INFORMES PRESENTATS DAVANT EL PLE

EL DICCIONARI DEL CATALÀ CONTEMPORANI

L'informe sobre l'estat del projecte del Diccionari del Català Contemporani fou presentat pel Dr. Joaquim Rafel davant el Ple el dia 21 d'octubre de 1994.

Origen del projecte. Supòsits metodològics i objectius

Amb el nom de Diccionari del Català Contemporani designem a l'Institut un programa de recerca d'un abast important, que té com a finalitat immediata de dotar l'Institut mateix en primer terme, però també el conjunt de la comunitat investigadora, d'una infraestructura bàsica de recerca lingüística en forma de corpus de referència de la llengua catalana. Abans, però, de fer una síntesi de les característiques del projecte i del seu desenvolupament al llarg d'aquests darrers anys, faré una breu al·lusió al seu origen i als plantejaments metodològics que foren tinguts en compte a l'hora de posar-lo en marxa.[1]

1. En altres llocs hem donat informació sobre l'origen i el desenvolupament d'aquest projecte: «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, p. 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, p. 147-161; «El Diccionari del Català Contemporani», *Serra d'Or* (1987), p. 428-431; «El "Corpus Textual Informatitzat de la Llengua Catalana" y el "Diccionari del Català Contemporani". Un proyecto del Institut d'Estudis Catalans». *Anthropos*, 1988, Documentación cultural e información bibliográfica, p. V-VII; «El Diccionari del Català Contemporani: treballs realitzats i previsions de futur», *Llengua & Literatura* [Revista anual de la Societat Catalana de Llengua i Literatura], núm. 5, 1992-1993 [1994], p. 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,

Hem de cercar l'origen d'aquest projecte en els moments en què l'Institut intentava incorporar-se, encara a poc a poc, a una activitat normal, després del lapse que havia suposat el llarg període de la dictadura, i, concretament, en la preocupació de la Secció Filològica, aleshores constituïda per molt pocs membres, per reprendre les activitats que l'havien caracteritzada abans de la interrupció total que provocà el desencadenament de la Guerra Civil. Un dels temes que havia quedat pendent de la primera època era el de l'anomenat «diccionari de l'Institut»; és conegut de tothom que en tot el període de la seva existència activa, aquesta institució no aconseguí elaborar el seu diccionari, i que aquesta mancança fou coberta per l'obra que redactà Pompeu Fabra, la qual, fins avui, és el diccionari reconegut com a normatiu. El desig, però, de la Secció Filològica en aquells moments, a principi dels anys vuitanta, era que la represa formal de la seva activitat creativa es produís d'acord amb els canvis que havia experimentat la societat en els cinquanta anys de pràctica inactivitat forçosa a què les circumstàncies de la història havien donat lloc; més concretament, si voleu, el desig de la Secció era que la represa de l'activitat productiva no fos insensible als avenços científics, metodològics o tecnològics que s'han produït en el camp de la recerca lingüística al llarg de tots aquests anys.

Fou dins aquest context que la Secció, al final de l'any 1982 o començament del 1983, decidí encarregar-me un informe sobre l'aplicació de la informàtica a la tasca lexicogràfica.[2] En aquest informe, que vaig lliurar el juliol de 1983, vaig posar en relleu la importància de servir-se dels mitjans que la tècnica moderna posa a la nostra disposició, per a acostar-nos al màxim a les exigències dels mètodes lexicogràfics més avançats. No és ara el moment de parlar de la profunda

---

Tübingen-Basel, 1993, p. 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, p. 131 i 135-137; *Memòria d'activitats 1984*, Barcelona, 1986, p. 136 i 137; *Memòria d'activitats: Curs 1988-1989*, Barcelona, 1990, p. 25-28; *Memòria d'activitats: Curs 1989-1990*, Barcelona, 1991, p. 137-139; *Memòria d'activitats. Curs 1990-1991*, Barcelona, 1992, p. 147-159 i *Memòria d'activitats: Curs 1991-1992*, Barcelona, 1993, p. 165-177. Un altre títol que forma part de la història del projecte és la publicació interna *Investigación lingüística 1989-1992*, Generalitat de Catalunya, Departament de Presidència, Comissió Interdepartamental de Recerca i Innovació Tecnològica (CIRIT). «Programes de recerca i desenvolupament de la Generalitat de Catalunya», núm. 3 (s. d.) [Anexo núm. 3. Diccionario del catalán contemporáneo. Corpus textual informatizado de la lengua catalana. Desarrollo general y estado actual del proyecto (43 p.). Anexo núm. 4. Diccionari del Català Contemporani. Informe sobre la fase de lematización. Descripción y métodos (34 p.)].

2. *Informe sobre les possibilitats d'aplicació dels mitjans informàtics a la confecció d'un diccionari del català contemporani i sobre els recursos necessaris*, juliol de 1983, 24 p. [inèdit].

renovació metodològica que ha experimentat la lexicografia en els darrers anys; només faré un breu esment d'aquells aspectes metodològics que justifiquen un projecte com el que ens ocupa.

És un sentir cada vegada més estès en el món de la lingüística que l'estudi sistemàtic dels diferents contextos en què un mot pot ocórrer és la millor manera, o l'única científicament vàlida, d'establir el significat o els significats que li podem atribuir; i això fins al punt de pensar que un mot aïllat no té sentit, o, si voleu, que un mot no té sentit si no és dins un context. És evident que una tasca lexicogràfica que vulgui tenir en compte aquest principi no pot ésser duta a terme per un lexicògraf o un grup de lexicògrafs que utilitzi la introspecció com a mètode principal de treball, ni tan sols a partir de buidats de textos establerts de manera selectiva segons criteris particulars. Fins fa relativament pocs anys, els lexicògrafs que volien treballar amb corpus o inventaris lingüístics establerts a partir de textos havien d'establir els seus fitxers amb criteris molt restrictius, si no volien trobar-se davant una quantitat de materials impossible de dominar, i amb un temps desmesurat per a la realització de l'obra.

Actualment, gràcies a les possibilitats que la informàtica posa al nostre abast en el camp del tractament de grans quantitats de dades, podem pensar en corpus molt més voluminosos, que siguin realment una mostra representativa de la llengua en el període de temps que volem estudiar; així, pel fet de no haver d'aplicar criteris tan restrictius com en els cedularis destinats a un tractament manual, podem obtenir resultats que reflecteixin d'una manera més objectiva i completa la realitat de la llengua. La utilització de mitjans informàtics permet, en efecte, d'emmagatzemar amb una relativa facilitat una gran quantitat de dades i permet de tractar aquestes dades amb una enorme flexibilitat, cosa indispensable per a poder-les sotmetre al tipus de tractament necessari per a un treball lexicogràfic basat en els supòsits que només he esbossat; aquests plantejaments preveuen no solament la determinació dels significats a partir de constatacions empíriques, sinó també l'estudi de les estructures sintàctiques de què pot formar part cada mot, les possibilitats combinatòries amb d'altres elements lèxics, etc., al costat de consideracions de caràcter estadístic, que poden arribar a tenir una gran importància en l'estudi científic d'una llengua.

En resum, doncs, en aquell moment vaig destacar la idea que la utilització de recursos informàtics va necessàriament lligada a la idea d'un diccionari de la llengua modern, no tant per a treballar més ràpidament amb els mateixos mètodes tradicionals, cosa també possible, sinó sobretot per a realitzar el treball a partir de mètodes lexicogràfics

nous; en conseqüència, vaig presentar com una condició indispensable el fet de poder utilitzar un corpus suficientment representatiu de la llengua com a font principal d'un futur diccionari descriptiu.

A la tardor de 1983 la Secció Filològica prenia l'acord d'iniciar els passos necessaris per a poder tirar endavant un projecte de creació d'un corpus textual de la llengua catalana com el que proposava en l'informe esmentat; a continuació, aquest programa fou assumit per l'Institut. Amb això, el que després hem anomenat Corpus Textual Informatitzat de la Llengua Catalana (CTILC) es convertia en la primera fase del projecte Diccionari del Català Contemporani (DCC).

Cal, potser, afegir encara en aquesta part introductòria que un dels aspectes tractats en aquell informe era la necessitat de preveure totes les aplicacions possibles d'un corpus d'aquesta naturalesa. És a dir, que, per bé que la motivació i alhora l'objectiu fonamental del corpus és de fer possible l'elaboració d'un diccionari descriptiu del català a partir dels supòsits metodològics d'allò que anomenem *la lexicografia moderna*, aquesta única finalitat no justificaria del tot l'enorme quantitat d'esforços ni els importants recursos que requereix la realització d'un projecte com aquest. És per això que sempre remarquem que el Corpus Textual Informatitzat de la Llengua Catalana, a part d'ésser la base o la font principal d'un futur diccionari descriptiu de la llengua, que tant de bo l'Institut pugui arribar a realitzar, és també útil, per no dir indispensable, per a qualsevol estudi sobre la llengua que vulgui basar-se en dades empíriques obtingudes a partir de textos escrits. Per això he dit al començ de la meva intervenció que els usufructuaris d'aquesta infraestructura bàsica de recerca són no solament l'Institut, que l'ha creada, sinó tota la comunitat d'investigadors que treballen sobre la llengua. De fet, en l'estat en què avui es troba el corpus, que tot seguit explicaré, ha estat utilitzat ja per a l'elaboració d'algunes tesis doctorals i per a d'altres treballs de recerca.

Desenvolupament del projecte

Fixació de les característiques del corpus i planificació dels treballs

A l'hora de posar en marxa aquest projecte, la Secció Filològica me n'encarregà la direcció i nomenà una comissió constituïda pels membres de la Secció senyors Joan Bastardas, Jordi Carbonell i el qui us parla, la qual havia de prendre les primeres decisions a propòsit d'una sèrie de qüestions fonamentals, com són l'establiment de les ca-

racterístiques del corpus i la determinació del procediment a seguir per a la selecció de les obres que l'havien d'integrar.

Deixant de banda tota una sèrie de qüestions tècniques o de detall que segur que serien d'interès dels presents, però que podrien allargar d'una manera excessiva aquesta exposició, faré només un breu esment de les decisions més rellevants que hom prengué en aquell moment; aquestes es refereixen als tres aspectes fonamentals que caracteritzen un corpus des del punt de vista general: el temporal, el qualitatiu i el quantitatiu.

L'aspecte temporal es refereix a les dates extremes dels textos que han de formar part del corpus. En el cas d'aquest, es tractava de determinar quina era la data més adequada per a constituir el punt de partida d'un corpus que havia de contenir textos publicats fins al moment més recent possible. Després de debatre a fons aquesta qüestió, la comissió acordà de fixar aquest terme *a quo* al voltant de 1833, com a data simbòlica representativa de la represa de l'ús literari del català en l'època moderna, amb la qual cosa l'abast temporal del corpus és de cent cinquanta-cinc anys, si mantenim el tancament a 1988, que és la data dels darrers textos actualment seleccionats.

L'aspecte qualitatiu es refereix a la naturalesa dels textos que s'han de prendre en consideració per a formar part del corpus. La caracterització general dels textos a tenir en compte fou la de textos o obres publicats, sigui quina sigui la seva forma;[3] contràriament al que havia estat habitual en corpus ja establerts per a d'altres llengües, a l'hora de definir el que ara presentem decidírem de prendre en consideració no solament textos de caràcter literari, sinó de qualsevol altre tipus, els quals hem anomenat «textos no literaris»; pel que fa a la proporció entre la llengua literària i la no literària, inicialment preveiérem que fos del 60 % i el 40 %, respectivament; després, però, a l'hora de seleccionar els textos a introduir en el corpus, comprovàrem que la gran riquesa i l'extraordinària varietat de les obres o publicacions de diversa naturalesa que no cabien sota la denominació de llengua literària ens obligaven a invertir pràcticament l'ordre d'aquests percentatges, donant així una importància o un pes més gran als textos no literaris, a fi de mantenir una representativitat adequada, en certa manera proporcional al volum de textos publicats i a la varietat temàtica.

Pel que fa a l'aspecte quantitatiu, que es refereix al que pròpiament

3. Els únics textos no publicats que han estat tinguts en compte són les cartes i les escriptures notarials.

anomenem *extensió del corpus* —la quantitat o volum de text de què ha de constar—, és evident que no pot ésser establert *a priori* amb exactitud en un corpus d'aquesta naturalesa, que no té com a característica principal el nombre de mots de què consta, sinó el seu caràcter representatiu i equilibrat; el que féu la comissió fou de prendre una referència orientativa a partir de la qual poder començar a treballar. Aquesta referència orientativa inicial fou la xifra de quaranta milions de mots del text. A l'hora, però, de seleccionar les obres que havien de formar part del corpus aplicant els criteris de selecció establerts, ens adonàrem que aquesta previsió es feia curta. L'estimació actual, en un moment en què tenim la major part del text del corpus introduït a l'ordinador, i, per tant, comptabilitzat, és d'entre cinquanta-tres i cinquanta-quatre milions de mots.

Pel que fa a la planificació dels treballs, establírem tres grans objectius successius: la selecció de les obres que havien de constituir el corpus, la introducció física dels textos a l'ordinador i la lematització.

Un corpus com aquest, que pretenia ésser un corpus general de referència de la llengua catalana moderna, havia de garantir un caràcter representatiu i equilibrat, tant des del punt de vista de la dimensió temporal, com des del punt de vista de la diversitat tipològica; l'objetiu era que en el corpus hi hagués el major grau de representativitat possible de tots els tipus de publicació en llengua catalana dins la franja temporal establerta. La comissió, doncs, establí una sèrie de grups cronològics (exactament vint-i-tres, vuit de deu anys cada un en l'època més antiga —1833-1913—, i quinze de cinc anys cada un en la més moderna —1914-1988—, a fi que a l'hora de fer la selecció es pogués assegurar, no solament des del punt de vista general o global, sinó també dins cada un d'aquests grups, 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 són: 0. Correspondència, 1. Filosofia, 2. Religió, 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).

Pel que fa a l'organització dels treballs relatius a la introducció dels textos a l'ordinador i a l'operació de lematització, a la qual em referiré tot seguit, fou deixada en mans del director del projecte i a l'equip de col·laboradors —lingüistes i informàtics— que començà a treballar-hi.

Execució del projecte

Una vegada establerts els criteris generals per part de la comissió que he esmentat, començaren pròpiament els treballs. Les primeres qüestions que abordàrem, per la urgència que tenien, foren la constitució d'un repertori informatitzat d'autors i obres i la fixació de criteris i elaboració dels programes informàtics per a la introducció dels textos a l'ordinador de manera que després permetessin un tractament adequat als objectius. El repertori d'autors i obres havia de servir per a fer més efectiva la selecció dels textos per al corpus, a base de fornir una informació al més completa possible i de permetre creuar dades, com la data de publicació, el tipus de text i l'autor. No puc parlar ara de detalls que donarien idea de la laboriositat d'aquesta tasca i de tot el procés de selecció; només afegiré que el Repertori d'Autors i Obres (RAO) constituït per a aquesta finalitat ha assolit l'extensió següent: 6.273 autors i 26.121 obres; d'aquestes obres, n'han estat seleccionades 3.302 (corresponents a 1.458 autors), repartides en 2.296 corresponents a la llengua no literària i 1.006 a la llengua literària.

Pel que fa a la incorporació adequada de les dades textuals al suport informàtic, he de dir que és també una operació amb un grau notable de complexitat; em referiré també només als seus aspectes més rellevants. Després de seleccionar les obres que han de formar part del corpus, cada una ha d'ésser objecte d'un estudi individualitzat a fi de posar de relleu les seves característiques més específiques, les quals poden tenir algun tipus de repercussió en aquesta fase d'introducció de la informació en l'ordinador. A causa de la gran diversitat dels textos, de les característiques materials dels originals, i, en molts casos, del seu estat físic, no ha estat possible la utilització de procediments de reconeixement òptic; el text s'introdueix, per tant, a través d'una operació de teclat, amb incorporació simultània d'una codificació que en permet el tractament informàtic adequat segons les previsions del projecte. Dins aquesta fase es realitzen també diversos tipus de revisions del text introduït, a fi de garantir-ne, per una banda, la fidelitat a l'original, que és un objectiu primordial, i, per una altra banda, la correcta incorporació dels codis que han de permetre una interpretació adequada de les dades per part dels programes informàtics que les han de tractar. En realitat, en aquesta fase es duu a terme, doncs, no solament la incorporació a l'ordinador de les dades textuals, degudament codificades, sinó també la validació de la informació introduïda. Aquesta fase culmina amb l'aplicació dels programes de segmentació del text en mots, els quals donen lloc a uns fitxers de mots, a partir dels quals hom pot obtenir ja diversos tipus de llistats (classificacions diverses, informació sobre freqüències, concordances). Notem, però, que fins aquí hem utilitzat el terme *mot* amb dos valors

diferents: el mot del text, o ocurrència, i el mot considerat com a cadena de caràcters (mot gràfic o grafia): d'acord amb aquest segon criteri, tot allò que correntment consideraríem mots repetits queda reduït a una sola unitat en aquests fitxers de mots; al final d'aquesta fase, doncs, podem saber que una obra, com, per exemple, *Nosaltres els valencians*, de Joan Fuster, que té una extensió de 69.914 mots (en el sentit de mots del text), consta de 9.753 mots (en el sentit d'unitats diferents des del punt de vista gràfic).

Una de les característiques que donen un valor especial al nostre corpus és que no ens limitem a emmagatzemar la informació tal com resulta al final de les operacions a què m'acabo de referir, sinó que la sotmetem encara a una nova operació, que representa donar un pas endavant en l'anàlisi de les dades; aquest és un pas qualitativament important: es tracta de la lematització. Per a aquells que no estiguin familiaritzats amb aquest concepte, només diré que la lematització representa una primera anàlisi de caràcter lingüístic de les dades; en la seva especificitat, però, pot ésser de naturalesa diversa, segons els criteris utilitzats; a més, el procés pròpiament dit es pot dur a terme d'acord amb més d'un procediment diferent.

A través de la lematització s'aconsegueixen 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 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 a través d'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 un corpus lematitzat, en canvi, d'acord amb aquests principis, l'usuari té accés a les dades agrupades adequadament a partir del substantiu masculí CAP, del verb CABRE, 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.

Doncs bé, un dels aspectes més importants en l'execució d'aquest projecte ha estat la determinació dels criteris per a resoldre els nombrosos aspectes problemàtics que presenta el procés de lematització, i, per una altra banda, l'establiment del procediment semiautomatitzat a través del qual es realitza. I això, al marge de la mateixa execució del procés, que es va acomplint progressivament després de l'operació anterior (introducció i validació de la informació textual i procés de separació de mots). Per raons òbvies de temps, no puc intentar en la meva exposició ni tan sols una aproximació a totes aquestes qüestions; crec, però, que el que he dit pot ajudar a fer-se càrrec de la complexitat i de la laboriositat d'aquestes operacions, sobretot si tenim en compte que tractant amb dades lingüístiques reals la casuística és gairebé infinita, o, si més no, és pràcticament impossible de preveure en la seva totalitat. Això ha fet que al llarg de l'execució del projecte hi hagi un continuat procés de retroalimentació, en virtut del qual, les noves dades que van apareixent enriqueixen els coneixements que tenim de la realitat lingüística i permeten de millorar i de matisar els criteris i els procediments.

La Base de Dades Textual de la Llengua Catalana

Una vegada resoltes totes les qüestions relatives a allò que podem considerar la constitució del corpus pròpiament dita, ens hem ocupat

dels aspectes relatius a la seva explotació, és a dir, de les possibilitats de consulta i d'obtenció i utilització de les dades que conté.

Com a resultat dels treballs fets fins ara en aquesta línia, després de les operacions que acabo de descriure d'una manera superficial, s'incorporen a una única base de dades totes aquelles informacions que són necessàries per a un aprofitament del contingut del corpus. Això és el que anomenem Base de Dades Textual de la Llengua Catalana (BDTLC). No puc entrar tampoc en detalls sobre la naturalesa de les informacions que conté aquesta base de dades; només afegiré, per a completar aquest informe, que han estat confeccionats també una sèrie de programes destinats a l'explotació del seu contingut; aquests programes permeten, per una banda, l'elaboració de llistats de naturalesa diversa, i, per una altra banda, la consulta interactiva a través d'una pantalla connectada a l'ordinador.

Els llistats que hom pot obtenir en aquest moment es refereixen a diferents tipus d'ordenacions (alfabètica directa, alfabètica inversa, per ordre de freqüències, amb distribució de les freqüències segons criteris cronològics o tipològics, segons el tipus de llengua, els codis gramaticals, etc.).

El sistema de consulta interactiva permet d'accedir a la base de dades a través del *lema*, a través de la *forma* o a través de la *localització* (la localització física en la font); aquesta darrera modalitat, però, 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 d'accedir, o bé al conjunt de tot el corpus, o bé a una part de les dades, que poden ésser seleccionades per l'usuari com un subcorpus; aquest subcorpus pot ésser definit 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 fer-ne una selecció prèvia, de caràcter aleatori, a partir d'un tant per cent del total d'ocurrències, o d'un nombre concret, fixats per l'usuari.

A través del sistema de consulta, podem obtenir, per exemple, si entrem per un lema determinat, la freqüència total del lema, les formes que té associades en el corpus i la freqüència de cada una, i, si accedim a una d'aquestes formes, les diferents ocurrències concretes que presenta en el corpus; podem veure a continuació el context que correspon a cada una d'aquestes ocurrències, amb una sèrie de dades relacionades, com són l'autor i el títol de l'obra a què correspon cada

ocurrència, 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.

Pel que fa al context corresponent a cada ocurrència examinada, en primera instància apareix l'extensió corresponent a tres línies físiques de l'edició de referència, del text originari; però l'usuari pot accedir immediatament al paràgraf sencer, si el context inicial és insuficient.

Aquest sistema de consulta funciona actualment i tots vostès són convidats a utilitzar-lo.

Evolució i estat actual dels treballs

Els treballs d'execució del projecte s'iniciaren al començament de 1985 amb recursos limitats.[4] A part d'aspectes materials, com el condicionament de locals i l'establiment d'una infraestructura informàtica adequada per al tractament de les dades, i de qüestions com la formació de personal especialitzat, durant els quatre primers anys treballàrem fonamentalment en la preparació del Repertori d'Autors i Obres, en la selecció de les obres que havien de formar part del corpus, en el disseny i elaboració del sistema d'introducció i de verificació de les dades i en el disseny i elaboració del sistema de lematització semiautomatitzada.

En l'apartat que he dedicat a la descripció dels treballs del corpus, ja m'he referit al RAO com un instrument necessari per a poder portar a terme, amb les màximes garanties de representativitat i equilibri, la selecció de les obres per a integrar en el corpus; doncs bé, la realització d'aquest fitxer informatitzat s'emprengué des del primer moment i aviat estigué en condicions d'ésser utilitzat en la fase de selecció. La selecció pròpiament dita s'inicià també molt al començament, i es perllongà al llarg de diversos anys, perquè és una tasca lenta i feixuga, tal com es desprèn de la breu descripció que n'he fet més amunt; ara és totalment acabada, i, tal com he dit també, el nombre d'obres seleccionades és de 3.302 (2.296 corresponents a textos no literaris i 1.006 a textos literaris).

4. Des del punt de vista pressupostari, aquest programa ha estat dotat al llarg del seu desenvolupament pel mateix IEC, per la Direcció General d'Investigació Científica i Tècnica (DGICYT) del Ministeri d'Educació i Ciència i per la Comissió Interdepartamental de Recerca i Innovació Tecnològica (CIRIT) de la Generalitat de Catalunya; ha rebut també subvencions del Departament de Treball de la Generalitat de Catalunya.

Durant aquests primers anys es realitzà i es perfeccionà el disseny del sistema d'introducció i verificació de dades i s'elaboraren els programes per a executar-lo; això implicà la presa d'una sèrie de decisions a propòsit del tipus de codificació a utilitzar, íntimament relacionades amb la naturalesa dels programes que s'havien d'elaborar per a tractar el text posteriorment. Aquests treballs donaren com a resultat, per una banda, els criteris per a la introducció del text a l'ordinador, que es materialitzen en un manual d'introducció de dades, i, per una altra banda, els programes d'introducció pròpiament dita i d'esmena de dades, i els programes de verificació automàtica (que permeten descobrir automàticament la presència de certs possibles errors), els de llistat del text introduït (per a una verificació manual), i el de separació de mots, que permet localitzar certs errors altrament difícils de trobar, i, a més, en el seu procés de realització dóna lloc als fitxers de mots sense lematitzar que serveixen d'*input* en la fase de lematització.

També durant aquest període inicial s'abordaren els diferents aspectes relatius al procediment de lematització. Aquesta operació, des del punt de vista lingüístic, és aparentment simple, però presenta una sèrie de problemes davant els quals s'han de prendre decisions, tant a l'hora d'establir els lemes de referència, com a l'hora de determinar l'atribució de determinades formes a un lema o a un altre; el conjunt d'aquests problemes arriba a ésser prou important; cada un d'ells ha estat estudiat detingudament i les decisions que s'han pogut prendre com a resultat d'aquest estudi han donat lloc a un manual de lematització, que conté explícitament les normes que s'han de posar en pràctica durant el procés, a fi d'assegurar, fins on sigui possible, la unitat de criteri. A part, però, dels criteris lingüístics, el procés de lematització requeria l'establiment d'uns instruments informàtics que permetessin la realització d'aquesta operació d'una manera semiautomatitzada; això implicava, per una banda, la constitució d'un diccionari de màquina, que anomenem Diccionari Bàsic Informatitzat (DBI) —el qual conté una relació exhaustiva de lemes i una relació exhaustiva de formes, degudament relacionades amb els lemes que els corresponen—,[5] i, per una altra banda, l'elaboració dels programes informàtics adequats per a aquesta finalitat, tots els quals han estat realitzats en l'àmbit del centre d'informàtica de l'Institut (Centre d'Estudis i

5. El DBI fou constituït a partir de fonts lexicogràfiques (pel que fa als lemes) i a partir de les previsions de la gramàtica (pel que fa al desenvolupament de les formes inflectives corresponents a cada lema). El DBI originari contenia 88.067 lemes i 631.286 formes, però al llarg del desenvolupament del projecte, des de l'inici dels treballs de lematització, s'han hagut de donar d'alta 72.885 lemes nous i 251.796 formes noves, que han anat apareixent en els textos tractats i no figuren en les fonts lexicogràfiques de referència.

Desenvolupaments Informàtics); doncs bé, a finals de 1988 s'havien aconseguit tots aquests objectius i s'havia començat a treballar productivament en la fase d'introducció de dades, de tal manera que en aquell moment comptàvem amb 6.500.000 mots introduïts i verificats; no, però, encara, lematitzats.

A principi de 1989, havent establert clarament aquestes fases prèvies (selecció d'obres, criteris d'introducció i de verificació, criteris de lematització) i havent resolt els aspectes tècnics que planteja un projecte d'aquesta naturalesa, un increment important dels recursos destinats a aquesta finalitat (gràcies a un conveni per tres anys signat entre l'Institut d'Estudis Catalans i la Secretaria d'Estat d'Universitats i Investigació, fruit d'una acció de política científica) permeté d'accelerar notablement la realització dels treballs d'introducció de dades i de lematització fins a finals de 1992, any en què col·laboraren en el finançament del projecte el Ministeri d'Educació i Ciència i la Comissió Interdepartamental de Recerca i Innovació Tecnològica (CIRIT) de la Generalitat de Catalunya; a partir d'aquest any, però, a causa de dificultats per a trobar una forma de finançament compartida, els recursos dedicats al projecte disminuïren. En les taules adjuntes i en la figura a part, podem observar l'evolució del volum de text introduït i lematitzat al llarg d'aquests darrers cursos, on s'observa la pràctica paralització en els dos darrers anys.

Si haguéssim pogut continuar amb el mateix nivell de finançament que els anys anteriors, el corpus s'hauria acabat completament a finals d'aquest any 1994.

TEXT INTRODUÏT

Data	*Llengua literària*	*Llengua no literària*	*Total*
31-XII-88	5.941.272	577.051	6.518.323
31-XII-89	7.316.331	5.862.195	13.178.526
31-XII-90	8.692.489	15.857.027	24.549.516
31-XII-91	9.258.576	27.054.538	36.313.114
31-XII-92	17.047.496	29.286.521	46.334.017
31-XII-93	18.110.247	29.234.544	47.344.791

TEXT LEMATITZAT

Data	*Llengua literària*	*Llengua no literària*	*Total*
31-XII-89	433.036	1.376.952	1.809.988
31-XII-90	2.486.692	7.629.382	10.116.074
31-XII-91	2.635.242	19.980.080	22.615.322
31-XII-92	2.746.863	27.427.599	30.174.462
31-XII-93	4.245.262	29.234.544	33.479.806

Pel que fa a la situació actual, és la següent: la part del corpus corresponent a la llengua no literària està completament acabada; la seva extensió és de 29.234.544 mots introduïts i lematitzats, que configuren, en realitat, un corpus de la llengua no literària; resta, però, en canvi, una part important de text literari per introduir (uns 6.500.000 mots), i una part encara més important per lematitzar (uns 20.300.000 mots).

Cal afegir, per a completar aquesta visió de l'evolució dels treballs del projecte al llarg del temps, que durant els anys 1990 i 1991 es dugueren a terme els treballs encaminats a la creació de la base de dades a què m'he referit fa un moment (Base de Dades Textual de la Llengua Catalana), que, com he dit, és el que permet pròpiament l'explotació del corpus; a final de 1992 quedà la base de dades completament constituïda, i a partir d'aquest moment s'hi incorporen d'una manera automàtica els resultats del procés de lematització a mesura que es donen per bons. Durant els dos darrers anys, per bé que no s'ha pogut treballar en l'increment del volum de dades introduïdes o lematitzades, ens hem ocupat de tasques de revisió de les dades acumulades i de manteniment de la base de dades, del millorament dels programes d'explotació i de la realització de càlculs relatius a la freqüència dels elements lexicals, que probablement donaran lloc a la publicació d'un diccionari de freqüències, eina fonamental per a l'estudi de certs aspectes del lèxic d'una llengua.

Resumint, doncs, en aquest moment tenim completament acabada la part del corpus que correspon a la llengua no literària; pel que fa a la llengua literària, tenim introduïts a l'ordinador una bona part dels textos seleccionats, però, en canvi, la lematització d'aquesta part del corpus és encara endarrerida. Tanmateix, la part del corpus que està completament acabada, tant la que correspon a la llengua no literària, com la part lematitzada de la llengua literària, s'integra en una base de dades única, que disposa d'uns programes d'explotació i de

consulta que permeten una utilització de les dades per a diverses finalitats d'estudi científic de la llengua.

Confiem que les dificultats aparegudes darrerament per al finançament del projecte trobaran aviat una solució que permeti la finalització del corpus en un termini raonable.

Joaquim Rafel i Fontanals
Director del Diccionari del Català Contemporani