filosofia	28	1	27-
2 Religió i teologia	49	4	45-
3 Ciències Socials	93	22	71-
4 Premsa	59	12	47-
5 Ciències Pures. Ciències Naturals	37	99	62
6 Ciències Aplicades	75	173	98
7 Belles Arts. Divertiments. Esports	47	42	5-
8 Llengua i Literatura	37	10	27-
9 Història i geografia. Biografia	58	123	65
0 Correspondència	2		2-
		Total desviació :	1

Freqüència total: 562 Freqüència NO LITERARI: 486
Dispersió: 0,63426284 Ús: 308,25174024

Lema: 17671 F idea Acumulat de principal i secundaris

Grup tipològic	Freq. prev.	Freq. real	Desviació
1 Filosofia	633	1.756	1.123
2 Religió i teologia	1.102	1.058	44-
3 Ciències Socials	2.071	2.239	168
4 Premsa	1.324	966	358-
5 Ciències Pures. Ciències Naturals	825	502	323-
6 Ciències Aplicades	1.660	1.076	584-
7 Belles Arts. Divertiments. Esports	1.042	1.181	139
8 Llengua i Literatura	827	1.060	233
9 Història i geografia. Biografia	1.298	935	363-
0 Correspondència	44	54	10
		Total desviació :	1

Freqüència total: 12.441 Freqüència NO LITERARI: 10.827
Dispersió: 0,85991584 Ús: 9.310,30879968

FIGURA VI.

Una de les possibilitats de consulta del corpus es refereix a la informació sobre la freqüència, i, especialment, sobre la repartició de la freqüència d'un lema entre els diferents grups tipològics. Això és el que ens ofereix aquesta imatge per als lemes CRESTA i IDEA. La columna central és la que ens mostra la freqüència del lema en cada grup; la columna de l'esquerra ens mostra la freqüència teòrica o previsible (calculada repartint la freqüència total de manera proporcional al nombre total d'ocurrències que constitueix cada grup); i la columna de la dreta ens mostra la diferència entre la primera i la segona. Aquesta darrera xifra ens indica si en un grup determinat aquell lema és sobreabundant o deficitari, o bé si el nombre d'aparicions s'acosta molt al previsible en el cas que fos uniformement distribuït. Les xifres que apareixen més avall (dispersió i ús) formen part de les recerques que es fan en aquest moment per a expressar les diferències de repartició en forma de coeficient únic i per a modificar el valor de la freqüència en funció d'aquest coeficient.

	1. Filosofia	2. Religió i teologia	3. Ciències socials	4. Premsa	5. Ciències pures i naturals	6. Ciències aplicades	7. Belles arts, diverti-ments, esports	8. Llengua i literatura	9. Història, geografia, biografia	0. Correspondència
cresta	0.03	0.08	0.23	0.20	2.67	2.30	0.89	0.27	2.12	0
idea	2.77	0.96	1.08	0.72	0.60	0.64	1.13	1.28	0.72	1.22
invocar	1.29	3.32	1.36	0.46	0.12	0.28	0.62	0.61	0.87	0.66
rambla	0	0.19	0.84	3.90	0.22	0.35	1.13	0.18	1.19	0.20
escórrer	0.10	0.20	0.20	0.30	2.18	3.58	0.59	0.37	0.77	0
sanció	1.35	0.67	2.56	1.10	0.10	0.44	0.53	0.40	0.58	0
jovenesa	2.09	1.84	0.44	0.95	0.14	0.24	0.77	2.71	1.34	1
esposa	0.59	2.67	1.11	1.35	0.12	0.14	0.47	0.74	1.42	2.50
telèfon	0.05	0.02	0.12	7.47	0.08	0.19	0.05	0.08	0.06	0.08
fauna	0.08	0.05	0.13	0.32	9.70	0.20	0.57	0.13	0.76	0
rima	0.23	0.20	0.07	0.45	0.31	0	1.17	9.54	0.22	0
sarment	0	1.11	0.05	0.21	0.25	5.07	0.24	0.05	0.25	1

FIGURA VII.

En aquesta taula podem veure, per a una petita mostra de lemes del corpus, les diferències de repartició de la freqüència expressada de manera proporcional a la freqüència teòrica (és a dir, el valor que té respecte de la previsió si la repartició fos uniforme); aquesta forma d'expressar els valors de la desviació de la freqüència té l'avantatge de poder comparar-los entre ells, tant per grups, com per lemes. Si comparem els valors d'aquesta taula amb els que apareixen a la figura anterior (per al primer lema), observarem que realment la freqüència de CRESTA en el grup 1 és només el 0.03 de la previsió que podem fer si partim de la hipòtesi que el lema està uniformement repartit; en el grup 2, és el 0.08; en el 3, el 0.23; en el 4, el 0.2; en el grup 5, en canvi, és 2.67 vegades la previsió; en el 6, 2.3 vegades; en el 7 s'acosta a la previsió (el 0.89); en el 8, torna a ésser deficitari (el 0.27) i en el 9 és altra vegada sobreabundant (2.12 vegades la previsió). Si comparem les dades de la taula per columnes, podem observar, per exemple, que en el grup 1 (Filosofia) els mots *idea, jovenesa, sanció* i *invocar* són sobreabundants (en ordre decreixent) i la resta són deficitaris progressivament per aquest ordre: *esposa, rima, escórrer, fauna, telèfon, cresta*; *rambla* i *sarment* tenen en aquest grup una freqüència nul·la. I així podem anar comparant la resta de les dades; el judici, els comentaris i les conclusions els deixo per al lector.

FIGURA VIII.1

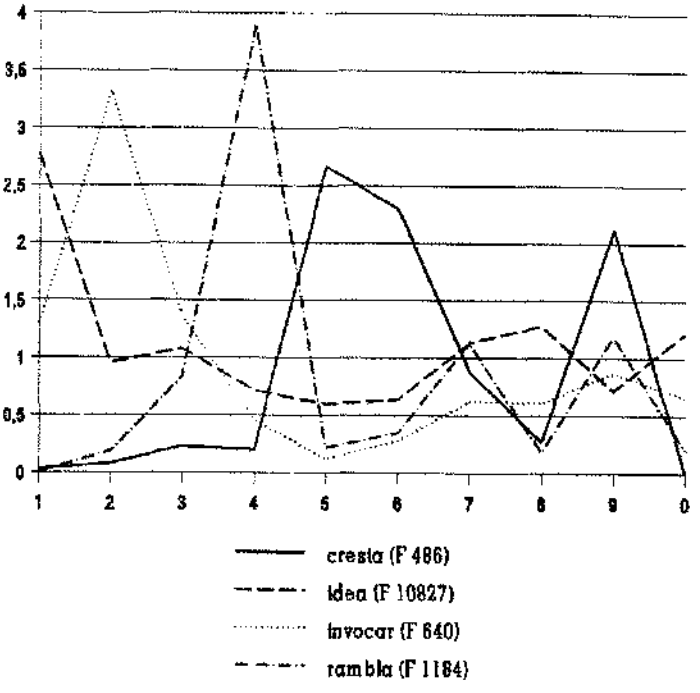

FIGURA VIII.2

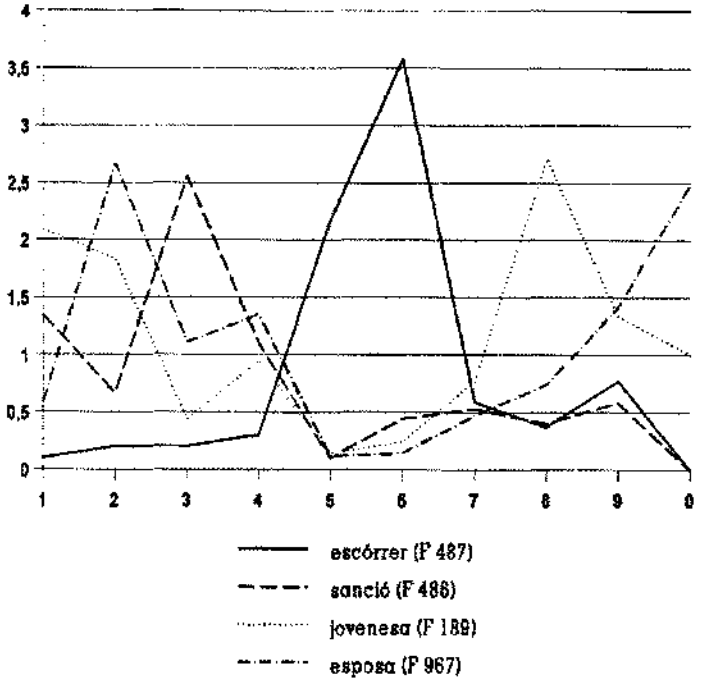

FIGURA VIII.

En aquestes imatges veiem la representació gràfica de les dades presentades en la figura anterior. Per a la interpretació d'aquests gràfics cal recordar que en el cas hipotètic que la freqüència d'un lema tingués una repartició completament uniforme entre els diferents grups, el valor proporcional de la freqüència seria 1, i, per tant, la seva representació gràfica seria una línia horitzontal situada en el valor 1 de l'escala de freqüències. Qualsevol desviació d'aquesta horitzontal hipotètica representa una variació, en més o en menys, respecte de la freqüència teòrica.

FIGURA VIII.3