JOAQUIM RAFEL I FONTANALS

EL «DICCIONARI DEL CATALÀ CONTEMPORANI»: TREBALLS REALITZATS I PREVISIONS DE FUTUR

Presentació

En altres llocs hem donat informació sobre l'origen i el desenvolupament d'aquest projecte.[1] Potser només caldrà recordar ara que el nom de *Diccionari del català contemporani* designa un programa de recerca —avui un centre de l'Institut d'Estudis Catalans— que sorgí del desig de la Secció Filològica de poder incorporar a l'elaboració d'un diccionari descriptiu de la llengua els avenços científics, metodològics i tecnològics que la lexicografia ha assumit en els darrers anys.

A l'hora de definir aquest projecte (anys 1983 i 1984) fou valorada d'una manera molt positiva i considerada una tasca preferent la constitució d'un *corpus textual informatitzat*, a fi de poder abordar per als diferents mots de la llengua una anàlisi basada fonamentalment en l'ús i determinar per aquest procediment tant els aspectes relatius a la freqüència d'aparició dels elements lèxics en els textos, com els que es refereixen pròpiament al significat. Hom prengué aquesta decisió des del convenciment que la determinació i la caracterització dels elements lèxics de la llengua ha de partir de l'anàlisi dels diferents contextos en què apareixen, d'acord amb la idea que, en general, un mot no té sentit sinó quan es troba en un context.

1. *Cap a un diccionari del català contemporani*, Comunicació presentada al II Congrés Internacional de la Llengua Catalana, àrea de *Lingüística social*, Palma de Mallorca 30 d'abril a 4 de maig de 1986 (en premsa a les actes del Congrés). *El «Corpus textual automatitzat de la llengua catalana»* (en col·laboració amb J. Ma. Solanellas), *Actas de las II Jornadas Españolas de Documentación Automatizada. 20-22 de novembre de 1986. Ponencias y comunicaciones* (Torremolinos-Màlaga 1986), ps. 147-161. *El «Diccionari del català contemporani»*, «Serra d'Or», 1987, ps. 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*, ps. V-VII. En les diferents memòries d'activitats de l'Institut d'Estudis Catalans, d'altra banda, hom pot trobar informació sobre l'estudi previ, l'aprovació, l'inici i el desnvolupament progressiu d'aquest projecte: *Memòria d'activitats (octubre 1982-desembre 1983)* (Barcelona 1984), ps. 131 i 135-137; *Memòria d'activitats 1984* (Barcelona 1986), ps. 136 i 137; *Memòria d'activitats. Curs 1988-89* (Barcelona 1990), ps. 25-28; *Memòria d'activitats. Curs 1989-90* (Barcelona 1991), ps. 137-139; *Memòria d'activitats. Curs 1990-91* (Barcelona 1992), ps. 147-159.

D'acord amb aquest plantejament fou definit el *Corpus textual informatitzat de la llengua catalana* com a primera etapa del programa *Diccionari del català contemporani*. Abans de poder elaborar pròpiament un projecte per a la redacció d'un diccionari descriptiu concebut a partir dels supòsits que acabem d'esbossar, calia comptar amb un corpus suficientment representatiu de la llengua del període de temps a estudiar, disposat sobre suport informàtic, a fi de poder extreure'n la informació adequada. A l'hora de prendre les decisions encaminades a aquests objectius hom tingué en compte que l'esforç que demana un treball d'aquesta naturalesa no podia reduir els seus objectius a la redacció d'un mer diccionari convencional, sinó que aquest corpus havia de quedar configurat de tal manera que, una vegada constituït, permetés d'extreure'n no tan sols les informacions específiques considerades necessàries per a la redacció del diccionari previst, sinó qualsevol altra informació sobre la naturalesa de la llengua escrita en el període considerat. Amb aquest plantejament, en la descripció del projecte, el *Corpus textual informatitzat de la llengua catalana* és definit com una base de dades lèxica amb informació textual, que podrà donar lloc a més d'un producte concebut en forma de diccionari, però que també serà explotable des d'altres punts de vista i permetrà, en conseqüència, la realització dels més diversos estudis sobre la llengua.

L'abast temporal del corpus fou fixat entre 1833 i 1988, o sigui els cent cinquanta-cinc anys darrers de la història de la llengua. L'extensió del corpus fou establerta inicialment en quaranta milions d'ocurrències de mots (o sigui, de mots del text). Aquest nombre de mots fou distribuït *a priori* en dues parts: un 60 % corresponent a textos de caràcter literari i un 40 % a textos de caràcter no literari; dins cada una d'aquestes dues modalitats es féu una distribució segons els gèneres (assaigs, narrativa, poesia, teatre) pel que fa a la llengua literària, i segons les diferents àrees temàtiques pel que fa a la llengua no literària. Les primeres prospeccions mostraren, però, que amb aquest nombre total de mots —que havia estat fixat en part tenint en compte experiències anàlogues fetes en altres llengües i en part tenint en compte la viabilitat del projecte— la representació dels diferents tipus de text de les diferents èpoques hauria resultat exigua; això féu pensar que calia partir des del principi d'una xifra més alta com a objectiu, i així fou com aquest quedà fixat en cinquanta milions d'ocurrències, mirant d'evitar amb aquesta decisió l'haver de fer modificacions substancials sobre la marxa. A la vegada, i també com a conseqüència de les primeres prospeccions, les previsions inicials sobre la proporció entre text literari i text no literari passaven a ésser aproximadament del 50 % de cada un dels dos tipus, donant així una major representació que la prevista inicialment per a la llengua no literària en relació amb la literària.

El projecte de constitució del corpus textual consta de tres fases fonamentals, que, des del punt de vista de l'execució, se sobreposen parcialment:

1. Selecció dels textos que han de formar part del corpus.
2. Introducció dels textos a l'ordinador i verificació de les dades.
3. Lematització.

Treballs realitzats

La primera i la segona d'aquestes fases començaren simultàniament amb l'inici de l'execució del projecte al principi de 1985 i hom hi treballà amb uns recursos limitats durant els anys 1985, 1986, 1987 i 1988. Durant el darrer d'aquests anys una part dels recursos foren dedicats a dissenyar i a posar a punt el sistema de lematització semiautomatitzada que després s'ha portat a la pràctica; aquesta operació implicà la creació d'un diccionari de màquina que conté 88.000 lemes (entrades de diccionari) i 630.000 formes (que constitueixen les sèries inflectives associades a aquests lemes), així com també l'elaboració dels diversos programes informàtics que permeten de portar a terme la fase de lematització segons el procediment establert.

A partir de l'inici de 1989, un increment important dels recursos destinats al projecte —gràcies a un conveni específic signat entre l'Institut d'Estudis Catalans i la Secretaria d'Estat d'Universitats i Investigació (Ministeri d'Educació i Ciència)— permeté d'accelerar la realització dels treballs, que s'han concentrat a partir d'aquest moment en la part del corpus corresponent a la llengua no literària.

L'evolució de l'extensió del text introduït a l'ordinador i verificat al llarg d'aquests darrers tres anys és la següent (expressada en nombre d'ocurrències):

	Llengua literària	*Llengua no literària*	*Total*
31-XII-88	5.941.279	577.051	6.518.330
31-XII-89	7.316.331	5.862.195	13.178.526
31-XII-90	8.692.489	15.857.927	24.550.416
31-XII-91	9.258.576	27.054.538	36.313.114

L'evolució de la quantitat de text lematitzat des que s'inicià l'execució d'aquesta operació és la següent (expressada també en nombre d'ocurrències):

	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

La selecció corresponent a la llengua no literària és en aquest moment ja acabada; l'extensió de text d'aquesta modalitat que manca introduir és avaluat en aproximadament 2.000.000 d'ocurrències. El total de text no literari del corpus arribarà, doncs, a 29.000.000 d'ocurrències, quantitat superior, encara, a la prevista després de les matisacions a què ens hem referit. La diversitat i l'interès dels textos disponibles ha aconsellat d'arribar com a mínim a aquesta xifra, que podria haver estat superada àmpliament si haguéssim donat una major presència a text de premsa i publicacions periòdiques; consideracions, però, lligades a la viabilitat del projecte han aconsellat d'aturar-nos aquí.

La constitució del corpus té una part complementària que consisteix en la integració en una base de dades comuna de tota la informació incorporada a l'ordinador durant les fases esmentades. En efecte, al llarg de la realització de les fases d'introducció i verificació i de lematització, a què ens acabem de referir, la unitat de tractament coincideix amb l'obra originària, de tal manera que tots els processos que es desenvolupen durant aquestes fases tracten de manera individualitzada els textos introduïts, sense relacionar-los entre ells. Aquest procediment, que ha estat el més convenient durant la fase d'elaboració del corpus, no és, en canvi, el més adequat per al tractament ulterior de les dades, és a dir, per a l'explotació del corpus.

Doncs bé, durant l'any 1990 treballàrem en l'anàlisi i el disseny d'aquesta base de dades conjunta (*Base de dades textual de la llengua catalana*), que ha de permetre l'adequada explotació del corpus de cara a les finalitats previstes. En el procés de creació d'aquesta base de dades hem previst dues fases diferenciades i successives: en primer lloc, la creació del que hem anomenat *Arxiu general de mots lematitzats*, que permet d'obtenir qualsevol tipus d'informació sobre els lemes i sobre les formes del corpus, amb diverses opcions de classificació i d'agrupació, sempre, però, sense context. La incorporació de la informació que permet de reconstruir el context, que és una operació prou complexa, s'ha reservat per a una segona etapa, amb la qual podrem donar per acabada la constitució de la infrastructura lògico-informàtica que permetrà d'abordar la tasca lexicogràfica pròpiament dita a partir de la informació continguda en el corpus textual.

Durant l'any 1991 podem dir que s'ha arribat a executar la primera d'aquestes dues fases: ha estat constituïda una base de dades integrada que conté en aquest moment 19.459.877 ocurrències, que corresponen a 314.308 formes, les quals s'agrupen al voltant de 94.283 lemes diferents. Els textos lematitzats amb posterioritat a la constitució de la base de dades s'hi van incorporant d'una manera progressiva a mesura que s'acaba la lematització de cada obra, i així es continuarà fent fins a la finalització del corpus.

Previsions per al futur immediat

A la vista del progrés dels treballs i del rendiment obtingut durant els darrers tres anys hem pogut fixar uns objectius concrets per al proper trienni:

1) El primer i més obvi d'aquests objectius és la culminació del corpus textual. Des del punt de vista de la selecció manca només portar a terme la corresponent a la llengua literària del període 1833-1913. Pel que fa a la introducció dels textos a l'ordinador i la consegüent verificació, manca un petit romanent de text no literari que no ha pogut ésser introduït durant el darrer any a causa de la notable ampliació que ha experimentat l'extensió del text d'aquesta modalitat (manquen uns 2.000.000 d'ocurrències per a arribar als 29.000.000 previstos definitivament; la part més important que manca introduir, però, és la corresponent a la llengua literària (manquen uns 15.000.000 d'ocurrències per a arribar als 24.000.000 que previsiblement constituiran el total de text literari); així, en resum, podem dir que manca introduir i verificar un total de 17.000.000 d'ocurrències aproximadament, o sigui, un 31,5 % del total del corpus, que, d'acord amb les darreres previsions adaptades a la realitat, arribarà a un extensió total de 53.000.000 d'ocurrències. La fase que requerirà, però, encara més temps i recursos, és la corresponent a la lematització: caldrà lematitzar encara durant aquest període unes 30.400.000 ocurrències, que representen aproximadament el 57,5 % del total del corpus.

Amb la incorporació de tots els mots lematitzats a la base de dades comuna a què ens hem referit quedarà completament acabat el corpus. Arribat aquest moment hom preveu la publicació de diversos resultats de caràcter estadístic en forma de diccionari de freqüències que ha de permetre d'observar la freqüència d'aparició de cada element lèxic i la seva distribució segons diferents criteris (temporal, tipus de llengua, nombre d'obres o altres).

2) D'altra banda hem de considerar la continuació de la constitució de la *Base de dades textual de la llengua catalana* amb l'execució de la segona de les etapes que hem establert, és a dir, la incorporació de la informació que ha de permetre de reconstruir el context.

3) Una vegada constituïda la base de dades textual pròpiament dita, hom iniciarà l'elaboració d'un conjunt de programes específics encaminats a la realització d'una anàlisi semàntica i, eventualment, sintàctica de cada una de les ocurrències del corpus —o d'una selecció aleatòria en el cas dels mots de freqüència elevada— a partir del context en què apareixen. Aquest *software* serà utilitzat en la fase següent del projecte per a l'elaboració del diccionari descriptiu, concretament per a l'estructuració dels articles (classificació dels diferents significats) i exemplificació.

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