RAFEL I FONTANALS, Joaquim [director] (1996-98) *Diccionari de freqüències. Corpus textual informatitzat de la llengua catalana*. Volum 1 (1996), volums 2, 3 i CD Rom (1998), Barcelona: Institut d'Estudis Catalans.

DISTRIBUCIÓ DE L'OBRA

Primer volum. Llengua no literària: Introducció (VII-LIX) signada pel Director de l'obra. Bibliografia (LXI-LXIII). Obres del corpus: Ordenació alfabètica (LXVII-CVIII). Ordenació cronològica (CIX-CLIII). Ordenació alfabètica (1-433). Diccionari: Ordenació per freqüència amb dispersió i ús (435-757). Ordenació per freqüència (freqüències inferiors a 10) (759-787). Ordenació per dispersió (977-1350). Ordenació alfabètica de lemes principals amb els secundaris (1351-1472). Ordenació alfabètica de lemes secundaris amb els principals (1473-1539). Índex (1541).

Segon volum. Llengua literària: Introducció (VII-XIII) signada pel Director de l'obra. Obres del corpus: Ordenació alfabètica (XV-XXX). Ordenació cronològica (XXXI-XLVI). Diccionari: Ordenació alfabètica (1-342). Ordenació per freqüència amb dispersió i ús (343-724). Ordenació per freqüència (freqüències inferiors a 4) (725-850). Ordenació per dispersió (851-1008). Ordenació per ús (1009-1196). Ordenació alfabètica de lemes principals amb els secundaris (1197-1306). Ordenació alfabètica de lemes secundaris amb els principals (1307-1367). Índex (1369).

Tercer volum. Dades globals: Introducció (VII-XIII) signada pel Director de l'obra. Obres del corpus: Ordenació alfabètica (XVII-LXXII). Ordenació cronològica (LXXIII-CXXXI). Diccionari: Ordenació alfabètica (1-540). Ordenació per freqüència (freqüències superiors a 19) (541-850). Ordenació per freqüència (freqüències entre 19 i 3) (851-1018). Ordenació per freqüència (freqüències inferiors a 3) (1019-1200). Ordenació alfabètica de lemes principals amb els secundaris (1201-1376). Ordenació alfabètica de lemes secundaris amb els principals (1377-1476). Índex (1477).

CD-Rom. Disquet i Manual d'utilització del CD-Rom, 37 p.

## 1. PREÀMBUL

Per raons pràctiques, la recensió que segueix no conté cap descripció de l'obra de referència més detallada del que es desprèn de la presentació precedent i es compon d'una selecció de comentaris crítics més o menys generals (des del disseny de la capsa del disquet fins a alguns usos de la noció de "llengua") que sovint pressuposen haver llegit els textos introductoris dels volums i haver utilitzat el CD Rom.

Aquest diccionari de freqüències (d'ara endavant, DdF) conté 52.375.044 mots en total, distribuïts en 29.266.353 de textos no literaris (que fan el 56 % del total) i 23.108.691 de textos literaris (el 44 % restant). Això ha donat lloc a 98.064 entrades o lemes principals i 9.833 lemes secundaris (alternances gramaticals i afixacions, juntament amb vulgarismes i dialectalismes, que no consten estadísticament). Tot el material, aparegut entre 1833 i 1988, està repartit en 23 períodes cronològics de deu anys fins a 1913 i de cinc a continuació (I, xxx-xxxv).

Com que l'obra és molt extensa, la impressió final que en pot causar una recensió depèn crucialment dels aspectes que més hagin atret l'atenció de qui la fa, i aquests, encara, de la seva formació i preferències (en conjunt, limitacions) personals. Això fa inevitable que passin detalls per alt, molts i tot. A més a més, ateses les perspectives metodològiques que aquesta mena d'obra té obertes a la tecnologia, en successives edicions pot canviar en qüestions tant fonamentals com de detall. De fet, tot i que s'afirma que "les possibilitats que [...] ofereix la consulta de la base de dades [...] són pràcticament il·limitades" (I, VII), la versió ressenyada només aprofita encara una petita part de les opcions que, sens dubte, pot oferir —entre elles, segurament, consultes sobre concor-

dances, que convindria fer en línia, a través de connexió amb Internet, com altres diccionaris més o menys similars.[1]

Cal tenir en compte, llavors, que aquesta és només una de les possibles recensions que se'n poden fer.

## 2. Presentació externa

2.1. L'obra, publicada en l'edició ressenyada entre 1996 i 1998, conté, com hem dit, tres volums i un CD-Rom. La manipulació dels volums és inevitablement feixuga, com no pot ser d'altra manera ateses les característiques. Ho seria igualment si se n'hagués desdoblat el nombre. Per això, quan es fa una consulta, especialment les primeres vegades, convé posar a la vista els índexs respectius, no massa fàcils de trobar. Potser se n'hauria d'afavorir la localització posant-los també a les solapes de les portades o al bell principi de volum o en paper colorat.

Si la capsa del CD-Rom dugués al costat oportú una incisió semicircular per on aplicar el dit s'obriria més fàcilment. Dintre hi ha el disquet i un full d'instruccions de 37 pàgines amb indicacions precises sobre la instal·lació i l'ús del programa, però també seria bo que afegís una llista de les moltes abreviatures utilitzades que, a més, fos fàcil de localitzar, apart de l'annex final sobre els codis morfològics. La presentació de les funcions és excel·lent: ordenada, pulcra i ben exemplificada.

El navegador del disc és prou clar, però podria millorar encara en alguns detalls. Li falta l'ajut en pantalla, cosa que obliga a tenir el full d'instruccions a la vora fins que no se'n domina prou el funcionament. Falta també una sortida general, és a dir, la possibilitat de tancar el programa des de qualsevol lloc. Tal com està, cal picar la icona de la porta oberta que duu, successivament, un pas enrere i que serveix tant per a repetir una cerca anàloga com per anar fins al menú principal i només des d'aquí, si es torna a picar, a fora del programa.

Finalment, la impressió es fa sempre directament, sense oferir les opcions habituals de control sobre la impressora, la qual, a més a més, un cop utilitzada, queda configurada tal com la deixa l'última impressió, cosa que n'obliga a restituir la configuració inicial.

L'aprofitament del disc és infinitament més àgil que no pas qualsevol cerca en els volums impresos —un cop, és clar, ens hàgim pres la 'molèstia' d'engegar l'ordinador—. Molt més important que això encara: el disc permet cercar resultats combinats que no són possibles en la consulta als volums.

2.2. Cal precisar, quant a això, una característica particular enfront d'altres obres aparentment similars. En l'Enciclopèdia Britànica, per exemple, l'oportunitat de la recerca en els volums impresos o en el CD Rom dependrà molt d'allò que es vol comprovar. Si es tracta de saber la data de naixement de Leibniz, és molt més efectiu obrir el corresponent volum de la Micropèdia que no pas el disquet, però si es volen aplegar els guanyadors del Nobel de literatura i mirar en quines llengües escrivien, el més assenyat és prendre's la molèstia d'engegar l'ordinador i consultar el CD Rom. I molt més encara si un vol navegar pels continguts sense rumb definit. En una enciclopèdia convencional sol ser fàcil decidir com i on és millor fer una consulta en funció de les moltes classes de requeriments que hom vol resoldre.

He arribat a la conclusió que, en canvi, en el DdF no és ben bé així. Primer de tot perquè, com deia abans, el disquet no sols fa les mateixes coses que els volums sinó moltes altres i, a més a més, molt més de pressa. I després perquè, tot i que per a comprovar quina freqüència absoluta té un mot

1. Precisament mentre redactava aquesta recensió s'estava enllestint la versió en línia que actualment tenim disponible a http:pdl.iec.es/ amb magnífiques prestacions, que tanmateix no coincideixen del tot amb les del CD Rom, amb concordances, però sense col·locacions.

aïllat —posem *estrep*, en la llengua literària— resulta més ràpid obrir el segon volum que no pas l'ordinador, el sistema operatiu, el programa i el disquet, el cert és que no sé trobar cap motiu raonable, que no sigui el de jugar, per a cercar la freqüència absoluta d'*estrep*, ja sigui en la llengua literària o no literària. En altres mots, aquí no veig l'equivalent a la necessitat de cercar una informació concreta i específica, com ara una data de naixement en el diccionari enciclopèdic convencional —tot sigui dit, evidentment, des de l'àmbit específic, i reduït, de les necessitats més habituals dels parlants com jo.

Dues coses aparentment contradictòries se m'ocorren encara arran d'això. Per una banda, que tanmateix no és desitjable prescindir dels volums impresos, encara que no sigui més que per donar suport físic a les introduccions. Les quals són massa importants, en diversos sentits, per a prescindir-ne, i massa llargues per a incloure-les en el fullet d'instruccions del disquet. La segona cosa és que aquestes introduccions no en són, d'importants, almenys des del punt de vista estrictament operatiu —mentre que sí que ho són, en canvi, les instruccions.

## 3. Plantejaments metodològics

### 3.1. *Marc estadístic*

A la vista dels plantejaments adduïts d'obres en alguna mesura semblants, aquesta ofereix pràcticament tots els avantatges possibles fins al punt de convertir-se, sens dubte, en un model. L'anàlisi raonada dels antecedents —tant de nocions bàsiques com de la seva articulació— és un exemple de fins a quin punt s'ha filat prim i s'han ponderat les possibles alternatives. Sempre que es justifica per què cal adoptar un determinat criteri, s'adopta. Almenys per a mi, gens versat en matèries estadístiques, els arguments són, a més d'indiscutibles, engrescadors i tot.

Ara bé, encara que només sigui per fer veure altres línies que es poden adoptar en aquesta mena d'obres, anomenaré tot seguit algunes mancances sense perdre de vista que, tal com està, l'obra és extensa i que afegir-hi res la convertiria en un patracol encara més difícil de manejar —en el text imprès, però no en el disquet—. Pitjor encara: la convertiria en una altra obra. Tenint això ben en compte, hi ha tanmateix un parell d'utilitats que satisfarien força les meves expectatives personals d'ús (i que cito perquè intueixo que molta gent les comparteix): les concordances i les anomenades col·locacions.

Les concordances permeten recuperar el context en què s'ha identificat cada entrada real, cada forma de mot. Això ens deixa col·legir el significat de les entrades sense lematitzar i destriar, per exemple, entre mots homònims i polisèmics, com ara el clàssic *banc*. Expliquem-nos. L'entrada *banc* dóna en el DdF una freqüència absoluta de 3.737 aparicions distribuïda entre formes alternatives del lema: *B.* (abreviatura), *banc*, *banch*, *bank*, *banchs*, *bancs*, *banquet*, *banquét*, *banquillet*, *bancalons*, *banquets* i *banquillos*. Presumiblement, per no dir segurament, aquí *banquet* només es refereix a un 'banc petit' i no a un menjar amb molts comensals —aquest sentit, amb el lema *banquet*, presenta, al seu torn, 381 aparicions pròpies—, però, tot i així, queda en peu establir, quan diu *banc*, a quina mena de 'banc' es refereix (banca, banc del parc, banc de fuster, banc de dades, banc de sardines...). ¿Què fer, encara, si un autor juga amb aquest doble sentit, com a mínim, de *banquet*? Més avall analitzarem una mica de prop uns exemples anàlegs.

El mal és que les concordances obliguen a introduir llistats amb totes les formes de mot, sense lematitzar, i això comportaria una ampliació molt important de l'abast actual de l'obra.

Lligades amb les concordances, les col·locacions són manifestacions tendencials dels mots a comparèixer juntament amb d'altres en concret per a formar significats més o menys unitaris. Sovint, aquesta tendència es consolida fins a convertir-se en una locució lèxica (*mà d'obra*, *medi ambient*...); d'altres, dóna lloc com a mínim a una construcció freqüent (*incendi forestal*, *cavall salvatge*...) enfront d'altres que pràcticament no es documenten mai (*incendi geològic*, *cavall*

*precoç*...). I, contra el que passaria amb les concordances, sembla que establir les col·locacions hauria de comportar un esforç i un espai relativament modest —ja que només caldria llistar els aparellaments de mots (fins i tot grups de tres, quan hi ha nexes preposicionals) que traspassessin un cert llindar freqüencial de coaparició— si no fos perquè les col·locacions pressuposen... les concordances. Tal com està ara, el DdF ens permet cercar, per exemple, *abans-d'ahir*, però no *demà passat* ni qualsevol expressió que estigui formada per dos o més mots gràfics separats. Això elimina de pas les formes verbals compostes, entre moltes altres coses.

Indubtablement, la utilitat de les col·locacions és immensa, sobretot en aplicacions lexicogràfiques i en l'anomenat processament del llenguatge natural, entre altres aplicacions informàtiques més trivials de la llengua. Imaginem l'ajut que suposa disposar de col·locacions per a confegir subentrades i accepcions o, si més no, exemples d'ús ben reals. Les col·locacions representen un nexe natural entre els significats lèxic i oracional, dos àmbits encara massa diferenciats per la dimensió de la sintaxi en moltes teories gramaticals.

### 3.2. *Marc teòric*

3.2.1. Hi ha un seguit de qüestions teòriques que em semblen si més no opinables. Sovint sembla que els especialistes en confegir obres 'modernes' que actuen en dominis tecnològics, sobretot computacionals, exhibeixen un desconeixement flagrant de certs principis teòrics, fins i tot elementals, que eren no sols indiscutibles, sinó, encara més, a l'abast de tothom qui tingués una formació mitjana en lingüística. Aquesta ignorància sembla a vegades fins i tot forçada, com si aquells especialistes es donessin de menys de mantenir una connexió amb la teoria tradicional per no propiciar una imatge obsoleta. Em consta molt bé que no és pas el cas d'en Joaquim Rafel.[2] Tanmateix trobo alguns aspectes opinables, com dic.

El que més en crida l'atenció és la insistència amb què s'anomena "llengua" el que no és més que un corpus, un recull de textos escrits (i alguns de parlats convertits en escrits). La "llengua", en el sentit en què els lingüistes l'han vingut utilitzant de Saussure ençà, és una entitat virtual que els parlants posseeixen intuïtivament i que serveix com a punt de referència perquè s'entenguin els uns amb els altres quan en fan ús.[3] Seguint una terminologia igualment convencional, la "parla" és la capacitat d'aplicar les virtualitats de la "llengua" per a satisfer en cada cas les necessitats particulars de comunicació. I quan aquesta capacitat s'aplica, amb l'activitat de parlar o d'escriure, es produeix un fet de parla més o menys complex, llarg, intel·ligible, encertat...

El DdF és essencialment un estudi estadístic sobre un corpus, una extensa col·lecció de fets de parla enregistrats al llarg d'un període de temps i classificats d'acord amb certs criteris. Des del punt de vista estadístic es pot discutir si la col·lecció és o no una mostra prou representativa dels fets de parla habituals, però mai no es pot insinuar ni donar a entendre que ho és de la "llengua". Des del punt de vista estricte de la "llengua", tenen exactament la mateixa virtualitat les formes verbals *fem* (del verb 'fer') i *estrinxolem* (del verb 'estrinxolar'), tot i que en el corpus la primera surt un bon grapat de vegades (més de 3.000 comptant les formes alternatives) i la segona no cap.

Això ens duu a l'afirmació, igualment repetida, que els mots tenen valors desiguals en funció de la seva freqüència d'ús i que, per això, uns són més importants que altres (I, VII): "[La freqüència] és tinguda generalment [...] com l'indici més clar del valor dels mots, i és sovint rela-

2. No he sabut esbrinar quina ha estat la contribució dels col·laboradors que figuren a l'obra. En qualsevol cas, les introduccions van signades només pel director.

3. La confusió es produeix perquè Saussure va utilitzar el terme 'llengua' (en francès 'langue', en oposició a 'parole', que traduïm per 'parla') en el sentit tècnic que dic, però la mateixa noció, o semblant, rep també altres solucions terminològiques —com ara la coneguda de 'competència' i 'performança' a Noam Chomsky.

cionada amb els conceptes d'utilitat i d'importància, de tal manera que hom parla de mots més importants i menys importants i de mots més útils o menys útils d'acord amb els valors freqüencials més alts o més baixos" (I, IX). Cal matisar-ho, si més no. Perquè si per valuosos entenem els mots que s'utilitzen amb tanta freqüència que surten pràcticament en tots els textos, un didacta de la llengua entendrà que els mots en qüestió han de ser ensenyats amb preferència i de bon començament, però un tècnic en informació dirà que aquests són precisament els mots menys comunicatius —típicament, els anomenats 'mots útils', mots funcionals, com articles, preposicions i similars—, aquells que, en l'època dels telegrames tothom, entès o no en informació, eliminava sempre que podia. Tot això obre —em sembla— la necessitat de matisar la noció d'importància estadística dels mots incorporant-hi altres nocions, com la de 'disponibilitat' tal com l'utilitza G. Gougenheim (manca relativa d'alternatives en certes situacions i temes) i la de 'vocabulari fonamental', entre altres.

3.2.2. També trobo apriorístic afirmar que els parlants tenen una visió errònia sobre la freqüència dels mots (I, XII). En primer lloc, cal acceptar que els resultats de qualsevol DdF estan plens de sorpreses per a qualsevol parlant, per ben dotat que tingui el sentit de la llengua, i, certament, tothom està exposat a cometre errors si pretén ordenar per freqüències una llista qualsevol de mots. Però fins un cert punt: si la llista es compon, per exemple, de tres substantius relatius al cos humà com ara *duodè*, *turmell* i *mà*, és molt probable que hi hagi un assentiment general a ordenar-los a l'inrevés de com estan —cosa que el DdF confirma plenament: *mà* 37.984, *turmell* 273, *duodè* 30—. De la mateixa manera tots els parlants, almenys els 'normals' o més estàndards, fan una selecció instintiva no sols dels mots més freqüents quan s'adrecen, per exemple, a una criatura petita o a un estranger amb pocs coneixements de la llengua, sinó dels més adequats a cada requeriment comunicatiu, en un sentit massa extens dels termes 'normal' i 'adequat' per entretenir-nos-hi ara. Sostinc, tot plegat, com molts altres lingüistes, que una noció de freqüència relativa dels mots forma part de les intuïcions espontànies i, per descomptat, bàsiques dels parlants. Això sí, prou bàsiques perquè mai no deixi de tenir utilitat un bon DdF, com el que comentem.

A propòsit d'això m'ha cridat l'atenció un detall presumptament contradictori. Quan es critica l'obra *Dictionnaire du fréquence du catalan* d'E. Guiter (I, xxv), se li retreu, entre altres defectes, que no s'adonés de la freqüència més elevada que té *diumenge* entre les denominacions dels dies de la setmana. Trobo que el diumenge potser ofereix algun atractiu superior, però només a posteriori, un cop comprovat al DdF (set vegades i escreix més que no pas el dimarts!), i no tant per intuïció. Perquè, llavors, no trobo cap justificació per a la freqüència relativa de la resta de dies ni perquè, entre les denominacions del mesos, no és el desembre, el més festiu, sinó el maig, amb el setembre, el d'ús més freqüent. Examinem els resultats del DdF en sengles quadres:

DIES DE LA SETMANA

| | *Freq. abs.* | *Ordre* |
|---|---|---|
| Dilluns | 1.051 | 4 |
| Dimarts | 630 | 7 |
| Dimecres | 655 | 6 |
| Dijous | 1.045 | 5 |
| Divendres | 1.061 | 3 |
| Dissabte | 2.937 | 2 |
| Diumenge | 4.697 | 1 |

MESOS DE L'ANY

| | *Freq. abs.* | *Ordre* |
|---|---|---|
| Gener | 3.112 | 9 |
| Febrer | 2.536 | 12 |
| Març | 3.148 | 7 |
| Abril | 3.518 | 3 |
| Maig | 3.821 | 1 |
| Juny | 3.226 | 6 |
| Juliol | 3.310 | 5 |
| Agost | 3.122 | 8 |
| Setembre | 3.707 | 2 |
| Octubre | 3.403 | 4 |
| Novembre | 2.557 | 11 |
| Desembre | 2.603 | 10 |

En general dono per entès que qualsevol contradicció entre la intuïció i el DdF no fa sinó posar de relleu la prioritat del DdF.

## 4. EL CORPUS

4.1. La caracterització de la parla en textos diferenciats, sovint anomenada tipologia textual, ha adquirit fa temps un gran interès per les importants aplicacions pràctiques que ofereix. En aquest DdF la selecció dels textos que componen el corpus ha rebut una atenció preferent gràcies a l'aplicació de diversos criteris que han dut a un conjunt considerat "equilibrat" des de tots els punts de vista.[4] El primer criteri ha consistit en la diferenciació temàtica entre llengua literària i no literària i, dintre de cada classe, en una distribució per èpoques i per temes. Encara que aquella primera divisió sembla, si més no, prou ponderada, el conjunt resultant de divisions no pot ser precisament equilibrat, per molts subcriteris i retocs que s'introdueixin, especialment des del punt de vista temàtic. Més aviat al contrari, com més divisions s'introdueixen més heterogeneïtat es forma, sobretot a partir d'un cert punt. I és que no hi pot haver cap homogeneïtat temàtica, posem només per cas, entre els grups 4) de premsa; 7) de belles arts, oci, esports; 8) de llengua i literatura, i 0) de correspondència. Si més no, aquell grup setè —de belles arts, oci i esports— ens pot ajuntar un tractat crític sobre el color a Kandinski, una partida de parxís i la legislació pel control del dopatge, cosa perfectament lògica que succeeixi també en la correspondència i encara més en la premsa, on solen concórrer aquests temes i molts altres, per no dir tots els altres. Al meu parer, convindria haver fet veure que s'és conscient dels problemes més o menys insolubles que planteja la tipologia textual —com a caracterització, insisteixo, de la parla, no de la llengua.

4.2. El problema és més transcendent del que sembla, perquè la tipologia textual hauria de permetre fer, a través del que aquí s'anomena dispersió, una adscripció semàntica, per bé que molt general, dels mots —probablement l'única aproximació explícita del DdF al tractament del significat—. Vegem-ho amb un exemple.

4. Fins i tot quantitatiu. En alguns casos, s'han recollit textos en forma només fragmentària per no desequilibrar la representativitat relativa global —cosa que comporta, es vulgui o no, una presumpció sobre els resultats.

Suposem que volem comparar les freqüències d'entrades com *alfil* i *torre*. El nostre coneixement de la llengua ens diu que totes dues comparteixen un mateix àmbit significatiu com a peces en el joc dels escacs; però mentre *alfil* no té cap altra aplicació (diccionari en mà), *torre* pot significar una bona sèrie de coses, a part d'una peça del joc. El Diccionari de l'IEC assenyala per a *torre* tres significats generals, alguns amb diverses accepcions i col·locacions (*t. de babel*, *t. de guaita*, *t. de control*...) al marge d'usos tròpics oberts a una abundant polisèmia (*una t. d'home*...). L'àmbit de la llengua no literària dóna com a freqüències absolutes 6 per a *alfil* i 1.983 per a *torre*. A més, *alfil* presenta com a formes de lema, *A* (abreviatura), *alfil*, *alfils*, *àlfil*, mentre que *torre* ofereix *T* (abreviatura), *torra*, *torre*, *torras*, *torres*, *torrassa*, *torreta*, *torrassas*, *torrasses*, *torretas* i *torretes*.[5] L'àmbit de la llengua literària dóna com a freqüències absolutes 6 per a *alfil* i 1.763 per a *torre* (totes dues amb les mateixes formes de lema). La consulta en la llengua literària i no literària suma lògicament les freqüències absolutes anteriors: 12 per a *alfil* i 3.746 per a *torre*. Això ens fa pensar que l'àmbit significatiu del joc d'escacs ha quedat repartit o, millor dit, entaforat sense caracterització entre els tipus, temes i gèneres textuals. A falta d'especificacions de significat, sembla que l'única manera de constatar això serà apel·lant a la noció de dispersió, definida com expressió de "la repartició dels mots en diversos tipus de text establerts d'una manera coherent" (I, IX). D'acord amb la dispersió, *alfil* ha de quedar molt més concentrat que *torre* en uns tipus molt limitats de text. Aquesta comprovació s'ha de fer per cada àmbit textual, literari i no literari:

| *Textos Literaris* | *Freqüència absoluta* | *Freqüència relativa* | *Dispersió* | *Ús* |
|---|---|---|---|---|
| Alfil | 6 | 0,000026 | 0,71506544 | 4,29 |
| Torre | 1763 | 0,007780 | 0,82730488 | 1458,54 |

| *Textos no literaris* | *Freqüència absoluta* | *Freqüència relativa* | *Dispersió* | *Ús* |
|---|---|---|---|---|
| Alfil | 6 | 0,000021 | 0,28899784 | 1,73 |
| Torre | 1983 | 0,006935 | 0,7593590 | 1497,04 |

A la dispersió, el valor 1 indica distribució uniforme, i baixa com menys uniforme és. Això significa que *alfil* és moltíssim més uniforme en textos literaris que no pas no literaris, mentre que *torre* ho és similarment. Per altra banda, l'ús, com més igual és a la freqüència absoluta, més equilibrat (o igualitari en els subcorpus) és. El desequilibri el marquen els valors inferiors. Els resultats demostren de nou que *alfil* té un ús molt més equilibrat en textos literaris que no pas en no literaris, mentre que *torre* el té similar en ambdós casos. I no és estrany perquè, de fet, la freqüència absoluta, l'ús i la dispersió són valors completament solidaris, atès que la dispersió és el quocient de dividir l'ús per la freqüència absoluta, i l'ús, lògicament, el resultat de multiplicar la freqüència absoluta per la dispersió.

Sigui com sigui, de les múltiples interpretacions d'això sembla que n'hi ha almenys una de contraintuïtiva, tenint en compte que esperàvem que *alfil* aparegués només en alguns tipus de text, en concret aquells que fessin alguna referència al joc dels escacs, i tingués, per tant, una dispersió molt baixa, cosa que només s'acompleix en textos no literaris, on més caldria esperar referències als escacs. No està clar, per tant, que a partir d'aquí puguem deduir res pel que fa als significats.

5. En molts indrets la *torreta* és un test per a plantar flors, a més d'una torre petita.

En efecte, si, per exemple, extraiem la llista de les dotze dispersions més altes de tot el corpus, almenys jo no trobo quina deducció conjunta es pot concloure sobre *encertar, postura, plet, embocadura, perseguir*...

| *Lema* | *F. absoluta* | *F. relativa* | *Dispersió* | *Ús* |
|---|---|---|---|---|
| encertar | 638 | 0.002815 | 0.99360182 | 633.92 |
| postura | 198 | 0.000874 | 0.98343000 | 194.72 |
| plet | 308 | 0.001359 | 0.98216880 | 302.51 |
| embocadura | 25 | 0.000110 | 0.98100577 | 24.53 |
| perseguir | 1200 | 0.005296 | 0.98075767 | 1176.91 |
| disfrutar | 299 | 0.001319 | 0.98064687 | 293.21 |
| tot | 111984 | 0.494181 | 0.98032827 | 109781.08 |
| vanitós | 206 | 0.000909 | 0.97909933 | 201.69 |
| temor | 1137 | 0.005018 | 0.97792390 | 1111.90 |
| commoure | 607 | 0.002679 | 0.97742215 | 593.30 |
| sever | 646 | 0.002851 | 0.97688475 | 631.07 |
| amulet | 35 | 0.000154 | 0.97644864 | 34.18 |

4.3. Hi ha encara un assumpte que presenta un interès afegit i més general. Es tracta de l'enorme casuística morfològica que presenten moltes grafies degut al tram cronològic que comprèn el corpus, des de 1833 fins a 1988, cosa que en temps representa una mica més d'un 50 % anterior a les Normes de 1912. Només per a l'infinitiu que avui escrivim com *fer* es comptabilitzen tretze variants (*fé, fé', fè, fê, feer, fér, fêr, fere* i *fert*, entre altres) suscitades sobretot per fidelitats fonètiques, dialectals i històriques a vegades barrejades amb especulacions etimològiques. En aquests casos se'n consigna la freqüència absoluta, però no queda, en canvi, constància de la distribució ni cronològica ni textual —tot i que seria, sens dubte, un aspecte força interessant de saber—. Aquesta casuística ens fa pensar en les errates ortogràfiques. I encara més en els mals usos, malapropismes i desencerts en general, tan freqüents a la pràctica. I encara, pel costat contrari, en els usos figurats i els tropismes com les clàssiques sinècdoques, metonímies i metàfores. I les ironies, que en general volen dir el contrari del que diuen. De tot això no en queda cap constància, perquè, com hem dit, aquest DdF no consigna els significats.

## 5. Usos i aplicacions possibles

Una altra qüestió general que m'agradaria veure més precisada al DdF és la presumpta utilitat de l'obra anomenant amb un cert detall algunes possibles aplicacions, és a dir, amb alguna menció del mètode, encara que fos molt en general, i de les fites científiques i tecnològiques que es poden cobrir. En diverses ocasions es diu que les aplicacions són múltiples (I, VIII); a part de la recerca lingüística "pròpiament dita", la lexicografia, l'experimentació psicolingüística, la didàctica de les llengües, el processament del llenguatge natural, etcètera. És indiscutible que un DdF així ofereix grans prestacions per a la lexicografia i el processament del llenguatge natural, però en aquest últim cas, al costat d'aspectes força útils (per a la lematització i per a confegir estratègies estocàstiques i tipologies textuals), n'hi falten d'altres que ho podrien ser molt (col·locacions, etiquetatge morfosintàctic). El mateix passa amb la didàctica de les llengües, almenys en l'aprofitament dels principals valors freqüencials, i segurament amb la psicolingüística en aspectes que se m'escapen. Tot i que no es diu en forma explícita, em sembla que una de les millors perspectives d'aplicació del DdF es troba en la caracterització lexicomètrica de textos en qualsevol moda-

litat.[6] No hi ha dubte que la previsió d'aplicacions té una importància cabdal que pot —que sol— fins i tot condicionar l'estructura general d'una obra així.

No crec pecar d'excés de suspicàcia si dono per suposat que la referència a la recerca lingüística "pròpiament dita" forma part de la mateixa confusió a què em referia més amunt sobre la noció de 'llengua'. Almenys el sentit més lògic que es pot donar a aquestes aplicacions és que un DdF (ben fet, com aquest) permet confegir no sols diccionaris, sinó també gramàtiques. Pocs temes han estat més debatuts que aquest a l'entorn de l'explotació computacional del lèxic. La idea sosté que un bon DdF reflecteix l'ús lingüístic dels parlants amb una abundància i una fidelitat úniques, superiors a les de qualsevol altre mètode conegut, i això fa que sigui la font ideal per a extreure diccionaris i gramàtiques basades en la realitat de la llengua. Com veieu, la idea subjacent és que el DdF és una representació de la llengua.

Més encara, cap lingüista, ni parlant assenyat, no dubtarà a creure que els 'textos' més representatius dels usos lingüístics són els orals, no sols pel nombre, sinó també per l'espontaneïtat i, fins i tot, per la 'naturalitat' de la seva manifestació. El mal és que tots els corpus, i en especial els orals, presenten un seguit d'inconvenients, començant perquè aquests últims són molt difícils d'enregistrar i aplegar. Un altre de pitjor és identificar-ne els paràmetres i interpretar-los, ja que, si es vol reunir un corpus representatiu de la parla, caldrà introduir variables fonètiques, com ara de registre, modulació de la veu i d'altres que incrementaran de molt els requeriments de la representativitat. Però hi ha un altre inconvenient encara, el pitjor de tots. I és que si les mostres recollides són realment espontànies, moltes d'elles contindran tantes irregularitats gramaticals que mai es podran considerar que siguin model de cap 'llengua'.[7] La immensa majoria de fets de parla espontània contenen rèpliques fragmentàries, oracions encavalcades unes amb altres, canvis continus d'estratègia expressiva, mancances de memòria, correlacions amb fenòmens paralingüístics... que desvirtuen els models abstractes amb què treballa la gramàtica, que és abstracta per definició. Altrament dit, d'un corpus així pot sortir, per exemple, una base de dades per a la identificació de veu i segurament molts altres productes, però no, com dic, una gramàtica. Per això, malgrat els meus esforços, també trobo inevitablement dubtosa l'aplicació del DdF a la lingüística 'en general', si més no en els apartats més clàssics de la lingüística.

## 6. Conclusió

Més amunt (2.2) he tingut la gosadia de dir que les introduccions teòriques al DdF no tenen importància operativa, ja que el diccionari funciona amb independència d'elles. Ara és el moment de reconèixer que, ben mirat, els meus comentaris fan referència gairebé exclusiva a aquestes introduccions; per tant, també accepto que els comentaris són, potser no ben bé frívols, però sí poc o molt tangencials, i que la meva recensió deixa bastant intacta l'anàlisi profunda del diccionari com a eina per extreure informació estadística d'un corpus. Això passa pel que també deia al preàmbul: cadascú guaita des de la seva perspectiva.

6. En aquest punt no em puc estar de citar una remarcable tesi de doctorat titulada *Tractament de corpus textuals lematitzats i estudi comparatiu del llenguatge científic amb la prosa estàndard*, on vaig compartir tribunal, a la Universitat Politècnica de Catalunya l'any 1994, amb en Joaquim Rafel. En ella l'autora, Anna Puig Montada, caracteritza el català —la llengua— per un interval de l'entropia obtingut a partir de quatre corpus i equivalent a I = (8,6 – 2. 0,3, 8,6 + 2. 0,3) = (8,0 9,2). En altres paraules, si calculem l'entropia d'un text i trobem que està entre 8,0 i 9,2, podem assegurar que el text està escrit en català.

7. L'expressió oral de textos escrits, com ara la declamació, el teatre, l'emissió d'oracions, fòrmules rituals, de discurs repetit (refranys...), etc. correspon pròpiament als textos escrits, i no als orals, atès que els textos en qüestió han estat concebuts i elaborats com a textos escrits. I a l'inrevés, la transcripció fonètica d'una expressió oral correspon als textos orals, encara que es manifesti per escrit. És la coneguda diferència entre concepció i elaboració d'un text, que distingeix entre el llenguatge oral i el llenguatge escrit, i el seu canal de transmissió.

No crec equivocar-me massa en insistir que les aplicacions més previsibles d'un DdF així són al domini cada cop més extens de la lexicometria, on es pot treure el profit més natural de les enormes possibilitats estadístiques que ofereixen les consultes complexes del CD Rom.

Ramon Cerdà
Universitat de Barcelona