Joaquim Rafel i Fontanals, *Diccionari de freqüències. 1 Llengua no literària. (Diccionari del catalá contemporani. Corpus textual informatitzat de la llengua catalana)*, Institut d'Estudis Catalans, Barcelona, 1996, 1539 p.

Con data de decembro do pasado ano vén de ve-la luz o *Diccionari de freqüències*, concretamente o primeiro volume dedicado á *Llengua no literària*, dirixido na súa elaboración por Joaquim Rafel.

Este tipo de obras pretenden ofrecer información cuantitativa sobre o léxico co fin de configura-los valores que nunha lingua teñen as unidades léxicas, partindo do cómputo das súas aparicións en enunciados lingüísticos reais. Xunto a estes estudios atopamos, en paralelo, os chamados "vocabularios básicos ou fundamentais", que tamén teñen como pretensión a delimitación das palabras máis importantes dunha lingua, xeralmente cara a un enfoque didáctico; pero os obxectivos entrámbolos dous son ben diferentes, posto que os diccionarios de frecuencias teñen a pretensión de achegar datos represen-tativos (dentro das posibilidades de cada obra) do uso dos elementos léxicos que fan os falantes na súa actividade comunicativa, utilizando, para iso, métodos estatísticos aplicados nunha análise cuantitativa do conxunto de textos considerados para tal fin, os vocabularios básicos só buscan a elaboración dunha lista de palabras tidas como das máis útiles ou necesarias para a comunicación en circunstancias moi concretas

A obra que aquí comentamos, é o primeiro paso visible, á luz pública, dun elaborado proxecto concibido pola "Secció Filológica del Institut d'Estudis Catalans", co fin da realización dun diccionario descritivo da lingua catalana dos últimos cento cincuenta anos, o *Diccionari del Catalá Contemporani* (*DCC*). Este proxecto vénse movendo na liña dos máis modernos métodos lexicográficos que xurdiron para palia-lo carácter pouco sistemático dos traba-llos lexicográficos tradicionais, os cales basean as súas técnicas na acumulación de datos tomados doutros diccionarios publicados con anterioridade. Isto vese solucionado coa utilización de fontes achegadas a partir de datos da frecuencia de uso do léxico, pasándose, así, do emprego de fontes subxectivas e intuitivas á obxectivización metodolóxica do traballo, onde teñen grande importancia destes datos obxectivos tirados de corpus textuais. Para que estes corpus poidan cumprir axeitadamente a función que deles se require, deben reunir certas características, as cales se poden resumir nunha: a representatividade da realidade lingüística que se quere describir, tanto cronolóxica como tipoloxicamente.

Para lograr este propósito concibiuse, alá polo 1984, a elaboración dun corpus textual, o *CTILC* (*Corpus textual informatitzat de la llengua catalana*), que comeza a elaborarse despois de catro anos nos que se establecen as bases metodolóxicas do traballo e se crea a infraestructura material necesaria para levalo adiante. Este corpus abrangue, temporalmente falando, desde 1883, data

simbólica da recuperación do catalán contemporáneo, ata 1988. Tipoloxica-mente recóllese a lingua escrita, contando con textos de carácter *literario* e *non literario*, dividíndoos en grupos que permitisen unha ponderada selección dos textos que se debían ter en conta para cada tipo (literario / non literario) e subtipo (ensaio, narrativa poesía e teatro no subcorpus literario, e correspon-dencia, filosofía, relixión e teoloxía, ciencias sociais, prensa, ciencias puras e naturais, ciencias aplicadas, belas artes, ocio e deportes, historia e xeografía para o non literario).

Excepto as cartas persoais e os textos notariais, as demais son obras publicadas en calquera soporte material, tendo en conta a primeira edición, sen correccións nin modificación respecto do orixinal. Isto faise buscando a máxi-ma representatividade, para o que tamén se fixeron certos cortes cronolóxicos para a selección dos textos, en grupos de dez anos, ata 1913, e de cinco anos de 1913 a 1988, contándose vintecinco grupos cronolóxicos nos que se tentou mante-la representatividade de cada un dos subtipos establecidos. Deste xeito, atopamos co primeiro feito deste traballo, que é a publicación deste volume dedicado ó subcorpus da lingua non literaria[1].

Unha vez seleccionado o corpus, pásase á súa informatización, lematizándoo e creando a *Base de dades textuals de la llengua catalana* (*BDTLC*), a partir da que se pode considera-lo léxico desde un punto de vista cuantitativo, analizando os índices dos valores das palabras. As unidades léxicas vense, pois, clasificadas segundo criterios cuantitativos referentes á súa maior ou menor posibilidade de aparición, para o que se recorre a termos estatísticos, como o de *frecuencia* dunha unidade léxica, ou número de veces que aparece nun texto ou nun corpus calquera determinado e delimitado. Non se adoita falar da frecuencia en termos absolutos, posto que hai que poñela en relación co conxunto de tódalas frecuencias das palabras analizadas e a extensión do corpus, tirándose así *frecuencia relativa* de cada unha delas. Da análise desta magnitude obxectiva dedúcese que a frecuencia das unidades léxicas non está repartida do mesmo xeito entre os textos compoñentes do corpus, así, xorden os conceptos de *repartición* e *dispersión*, usados en lexicoloxía cuantitativa. Chámaselle *dispersión simple*[2] a un xeito de ter en conta as diferencias de distribución das palabras nun corpus á hora de establecer unha xerarquía do léxico, e se se considera esta distribución nos grupos nos que aparece en función do reparto da frecuencia absoluta, témo-la chamada *dispersión complexa*[3], reductible a un índice numé-

[1]No presente momento debe estar a piques de remata-lo traballo para o subcorpus da lingua literaria, de térense acadadas as previsións do director do proxecto, tal e como foron expostas no seu traballo "Diccionarios y corpus textuales. Perspectivas para el catalán", publicadas en *Actas do Simposio de Lexicografía actual: elaboración de diccionarios*, organizado pola Real Academia Galega e o Centro de Investigacións Lingüísticas e Literarias "Ramón Piñeiro", Constantino García, Isabel González Fernández e Manuel González González (editores), *Cadernos de Lingua*, Anexo 3, A Coruña, RAG, 1995, páx. 157-196.

[2]Juilland, A. e Chang-Rodríguez, E., *Frequency dictionary of Spanish words,* The Romance Languages and their Structures, First Series, The Hague, 1964, páx. XLV e ss.

[3]Tamén teorizada por Juilland. Para ter unha idea máis precisa da teoría empregada, consúltese a este respecto a bibliografía que se achega nas páxinas LXI a

rico chamado *índice de dispersión*. Partindo da repartición da frecuencia de cada palabra, e combinándoa coa frecuencia total, podemos chegar a un novo concepto, o *uso*, que integra as diferencias de repartición.

No presente traballo aplícase o procedemento de Juilland[4] coas modificacións oportunas enfocadas a un corpus como o *CTILC*, na procura dunha clasificación obxectiva nos datos de xeito que permitisen a súa compa-ración, se fose preciso, con outros traballos sobre outras linguas románicas nos que se teña empregado este sistema[5].

Unha das características do *CTILC* é que se trata dun corpus lematizado, cun alto grao de funcionalidade, no que se levou a cabo unha minuciosa operación de análise lingüística de cada unidade léxica, que foi categorizada gramaticalmente en cada unha das ocorrencias de cada forma gráfica coa que se asocia, co que se cumpren dous obxectivos: evita-la ambi-güidade entre determinados casos, e relacionar formas dunha serie flexiva, as cales quedan asociadas a unha forma de referencia chamada *lema*. Este lema consta duns caracteres que constitúen a súa grafía, e dun código gramatical marcado cun máximo de tres caracteres. Podemos atopar elementos homó-grafos que se distinguen, nas listas que se achegan polo seu código gramatical. Estes códigos foron tomados das fontes lexicográficas xerais que se empregaron como referencia.

No volume inclúense tódolos termos analizados no corpus, normativos ou non, desde os que teñen unha frecuencia máis elevada ata os que só aparecen unha soa vez, pero no volume impreso (posto que ademais se acompaña dun CD-ROM) non se atopan algunhas informacións que si están nos textos e na base de datos do *CTILC*, tales como termos de taxonomía, nomes químicos con símbolos, signos, sufixos, prefixos e infixos, nomes propios, expresións numéricas, etc., que se marcan co código *nc* (non codificado), pero ós que si se pode acceder na consulta en soporte magnético. Estes veñen ser un total de 107.897 lemas (98.064 ocorrencias principais e 9.833 secundarias). Debido ó proceso lematizador, á vez que temos agrupadas formas da mesma serie flexiva, ocorre o mesmo coas da series homógrafas, así que xunto a ocorrencias da forma gráfica ***portes***, verbo, aparecen formas gráficas de ***porta***, substantivo feminino (***portes***, en plural). Para remediar moitos destes problemas tívose que elaborar un "ma-nual de lematización" para o persoal que fixo o proceso de lematizador. Xunto a cada lema están asociadas formas flexivas e derivadas (diminutivos, aumen-tativos, despectivos, etc.), xunto coas súas variantes grafías recollidas dos textos. Estas formas están codificadas baixo o código *met*, metalingüisticamente. A suma do número de formas referida a cada lema pode ser moi elevada en moitos casos, polo que só se achegan informacións cuantitativas referentes ós lemas principais, mentres que as dos secundarios danse en listas á parte, pero podén-dose consultar por xunto. Esta separación entre lemas principais e secundarios faise co fin de podelos tratar independentemente con cada un dos seus propios datos

---

LXIII da obra que comentamos, concretamente a páxina LXII posúe as referencias ás obras deste autor.

[4]Véxase a nota anterior.

[5]Para unha información máis detallada sobre os procedementos estatísticos empregados neste traballo, véxanse as páxinas XLV a XLIX do prólogo.

(frecuencias, formas asociadas, etc.), posto que se poden dar estructuras moi complicadas, coma nos casos dos verbos con lemas principais e secundarios e con variantes estructurais e gráficas para cada un deles.

Como se pode deducir polo dito ata o de agora, este *Diccionari de freqüències* conta con dúas partes, ou mellor dito, ofrécesenos en dous formatos diferentes de consulta: un impreso, o volume publicado, e outro en soporte magnético, concretamente en CD-ROM. Ademais, dáse a circunstancia de que tódolos datos do *CTILC* poden ser consultados a distancia por vía informática, poñéndose en contacto co Institut d'Estudis Catalans.

A parte impresa consta dun volume que, como xa vimos, está dedicado á lingua non literaria, abríndose cunha introducción pormenorizada e clara realizada polo director do proxecto, Joaquim Rafel, onde se explica o proceso de elaboración, fins, metodoloxía, bases teóricas e materiais empregados (páx. VII-LIX), á que segue a bibliografía básica empregada (páx. LXI-LXIII) e dúas listas das obras que configuran o corpus textual, unha por orde alfabética (páx. LXVII-CVIII) e outra cronolóxica (páx. CIX- CLIII).

Despois desta parte introductoria está o *diccionari* propiamente dito, estructurado en sete seccións que achegan os tatos tirados da análise do corpus.

- *Ordenació alfabètica* (páx. 1-433) dos lemas principais do subcorpus non literario, acompañados da frecuencia acumulada dos lemas principais e secundarios, sendo un total de 98.064, concretados en 28.554.142 ocorrencias. Esta lista ofrece a información básica para acceder ós datos, isto é, se a ocorrencia que se pretende investigar está ou non no corpus, e cal é a súa frecuencia absoluta. Cada lema grafiado vai seguido do seu código gramatical (dos que se dá unha relación na páxina XXXIV), o código de procedencia e a frecuencia absoluta.

- *Ordenació per freqüència*, que clasifica os mesmos lemas en orde decrecente de frecuencia (aparecendo distribuídos alfabeticamente aqueles que teñan a mesma), dividíndose en dúas listas complementarias, unha inclúe os lemas cunha frecuencia igual ou superior a dez (33.993), presentando tódolos seus datos cuantitativos, isto é, frecuencia absoluta e relativa, índice de dispersión e uso (páx. 435-757), e a outra sección (páx. 759-975) inclúe só os lemas con frecuencia inferior a dez (64.071), dividíndose esta en subseccións, unha para cada frecuencia diferente, indicadas a frecuencia absoluta e relativa só ó comezo de cada unha delas.

- Segue unha *ordenació per dispersió* (páx. 977-1151) dos lemas en función do índice de dispersión, marcándose, ademais, a frecuencia absoluta e o uso. Nesta lista non se indican aqueles lemas que teñen un índice de dispersión igual ou superior a 0'5, que serían 26.191.

- *Ordenació per ús* (páx. 1153-1350). Preséntanseno-los datos referentes ó uso, á frecuencia absoluta e ó índice de dispersión, sempre e cando o seu uso sexa igual ou superior a cinco, o que sumarían 29.703 lemas.

- Pechan o *diccionari* dúas listas de lemas que diferencian aqueles que son principais e secundarios. A primeira fai unha ordenación alfabética de lemas principais e secundarios (páx. 1351-1472), en función do lema principal, cos seus datos identificativos, a súa frecuencia particular e a acumulada. Debaixo de cada un deles van os seus lemas secundarios asociados con cadansúa propia

frecuencia. A lista consta de 8.348 entradas, das que 688 presentan en branco o espacio reservado para a frecuencia, o que significa que estes lemas non aparecen no subcorpus non literario, pero si no conxunto total do corpus, aparecendo neste tramo só lemas secundarios. En segundo lugar témo-la lista alfabética dos lemas secundarios cos principais (páx. 1473-1559), onde aparecen os lemas secundarios nunha columna (9833), e na outra os principais cos que están asociados. Isto facilita a localización de cada lema secundario con respecto ó seu principal, tanto para saber se algún deles está recollido no corpus, como para ir busca-los seus datos cuantitativos nas entradas pertinentes das listas anteriores.

Os datos en soporte magnético achéganse nun CD-ROM[6] que vén conxuntamente co volume, onde aparecen tódolos datos cuantitativos e léxicos recollidos, así como información cronolóxica e tipolóxica non reflectida no volume impreso.

O deseño informático desta parte está pensado para permitir unha busca da información o máis aberta e variada posible, polo que o seu programa de consulta está pensado para poder tira-los datos requiridos, xa pola pantalla, imprimilos ou transportalos a outro soporte magnético, contando tamén coa posibilidade de traspasa-las informacións a outro tipo de base de datos.

Son dous os xeitos de consulta permitidos. Coa consulta simple pódense obter datos dun lema ou varios, partindo da súa grafía total ou parcial, da frecuencia, índice de dispersión ou uso. Coa consulta complexa pódense combinar calquera dos rexistros codificados, tanto cuantitativos como léxicos e gráficos. O usuario elixe o formato e os criterios de presentación das consultas ata un número case ilimitado de combinacións.

A finalidade deste traballo, como xa dixemos, é a de achegar uns datos determinados que sirvan como fonte para a elaboración dun *Diccionari del Catalá Contemporani*. Dependendo de cómo se utilicen estes materiais, e a información que deles se tire, así se determinará a estructura e o contido do diccionario pretendido. Así, poderase fixa-la macroestructura da obra partindo da comparación dos datos obtidos do corpus, co léxico recollido nos diccionarios de catalán máis representativos[7], podéndose observa-lo punto de intersección entre os datos do corpus e as anteriores obras lexicográficas, comprobándose que palabras que se atopan na maioría dos diccionarios case non se usan, e outras, que si se empregan, non se recollen. Unha análise moi detallada por grupos léxicos axudará a fixa-la selección oportuna de termos que se incluirán no futuro *Diccionari*, así como a disposición dos artigos.

Pero este traballo tamén pode condiciona-la natureza da microestructura, posto que mostra unha información obxectiva e exhaustiva sobre o uso que se fai dun elemento léxico, o que pode axudar, nunha altísima porcentaxe, na redacción do artigo. Así, é posible establecer unha rede de significacións entre os valores léxicos de cada palabra, fronte ó sistema de traballo lexicográfico tradicional

[6]Require para o seu manexo un material informático bastante accesible: procesador Intel 386-33 ou superior, memoria RAM de 4 Mb, espacio no disco duro de 4 Mb, contorno Windows 3.1 ou superior.

[7]O *Diccionari general de la llengua catalana* de Pompeu Fabra e o *Diccionari de la llengua catalana* da Enciclopedia Catalana, cos que se elaborou un diccionario máquina, o *Diccionari Bàsic Informatitzat*.

baseado nas obras precedentes e na compe-tencia que o lexicógrafo ten da lingua. Tamén se poden conseguir informacións sobre estructuras sintácticas, ocorrencias léxicas coincidentes e derivación e composición de palabras, estas últimas tan pouco tratadas nos diccionarios (si exceptuamos aquelas que se atopan lexicalizadas) ó non se constituíren como unidades léxicas cun significado primitivo diferente ó da palabra orixe[8].

Unha obra lexicográfica elaborada partindo do material ofrecido neste volume do *Diccionari de freqüències* (e nos futuros volumes), previamente analizado e traballado convenientemente, pode, pois, ofrecer un alto grao de garantías sobre a fiabilidade e representatividade da lingua que pretende describir, afastándose da subxectividade e das limitacións que ofrecen os méto-dos lexicográficos tradicionais, por iso, é grandemente eloxiable a aparición deste traballo, tanto polo que ten de planificación e minuciosidade, como polos campos metodolóxicos que abre, tan necesarios e desexables tamén para a lexi-cografía galega.

XESÚS DOMÍNGUEZ DONO

[8]Para unha descrición un pouco máis detallada sobre o proxecto do *DCC*, véxanse as páxinas 168-178 do artigo de J. Rafel citado na nota 1.