Els Marges *107, Tardor 2015, Barcelona, ISSN 2339-8256, p. 32-56*

La riquesa lèxica dels escriptors catalans contemporanis: dades quantitatives

JORDI GINEBRA *Universitat Rovira i Virgili*

RESUM: Aquest treball té com a objectiu mostrar l'interès que ofereix per als estudis de llengua i literatura catalanes l'explotació sistemàtica de dades quantitatives de lèxic que avui poden obtenir-se gràcies a les noves possibilitats tècniques de formació i tractament de corpus digitals. Específicament, l'article mostra que l'aproximació quantitativa al concepte de riquesa lèxica permet objectivar judicis i valoracions en l'àmbit de la llengua i la literatura catalanes.
PARAULES CLAU: lingüística quantitativa, lingüística de corpus, riquesa lèxica, literatura catalana.

ABSTRACT: The aim of this paper is to show how the systematic exploitation of quantitative lexical data, nowadays made possible through the use of digital resources, can be usefully applied to the study of Catalan language and literature. Specifically, the article shows how a quantitative approach to lexical richness can make for a higher degree of objectivity in value judgements.
KEYWORDS: quantitative linguistics, corpus linguistics, lexical richness, Catalan literature.

Introducció

Aquest treball –que presenta un caràcter de recerca marc– té com a objectiu general mostrar l'interès que ofereix per als estudis de llengua i literatura catalanes l'explotació sistemàtica de dades quantitatives de lèxic que avui poden obtenir-se gràcies a les noves possibilitats tècniques de formació i tractament de corpus digitals.

Els objectius específics són els següents. En primer lloc, mostrar que l'anàlisi quantitativa del lèxic pot ajudar a estudiar la història de la llengua contemporània i de la codificació del català modern i, per tant, a respondre a preguntes com aquestes: ¿Què va representar la normativització en l'evolució del lèxic contemporani? ¿Quins autors han estat considerats modèlics i amb quina justificació? ¿És cert que la gran riquesa lèxica de Verdaguer va influir en la codificació normativa del lèxic català?

En segon lloc, aportar dades generals sobre l'estructura quantitativa de la llengua catalana. Si el progrés en la descripció lingüística té a veure amb el progrés en la

caracterització fonètica, fonològica, morfològica, lèxica i sintàctica de les llengües, també cal considerar que la caracterització dels idiomes des del punt de vista de la freqüència de les seves unitats pertany al treball del lingüista (MULLER 1968, p. 347).

En tercer lloc, mostrar que l'anàlisi quantitativa del lèxic pot ser útil als especialistes en literatura. Determinades anàlisis d'escriptors realitzades des de l'àmbit de la lingüística no interessen gaire als historiadors i crítics de la literatura. En canvi, sembla que els resultats d'anàlisis com la que es presenta els poden servir per validar o completar hipòtesis i conclusions.

Finalment, i encara que pot semblar que hi té poca relació directa, el treball és un primer pas per penetrar posteriorment en un camp més complex, que és el de la fraseologia quantitativa. De moment, com es veurà, només es faran en aquest àmbit uns quants comentaris generals.

La recerca va consistir a mirar d'objectivar amb dades quantitatives, tant concretes com generals, per als escriptors catalans contemporanis, el concepte de riquesa lèxica. De fet, la idea de riquesa lèxica apareix sovint tant en estudis de literatura com en treballs relacionats amb l'aprenentatge de llengües, i fins i tot en la conversa de persones no relacionades professionalment amb la llengua i la literatura: és un concepte que normalment serveix per valorar positivament escriptors, oradors, conversaires i tertulians.[1]

El concepte de riquesa lèxica: consideracions i estat de la qüestió

La idea bàsica de riquesa lèxica és de caràcter quantitatiu: l'obra d'un autor tindrà més riquesa lèxica com més extens sigui l'inventari de mots usats per aquest autor. L'inventari dels mots d'un autor, però, és poc significatiu si no el podem comparar amb el d'altres autors. Si disposem de l'inventari dels mots usats per dos autors, podrem dir que el lèxic de l'autor *A* és més ric que el lèxic de l'autor *B* si el seu inventari és més extens. Tot i així, la quantificació comparada de la riquesa lèxica de dos o més autors continua sent poc significativa si l'extensió de l'obra de l'un i de l'altre no és comparable. Generalment, com més extens és un text, més nombre d'unitats lèxiques conté. Un text de 200.000 paraules pot tenir 10.000 unitats lèxiques diferents, però un text de 20.000 paraules és molt difícil que arribi a les 5.000 unitats lèxiques diferents. Aquesta relació entre extensió del text i quantitat de mots diferents s'expressa tècnicament dient que *V* (nombre d'unitats diferents) és funció

1. L'autor ha d'agrair, a propòsit de dues exposicions públiques d'aquest treball, els comentaris d'Albert Rossich, Salvador Oliva, Antoni Serrà Campins, Vicent Simbor, Albert Hauf, Ferran Carbó, Joan Mas i Vives, Vicent Salvador, Joan Alegret, Brauli Montoya, Josep Murgades, Joaquim Viaplana i, especialment —també per les seves orientacions bibliogràfiques—, a Teresa Cabré.

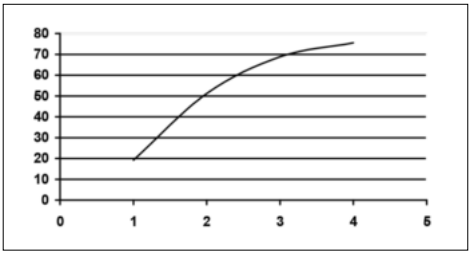
de *N* (extensió o nombre total d'unitats). D'acord amb això, quan comparem dos autors amb una obra completa molt desigual quant a l'extensió, el que en resulta és que l'autor amb l'obra més extensa és sempre el que presenta un lèxic més ric.

Per aconseguir que la riquesa lèxica esdevingui un concepte amb sentit, més enllà, doncs, de ser una simple projecció de l'extensió d'un text, s'introdueixen factors de correcció. Es pot parlar, llavors, de *riquesa lèxica relativa*, que és la proporció entre extensió de l'obra i el nombre d'unitats lèxiques diferents. Aquesta proporció és anomenada generalment *índex V/N*: el resultat de dividir el nombre de mots diferents d'un text (*V*) pel nombre d'ocurrències totals del text (*N*). L'índex sovint es multiplica per 100 per fer-lo més manejable. Aquest índex –que en aquest treball també serà anomenat *IRLR* (índex de riquesa lèxica relativa)– podria ser considerat l'indicador de la riquesa lèxica d'un autor o d'un text, i es podria convenir, doncs, que a l'hora de valorar la riquesa lèxica de l'obra d'un autor és més significatiu que no pas la xifra que expressa la riquesa lèxica absoluta. Per exemple, el *Bestiari* de Pere Quart conté 715 unitats lèxiques, i *Les flors de muntanya* de Marià Aguiló en contenen només 471. Però l'índex *V/N* és favorable a Marià Aguiló: un 46,04 contra un 45,36.

Si es compara l'índex *V/N* d'una mostra qualsevol de textos, el que es detecta és el següent: aquest índex és, en general, més baix com més extens és el text. Dit d'una altra manera: si bé un text més llarg conté generalment un nombre més alt d'unitats lèxiques diferents que no pas un text més curt, *proporcionalment a l'extensió* un text curt conté més unitats lèxiques diferents que no pas un text llarg (aquesta constatació de fet coincideix amb la percepció del sentit comú: es repeteixen més les paraules). Per exemple, l'índex *V/N* de *Les flors de muntanya*, com hem vist, és 46,04. És un índex que contrasta amb l'índex *V/N* de les *Memòries* de Sagarra –una obra molt més extensa–, que és 3,96.

Aquest estat de coses se sol representar en abstracte per mitjà del que s'anomena la línia de creixement lèxic: en un eix de coordenades es col·loca *N* a l'abscissa i *V* a l'ordenada. En aquesta representació, la línia que indica l'increment no és recta, sinó corba i, més en concret, parabòlica (GUIRAUD 1960, MULLER 1968, LABBÉ; HUBERT 1997, MALVERN i altres 2004), com es representa al gràfic 1.

Gràfic 1. Línia de creixement lèxic



Això significa que, de fet, fins i tot posseint l'índex *V/N* com a eina, només podem comparar, quant a riquesa lèxica, obres de la mateixa extensió. Aquest ha estat l'escull més gros dels estudis de lexicometria i lingüística quantitativa dels darrers seixanta anys. I ha provocat una intensa recerca per mirar de trobar un altre índex que no fos sensible a l'extensió i que permetés fer avaluacions comparatives de textos d'extensió desigual. De fet, es pot dir que ha estat i és el principal tema de recerca en estadística lingüística i estilística computacional des dels pioners en aquest camp. Les anàlisis que s'han fet tenen en compte altres variables a més de *V* i *N*, com ara la distribució interna de les freqüències, i el que en general pretenen és, amb l'ajut de l'estadística, buscar constants que permetin predir la riquesa lèxica d'un text més extens a partir d'un text més curt de les mateixes característiques. Tot plegat ha donat lloc a diverses propostes d'índexs, que s'expressen generalment per mitjà de fórmules matemàtiques complexes: característica *K* de Yule, índex *D* de Simpson, índex *R* de Guiraud, índex *W* de Brunet i uns quants més (MULLER 1968, BRUNET 1978, MULLER 1970, HOLMES 1988, TWEEDIE; BAAYEN 1998, RIBA; GINEBRA 2000).

En els darrers anys, una bona part dels treballs apareguts en aquest camp són treballs experimentals de verificació: estudis que apliquen els diferents índexs als mateixos textos i miren de descobrir quin és menys sensible a l'extensió (MENARD 1983, COSSETTE 1994). Amb tot, no hi ha encara un acord general en aquesta matèria. Un article recent posa en dubte que cap d'aquestes índexs sigui realment representatiu de la riquesa lèxica d'un text (HOOVER 2003).

Una altra manera d'introduir factors de correcció per evitar el problema de la desigualtat d'extensió ha estat segmentar les diferents obres en parts de la mateixa extensió (de vegades en parts iguals a l'obra més curta del conjunt d'obres que es vol analitzar), i establir llavors comparacions a partir d'aquestes parts (vegeu, per exemple, LABBÉ; HUBERT 1997). Però aquest procediment també ha estat criticat, perquè la distribució del lèxic en una obra no és necessàriament homogènia en cada una de les parts en què podem segmentar-la artificialment, per això s'ha dit que la segmentació distorsiona el text i no permet obtenir informació real de la seva riquesa lèxica. Serant i Thoiron (1988), en aquesta línia, sostenen que la riquesa lèxica no es pot mesurar només amb el recompte de freqüències d'un text, sinó que convé tenir en compte la «topografia» de les repeticions, és a dir, en quins llocs del text es concentren les repeticions.

La recerca: preliminars i dades

Com s'ha dit, tots els índexs complexos que s'han proposat per mirar de corregir l'índex simple *V/N* operen amb dades relatives a la freqüència de les unitats lèxiques dins dels textos i a la distribució d'aquestes freqüències. Per poder treballar

amb aquests sistemes de mesurament cal, per tant, disposar de textos analitzats de manera que se n'hagin obtingut les variables necessàries: nombre de paraules, freqüència i rang de les unitats, distribució de les freqüències, etc. No hi ha de moment, per a la llengua catalana, cap conjunt extens de textos analitzats d'aquesta manera, i per tant, amb independència de la utilitat d'aquests índexs, el fet és que no es poden aplicar perquè no hi ha dades disponibles.

¿Quines dades disponibles hi ha per a la llengua catalana? Si bé no hi ha un conjunt extens de textos amb la informació per a cada text de les variables necessàries per obtenir els índexs esmentats, el fet és que sí que hi ha un conjunt extens de textos amb, almenys, la informació relativa a les magnituds *N* (extensió) i *V* (vocabulari), que permeten, per tant, obtenir l'índex *V/N* i fer una fotografia general de la riquesa lèxica dels escriptors catalans contemporanis. El treball que tot seguit s'exposa, doncs, parteix d'unes limitacions molt clares quant a les dades, però tot i així, com es mirarà de mostrar, sembla que té un interès indubtable.

La informació sobre l'extensió (nombre d'ocurrències o paraules de text) i sobre el nombre d'unitats lèxiques d'un conjunt extens d'obres literàries catalanes la podem obtenir del Corpus Textual Informatitzat de la Llengua Catalana de l'Institut d'Estudis Catalans. Com se sap, aquest corpus (des d'ara CTILC) està constituït pel text, en suport digital, d'obres literàries i no literàries catalanes publicades entre els anys 1833 i 1988. El CTILC inclou 3.299 obres diferents, que sumen més de 50 milions d'ocurrències.

El CTILC –que es pot consultar en línia– ofereix a l'investigador una fitxa per a cada una de les obres que conté, que informa —en relació amb el que ara interessa— del nombre d'ocurrències i el nombre d'unitats lèxiques de cada obra. A més, també informa del nombre de formes (*types*) de cada text.[2]

Abans de continuar, és imprescindible fer un petit aclariment terminològic. Fins ara s'ha utilitzat l'expressió *unitat lèxica* per evitar començar l'exposició fent precisions. A partir d'ara, es farà servir el terme *lema*, d'acord amb un ús prou general i amb el que trobem al CTILC. D'altra banda, se substituirà el símbol *V*, utilitzat fins ara per indicar el nombre de lemes d'un text, pel símbol *L*. La raó és que en els estudis lexicomètrics i de lingüística quantitativa el símbol *V* es fa servir indistintament per indicar el nombre de formes d'un text i el nombre de lemes.

2. Per conèixer les característiques generals i altres dades quantitatives del CTILC, es pot consultar Joaquim Rafel (1996, 1998*a* i 1998*b*). Es pot accedir en línia a altres documents que n'informen (<http://ctilc.iec.cat/>). Actualment l'accés a la fitxa de cada obra no és immediat, i requereix unes quantes operacions de navegació dins el web.

La recerca: el pla de treball

El pla de treball previst era senzill, per bé que de realització lenta. Consistia a: *a*) accedir —per mitjà de la Xarxa— al Portal de la llengua de l'IEC, consultar cada una de les 3.299 fitxes del CTILC i anotar el nombre d'ocurrències i el nombre de lemes de cada una de les obres; *b*) ordenar les obres en funció de l'extensió i associar cada obra amb el valor *L* (nombre de lemes) corresponent per tal d'obtenir informació comparativa general; *c*) obtenir l'índex de riquesa lèxica relativa (*L/N*) de cada obra dividint, com s'ha dit, el nombre de lemes pel nombre d'ocurrències; *d*) agrupar les obres en conjunts segons l'extensió; i establir el resultat en forma d'una llista jerarquitzada per a cada conjunt; *e*) fer les valoracions oportunes i obtenir les conclusions pertinents.

Tot seguit es comenten breument uns quants problemes i incidències d'alguns d'aquests processos. Pel que fa a l'accés al Portal de la llengua de l'IEC i a la consulta de les fitxes corresponents a cada una de les 3.299 obres del CTILC, cal dir que l'aplicació informàtica permetia accedir a cada una de les fitxes de les obres si l'usuari proporcionava el nom de l'autor o el títol de l'obra. L'aplicació, doncs, no oferia la possibilitat d'accedir directament a la llista de les obres. Això va fer que el procés d'obtenció de les dades de les fitxes fos lent. (A més, el procés es va alentir encara més perquè de tant en tant el Portal avisava l'usuari i li deia el següent: «Disculpeu per les molèsties. El nostre servidor no pot atendre la vostra petició en aquest moment. Torneu-ho a intentar d'aquí una estona».) Aquestes circumstàncies van fer desistir, almenys per a aquest primer estudi, d'examinar les fitxes de tot el CTILC, i el treball es va limitar a analitzar les fitxes de les obres del subcorpus literari, una de les dues seccions en què està dividit el CTILC (l'altra secció és el subcorpus no literari, que conté obres que no són de literatura). El subcorpus literari conté 23.105.591 ocurrències (una mica menys de la meitat del total) i 1.011 títols. La consulta de les fitxes es va fer seguint la llista impresa de les obres del subcorpus literari que figura a Rafel (1998*a*: XVII-XXX).[3]

Quant a la jerarquització de les obres en funció de l'extensió i quant a l'assignació a cada obra del valor *L* (nombre de lemes) corresponent per tal d'obtenir informació comparativa general, cal indicar que gràcies a l'ajut de Josep Ginebra, professor del Departament d'Estadística i Investigació Operativa de la UPC, aquest pas no va suposar cap dificultat, i es va materialitzar en la representació de les dades per mitjà d'uns eixos de coordenades que es mostraran tot seguit. Pel que fa a l'obtenció de l'índex de riquesa lèxica relativa (*L/N*) de cada obra dividint el nombre de

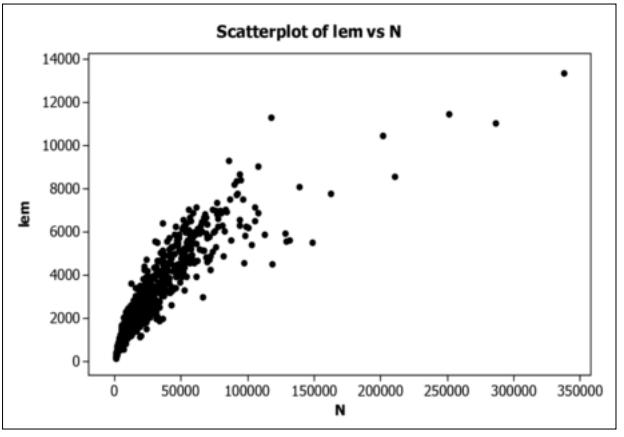
3. Actualment ja no s'accedeix al CTILC per mitjà del Portal de la llengua, sinó que es fa per l'adreça web <http://ctilc.iec.cat/> i, per tant, alguns dels problemes comentats s'han resolt. La interfície de consulta també ha canviat.

lemes pel nombre d'ocurrències, no hi va haver cap problema, tot i que el procés va ser llarg.

La recerca: resultats

Per fer la jerarquització de la riquesa lèxica, es van ordenar les obres en funció de l'extensió i es va associar cada obra amb el valor *L* (nombre de lemes) corresponent per tal d'obtenir informació comparativa general. El resultat es presenta al gràfic 2, el diagrama de dispersió dels valors *N* i *L* de les 1.011 obres estudiades. Tal com es preveia, la figura és més alta per a valors de *N* més grans, i la forma general de la línia subjacent és corba i tendeix a la paràbola.

Gràfic 2. Creixement i dispersió de la riquesa lèxica del subcorpus literari



Perquè el lector pugui fer-se una idea ni que sigui molt general de la correspondència dels punts, indicarem que el punt del gràfic que apareix més a la dreta correspon a les *Memòries* de Sagarra, que tenen 337.454 ocurrències i 13.388 lemes. El segon correspon a *Camins de França* de Joan Puig i Ferreter, amb 285.692 ocurrències i 11.037 lemes. El tercer, a *El quadern gris*, de Josep Pla, que, tot i ser una obra més curta que l'anterior (251.198 ocurrències), és més rica lèxicament (11.477 lemes).

A més, es va obtenir l'índex de riquesa lèxica relativa (*L/N*) de cada obra dividint, com s'ha indicat, el nombre de lemes pel nombre d'ocurrències i multiplicant per 100; es van agrupar les obres en conjunts segons l'extensió, i es va establir el resultat en forma d'una llista jerarquitzada per a cada conjunt. Tal com es preveia, les obres més extenses tenen un índex *L/N* més baix. L'índex oscil·la entre el 3,96 i

el 64,08. Pel que fa a les agrupacions, el resultat és que les 1.011 obres es distribueixen, quant a extensió, segons la taula 1:

Taula 1. Obres i ocurrències (general).

Obres de més de 300.000 ocurrències	1
Obres que tenen entre 200.001 i 300.000 ocurrències	4
Obres que tenen entre 100.001 i 200.000 ocurrències	14
Obres que tenen entre 1 i 100.000 ocurrències	992
TOTAL OBRES	1.011

Atès que la majoria d'obres no arriben a les 100.000 ocurrències, va semblar convenient formar conjunts establint franges quantitatives més petites. Les 992 obres que tenen entre 1 i 100.000 ocurrències es van classificar com mostra la taula 2:

Taula 2. Obres de fins a 100.000 ocurrències.

Obres que tenen entre 90.001 i 100.000 ocurrències	12
Obres que tenen entre 80.001 i 90.000 ocurrències	9
Obres que tenen entre 70.001 i 80.000 ocurrències	17
Obres que tenen entre 60.001 i 70.000 ocurrències	22
Obres que tenen entre 50.001 i 60.000 ocurrències	50
Obres que tenen entre 40.001 i 50.000 ocurrències	55
Obres que tenen entre 30.001 i 40.000 ocurrències	62
Obres que tenen entre 20.001 i 30.000 ocurrències	114
Obres que tenen entre 10.001 i 20.000 ocurrències	235
Obres que tenen entre 1 i 10.000 ocurrències	416
TOTAL D'OBRES QUE TENEN ENTRE 1 I 100.000 OCURRÈNCIES	992

L'últim dels conjunts, el més nombrós, es va dividir al seu torn en franges de 1.000 ocurrències (taula 3):

Taula 3. Obres de fins a 10.000 ocurrències.

Obres que tenen entre 9.001 i 10.000 ocurrències	24
Obres que tenen entre 8.001 i 9.000 ocurrències	27
Obres que tenen entre 7.001 i 8.000 ocurrències	31
Obres que tenen entre 6.001 i 7.000 ocurrències	55
Obres que tenen entre 5.001 i 6.000 ocurrències	41
Obres que tenen entre 4.001 i 5.000 ocurrències	66
Obres que tenen entre 3.001 i 4.000 ocurrències	60
Obres que tenen entre 2.001 i 3.000 ocurrències	48
Obres que tenen entre 1.001 i 2.000 ocurrències	46
Obres que tenen entre 1 i 1.000 ocurrències	18
TOTAL D'OBRES QUE TENEN ENTRE 1 I 10.000 OCURRÈNCIES	416

Aquesta agrupació tenia l'objectiu de fer prediccions, per mitjà del càlcul de la mitjana de riquesa lèxica relativa de cada conjunt, sobre el nombre de lemes d'una obra d'una determinada extensió. Té un problema teòric, que és la justificació dels talls que donen lloc a cada grup, però el resultat és vàlid com a aproximació general, com es veurà. Els talls fets serien rebutjables per inoperativitat si resultés que en cada grup les obres més extenses fossin les d'un índex més alt, però això no és així. Per tant, les dades empíriques validen l'operativitat dels talls. L'índex *L/V* (o IRLR) és el següent:

1. Grups d'obres que tenen entre 1 i 10.000 ocurrències

Ocurrències	*Obres*	*Mitjana de L/V*
entre 9.001 i 10.000	24	18,37
entre 8.001 i 9.000	27	19,01
entre 7.001 i 8.000	31	19,66
entre 6.001 i 7.000	55	20,94
entre 5.001 i 6.000	41	21,16
entre 4.001 i 5.000	66	24,27
entre 3.001 i 4.000	60	26,02
entre 2.001 i 3.000	48	28,95
entre 1.001 i 2.000	46	33,98
entre 1 i 1.000	18	44,03

Per il·lustrar el que signifiquen aquests índexs quant al nombre concret d'unitats lèxiques que presenten les diferents obres, es pot dir que, si adoptem la mitjana de riquesa lèxica relativa de cada grup com a índex de referència per a les obres de grup, això significaria, per exemple, que esperaríem les correlacions següents (s'eliminen els decimals):

Una obra de	500	*ocurrències*	*conté*	220	*unitats lèxiques*
	1.500			509	
	2.500			723	
	3.500			910	
	4.500			1.092	
	5.500			1.163	
	6.500			1.361	
	7.500			1.474	
	8.500			1.615	
	9.500			1.745	

2. Grups d'obres que tenen entre 1 i 100.000 ocurrències

Ocurrències	*Obres*	*Mitjana de L/V*
entre 90.001 i 100.000	12	7,44
entre 80.001 i 90.000	9	8,14
entre 70.001 i 80.000	17	8,15
entre 60.001 i 70.000	22	8,71
entre 50.001 i 60.000	50	9,85
entre 40.001 i 50.000	55	10,35
entre 30.001 i 40.000	62	11,08
entre 20.001 i 30.000	114	12,30
entre 10.001 i 20.000	235	14,36
entre 1 i 10.000	416	25,22

Una obra de	5.000 15.000 25.000 35.000 45.000 55.000 65.000 75.000 85.000 95.000	*ocurrències*	*conté*	1.261 2.154 3.075 3.878 4.657 5.417 5.661 6.112 6.919 7.068	*unitats lèxiques*

3. Grups d'obres que tenen entre 1 i 400.000 ocurrències

Ocurrències	*Obres*	*Mitjana de L/V*
entre 300.001 i 400.000	1	3,96
entre 200.001 i 300.000	4	4,43
entre 100.001 i 200.000	14	5,68
entre 1 i 100.000	992	17,65

Una obra de	50.000 150.000 250.000 350.000	*ocurrències*	*conté*	8.825 8.520 11.075 13.860	*unitats lèxiques*

A l'apèndix hi ha una mostra del resultat exhaustiu d'aquesta operació (l'espai impedeix presentar-ne tot el resultat). Les obres són les del grup que va de 20.001 a 30.000 ocurrències, i estan ordenades per ordre decreixent d'IRLR. Encara que la totalitat de les dades, com a conjunt, respongui a la llei prevista de decrement progressiu de l'índex de RLR, pel que fa a les obres concretes hi pot arribar a haver grans contrastos.

Valor i interpretació dels resultats I: la dispersió morfològica a través de la història

Al començament d'aquesta exposició es deia que treballs com aquest poden ser útils per respondre a preguntes de l'àmbit de la història i la codificació del català contemporani. Tot seguit s'il·lustra aquest punt. Seria raonable, per exemple, fer la

hipòtesi que els textos anteriors a la normativització presenten un grau de dispersió morfològica superior al dels textos de després de la normativització. Com s'ha dit, es pot obtenir del CTILC, per a cada obra literària, no sols el nombre de lemes, sinó també el nombre de formes, que és sempre superior al nombre de lemes. La dispersió morfològica es pot mesurar, en cada obra, per mitjà de l'índex *F/L* (forma/lema), que indica el nombre de formes per a cada lema. Com més alt és l'índex, més dispersió hi ha.

Per tal de fer una primera prospecció, es van dividir les obres del subcorpus en tres períodes:

Primer període (A):	1833-1882
Segon període (B):	1883-1932
Tercer període (C):	1933-1988

Tot seguit es va comparar la mitjana de l'índex *F/L* per al primer i el tercer períodes. El resultat és el següent:

Períodes	*F/L*
Primer (A):	1,66
Tercer (C):	1,53

S'observa que la dispersió morfològica és més alta en el primer període. El contrast és més marcat en la franja d'obres poc extenses. Es verifica la hipòtesi, doncs, tot i que el contrast és potser menor del que s'esperava (una dada d'interès, que potser hauria de fer qüestionar la idea que el català escrit anterior a la normativització era anàrquic des del punt de vista formal). Amb altres talls cronològics i combinant paràmetres, es pot arribar a detalls més precisos.

Valor i interpretació dels resultats II: la riquesa lèxica a través de la història

Encara en l'àmbit de la història de la llengua podríem fer-nos una altra pregunta lògica, que és la següent. La riquesa lèxica, amb el temps, ¿creix o decreix? ¿Hi ha diferències entre la riquesa lèxica dels escriptors des del punt de vista cronològic? ¿Podem fer afirmacions com ara que el lèxic de la Renaixença és més pobre/ric que el del Noucentisme?

La intuïció ens deia que la riquesa lèxica havia de créixer amb el pas del temps, és a dir, que la mitjana de l'IRLR del primer període havia de ser inferior a la del

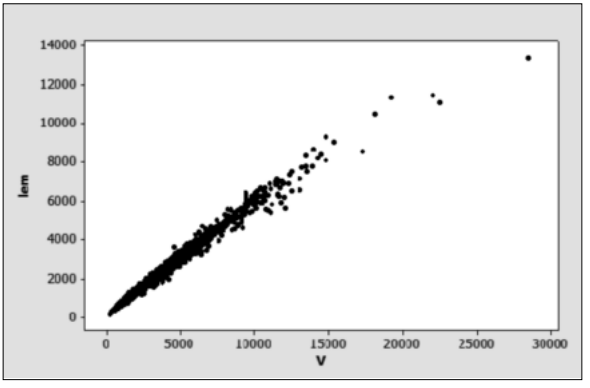
darrer període. Les dades, un cop calculada la mitjana dels períodes per a obres d'extensió comparable, són les següents (mitjana de l'IRLR per períodes per a les obres de 10.001 a 20.000 ocurrències):

Primer període: 1833-1882	Segon període: 1883-1932	Tercer període: 1933-1988
15,02	14,40	14,12

En aquest cas, la hipòtesi que es derivava de la intuïció inicial no es va poder verificar. Segurament, el que tindria interès en aquest punt és poder comparar, a més, grans períodes històrics. Per exemple, seria interessant comparar la llengua d'avui amb la llengua medieval. Dissortadament, no tenim corpus prou extensos de català medieval lematitzats que ens puguin servir per a la comparació. De tota manera, es va fer una petita recerca que sí que fa llum. El corpus de català antic més extens que existeix és el CICA (Corpus Informatitzat del Català Antic), impulsat per Joan Torruella i dirigit avui per ell juntament amb Manuel Pérez Saldanya i Josep Martines (CICA 2005, TORRUELLA 2005). Aquest corpus no té els mots lematitzats. Únicament podem obtenir, per a cada obra, el nombre d'ocurrències i el nombre de formes.[4]

La petita recerca va consistir en el següent. Es va pensar que si, per al CTILC, la proporció entre lemes i formes fos constant, es podria aplicar aquesta constant a les obres del CICA i fer una estimació del nombre de lemes de les obres d'aquest corpus sense haver de lematitzar-les. El gràfic 3 ens mostra la relació forma/lema del subcorpus literari del CTILC.

Gràfic 3. Relació entre V (formes) i L (lemes)



4. Avui (però encara no quan es va fer aquest estudi) aquest corpus, prou més ampliat, es pot consultar en línia: <http://www.cica.cat/index.php>.

La figura no és sinó la representació gràfica d'una fórmula obtinguda segons el model de regressió lineal. El que es pot apreciar per la forma de la figura és que la línia subjacent és una diagonal gairebé perfecta i, per tant –efectivament–, en termes estadístics generals es pot dir que la proporció entre formes i lemes és constant. Aplicada la fórmula a uns quants textos del CICA triats a l'atzar, obtenim aquest resultat:

Textos obtinguts del CICA (2005)

N = extensió

V = nombre de formes

L* = nombre de lemes estimats (model de regressió lineal obtingut de l'anàlisi del CTILC)

L** = nombre de lemes previsible segons mitjana de l'IRLR del grup corresponent del CTILC

1. Textos extensos

	N	V	L*	L**
Vides de sants rosselloneses	162.753	13.393	7.955	9.244
Furs de València (Furs de Jaume I)	152.645	6.920	4.110	8.670
Epistolaris d'Hipòlita Roís de Liori i d'Estefania de Requesens	157.549	9.493	5.638	8.948

2. Textos breus

Jurament de compareixença	244	170	101	107
Greuges de Guitard Isarn, senyor de Caboet	752	305	181	331
Jurament de pau i treva del comte Pere Ramon de Pallars Jussà al bisbe d'Urgell	231	135	80	101
Llibre dels judicis. Fragment primer	457	225	133	201
Capbreu de Castellbisbal	289	154	91	127
Greuges dels homes de Sant Pere de Graudescales	793	207	123	349

El resultat, que cal considerar merament aproximatiu, ens indica que el lèxic del català antic és més pobre que el del català contemporani. Aquest resultat no és coherent amb l'anterior, però sí que ho és amb el resultat que va obtenir Étienne Brunet per al francès contemporani a partir del corpus del *Trésor de la langue française*, que li va servir per crear el concepte d'*inflació lexical*. Brunet (1988: 25, 51-79) arriba a dir que, per causa de la inflació lexical, els escriptors del segle XIX i els del XX no es poden comparar (o que la comparació s'ha de fer tenint en compte la «disponibilitat lèxica» de cada segle).

També és cert que Brunet assenyala que hi ha dos moments d'inflexió: cap al 1865, coincidint amb la segona meitat del Segon Imperi, i cap al 1900. El cas del segle XIX català podria explicar-se d'aquesta manera, com un punt d'inflexió. Això no vol dir que no calgui buscar-ne la causa.

Valor i interpretació dels resultats III: elements per a la descripció de l'estructura quantitativa de la llengua catalana

Com s'ha dit, amb el treball també es pretenia aportar dades generals sobre l'estructura quantitativa de la llengua catalana. De fet, ja se n'han presentat unes quantes. Les dades, en realitat, poden ser explotades de moltes maneres. Ara ens limitarem simplement a mostrar que també permeten fer comparacions quantitatives entre llengües.

El que s'ofereix tot seguit és una taula comparativa elaborada a partir de les dades del francès, procedents de Brunet (1978), i les dades del català, elaborades per a aquesta recerca. Cal tenir en compte diversos factors que sens dubte distorsionen les dades i aconsellen prudència a l'hora de fer judicis categòrics, però tot i així val la pena de fer un cop d'ull a la taula.[5]

francès		**català**	
ocurrències	*lemes*	*ocurrències*	*lemes*
1.538	500	1.500	509
2.511	717	2.500	723
4.430	1.070	4.500	1.092
9.491	1.760	9.500	1.745
45.317	4.370	45.000	4.659
150.814	8.090	150.000	8.520

5. Els factors que distorsionen les dades són, d'una banda, els criteris de lematització dels materials lèxics, que no coincideixen, i els mètodes de càlcul del nombre de lemes, que també són diferents.

El que s'observa de manera notable és la proximitat dels resultats. Si ens endinséssim en els viaranys de la història social de la llengua, esdevindria obligatòria aquesta reflexió ponderativa: resulta increïble que el català presenti aquests paràmetres de normalitat, si tenim en compte fins a quin punt ha estat una llengua maltractada.

Valor i interpretació dels resultats IV: utilitat per als estudis de literatura catalana

Els resultats obtinguts permeten ser optimistes a l'hora de reflexionar al voltant de la possibilitat que aquesta mena d'anàlisis puguin ser útils per als estudiosos de la literatura. En aquesta línia, sembla interessant presentar també unes quantes dades. Per exemple, entre els experts verdaguerians es repeteix l'afirmació que l'obra de Verdaguer és d'una gran riquesa lèxica. ¿Quin és l'abast objectiu, des del punt de vista quantitatiu, d'aquesta afirmació? El CTILC conté 12 obres de Verdaguer. Les dues més extenses, *Excursions i viatges* i *Dietari d'un pelegrí a Terra Santa*, tenen un IRLR, respectivament, d'11,44 i de 10,84. La mitjana del grup (30.001 a 40.000 ocurrències) és d'11,08. La primera obra supera lleugerament aquesta mitjana (tot i que queda molt lluny de la que ocupa el primer lloc, *Flor de card* de Salvador Galmés, que té un IRLR de 18,38), però la segona no hi arriba. També se situen al voltant de la mitjana (12,3) les dues obres de la franja següent (20.001 a 30.000 ocurrències), *L'Atlàntida* i *Canigó*, amb un IRLR de 14,64 i 12,37. En la franja següent hi torna a haver dues obres de Verdaguer, *Flors del Calvari* i *Roser de tot l'any*, que se situen (12,83 i 11,34) per sota de la mitjana (14,36). En el grup de 6.001 a 7.000 hi ha tres treballs verdaguerians. Destaca *Dos màrtirs de ma pàtria* (26,73), una de les obres amb l'IRLR més alt del grup, però les altres dues, *Al cel* (20,74) i *Jesús infant* (19,06) no arriben a la mitjana (20,94). També se situa per sota de la mitjana del grup corresponent la *Passió de Nostre Senyor Jesucrist* (27,09/28,95). I, pel que fa a les obres més curtes, *Romanç llegit a l'ombra de la font del Desmai* i *A Barcelona*, tenen un IRLR de 37,42 i 32,78 (mitjana del grup: 33,98).

Verdaguer, doncs, no destaca especialment per la riquesa quantitativa del seu lèxic. En canvi, un autor de la mateixa època, del qual s'ha afirmat sovint que escrivia amb una llengua pobra i insegura, Narcís Oller, presenta les dades següents (vegeu també Ginebra 2010). El CTILC conté 11 obres d'Oller, que són *Croquis del natural*, *Isabel de Galceran* (l'embrió de *Vilaniu*), *La papallona*, *Notes de color*, *L'escanyapobres*, *Vilaniu*, *La febre d'or*, *La bogeria*, *Pilar Prim*, *Rurals i urbanes* i *Al llapis i a la ploma*. En el cas de *La febre d'or*, els recomptes d'ocurrències i lemes no es van fer de l'obra completa sinó de cada un dels volums, com si fossin tres obres independents. Per tant, amb relació al que ara ens ocupa, hem de conside-

rar com si al subcorpus literari del CTILC hi hagués 13 obres d'Oller. Doncs bé: totes superen la mitjana del grup corresponent, com mostra la taula següent.

	Obra	*N*	*IRLR*	*Mitjana del grup*
1	Isabel de Galceran	6.923	24,38	20,94
2	Croquis del natural	21.295	14,67	12,3
3	L'escanyapobres	22.548	15,96	12,3
4	La bogeria	30.398	13,37	11,08
5	La febre d'or I	43.057	12,33	10,35
6	Notes de color	46.611	11,72	10,35
7	Rurals i urbanes	47.429	12,09	10,35
8	Al llapis i a la ploma	47.628	12,33	10,35
9	La papallona	48.858	10,53	10,35
10	La febre d'or II	58.303	9,93	9,85
11	La febre d'or III	63.664	9,30	8,71
12	Pilar Prim	79.210	8,69	8,15
13	Vilaniu	89.446	9,16	8,14

Aquestes dades no necessàriament desmenteixen l'afirmació generalment admesa. En tot cas, sembla que ajuden a objectivar-la: la riquesa lèxica de Verdaguer no es mesura per l'IRLR sinó per mitjà d'altres paràmetres (Ginebra 2006). Per exemple, potser per la qualitat i la precisió del lèxic. El que sembla interessant és precisament això: amb anàlisis com aquesta es pot ajudar els experts a objectivar els seus judicis.

Vegem-ne molt breument dos casos més. Les obres de Mercè Rodorera que figuren al CTILC permeten comprovar que la riquesa lèxica d'aquesta autora és extremament baixa. Passem, però, a la interpretació. Aquesta pobresa lèxica, ¿és un factor negatiu a l'hora de valorar l'obra de Rodoreda? Potser en la mesura que, com ha indicat la crítica, Rodoreda vol utilitzar un llenguatge col·loquialitzant i senzill, les dades indiquen precisament que hi va reeixir. En el seu cas, doncs, la pobresa lèxica caldria interpretar-la com una qualitat literària. En canvi, les obres de Pru-

denci Bertrana són d'una riquesa lèxica altíssima. ¿Com cal interpretar aquest fet? ¿Aquesta riquesa és un factor positiu a l'hora de valorar l'obra de Bertrana? ¿O podria ser un indicador de l'artificiositat gratuïta de la seva prosa?

Lògicament, les dades que proporcionem, tot i que objectives, són dades, diguem-ne, «brutes». Caldria comparar textos comparables, i els textos no es fan comparables només per l'extensió, sinó pel gènere i per altres característiques. En el cas de Verdaguer, potser el que hauríem de fer, almenys per a algunes de les seves obres, les que pertanyen al subgènere de la poesia pietosa, és comparar-les amb les dades procedents d'un corpus format per textos de poesia pietosa. No hem pogut anar encara gaire enllà en aquesta recerca. Ara: les dades indiquen que la riquesa prototípica, en efecte, varia en funció dels diferents gèneres. Aquest contrast es pot expressar en termes numèrics per a obres d'extensió equiparable. Vegeu la taula següent, que expressa la mitjana de l'IRLR per gèneres (obres que tenen entre 10.001 i 20.000 ocurrències):

Poesia	Assaig	Narrativa	Teatre
17,71	15,50	15,23	12,79

Aquest punt és, sembla, el que pot arribar a donar més fruits de cara a aquesta col·laboració que es demana entre lingüistes i especialistes en literatura. Així, segons Brunet (1988: 29), el teatre té un índex més baix de riquesa lèxica perquè l'oient no pot consultar el diccionari i no pot tornar a «llegir» la peça. Per tant, una obra de teatre amb un índex de riquesa lèxica especialment alt potser no seria un indici de perícia dramàtica, sinó d'imperícia. Estudis fets per a la llengua francesa mostren que la tragèdia és més pobra lèxicament que la comèdia: la tensió tràgica col·lapsa la creativitat lèxica. Cossete (1994, p. 168-179) estudia l'obra dramàtica canadenca *Bilan* (a partir de talls de 1.000 mots), i conclou que els passatges de menor riquesa lèxica coincideixen amb els d'intensitat dramàtica més alta.

Una via que sembla de gran interès, en aquesta línia, és la distinció que estableix aquest autor, basant-se en formulacions prèvies de Muller, entre *riquesa temàtica* (o lèxic de situació, que no depèn de l'autor) i *riquesa estilística* (que és individual). Cossette (1994, p. 9-15) afirma que la riquesa lèxica no s'ha d'avaluar en brut sinó en funció dels diferents camps lèxics. Si un autor tria un determinat tema, la riquesa lèxica del que escrigui estarà condicionada pels límits del camp lèxic d'aquell tema (i, encara, un mateix text pot tenir parts més especialitzades que altres). Afegeix que mentre no tinguem instruments per mesurar cada una d'aquestes variables, el mesurament de la riquesa lèxica serà poc significatiu. (Potser la gran riquesa lèxica d'Oller caldria relacionar-la amb l'eixamplament temàtic –la incorporació de les

classes socials de la modernitat industrial i l'agitació de la vida urbana– que va significar en la literatura catalana la irrupció del novel·lista vallenc.)

Charles Bernet (1988, p. 3, 11) afirma, a propòsit de Zola —del qual s'havia dit, amb poc fonament, que tenia un vocabulari més ric que el de Chauteaubriand, Giraudoux o Proust—, que la il·lusió de la riquesa lèxica pot provenir de la presència de vocables rars o de termes tècnics, és a dir, d'un vocabulari marginal i perifèric. Menard (1983: 10) i Hoover (2003) fan comentaris en la mateixa línia.

Valor i interpretació dels resultats V: indicis per als estudis de fraseologia

Al començament de l'exposició s'ha dit que el treball que es presentava també tenia relació amb la fraseologia. Els resultats obtinguts proporcionen pistes metodològiques clares, potser ja previsibles però no per això menys interessants. Com ara que a l'hora de buscar recurrències —pas instrumental clau per a la detecció de concurrències significatives i, per tant, d'unitats fraseològiques— és poc útil recórrer a obres curtes. Si es vol formar corpus per explotar-los de cara a l'obtenció de material fraseològic, cal aplegar obres extenses.

Però els resultats obtinguts també conviden a fer-se altres preguntes interessants. Per exemple: ¿podria ser que una autora com Rodoreda, més pobra lèxicament que la majoria de novel·listes catalans, fos més rica des del punt de vista fraseològic? És una pregunta que ara no es pot respondre, però que es podrà respondre objectivament, amb els instruments que comencem a tenir, en un termini no gaire llunyà. De moment, es pot fer un raonament indirecte: si la riquesa lèxica és menor, la recurrència lèxica és més alta, i la recurrència lèxica alta sol ser signe de «fraseologicitat» d'un text.

Conclusions

Les conclusions de l'exposició són les següents. La recerca feta *a*) mostra l'interès que per a la investigació en lingüística catalana tenen les noves possibilitats tècniques d'explotació sistemàtica de dades quantitatives relatives al lèxic; *b*) aporta dades útils per a l'estudi de la història de la llengua contemporània i de la codificació del català modern; *c*) proporciona informació descriptiva relativa a l'estructura quantitativa de la llengua catalana i possibilita anàlisis contrastives amb altres idiomes; *d*) assenyala camins possibles per aconseguir que els lingüistes realitzin estudis que poden ser altament profitosos per als especialistes en literatura; *e*) suggereix àmbits d'investigació i vies d'anàlisi que caldrà tenir en compte en els estudis de fraseologia.

Bibliografia

C. BERNET, 1988: «Faits lexicaux. Richesse de vocabulaire. Resultats», dins P. THOIRON; D. LABBÉ; D. SERANT (ed.), p. 1-11.

E. BRUNET, 1978: *Le vocabulaire de Jean Giraudoux. Structure et évolution*, Ginebra: Éditions Slatkine.

E. BRUNET, 1981: *Le vocabulaire français de 1789 à nos jours*, prefaci de Paul Imbs, 3 vol., Ginebra-París: Slatkine-Champion.

E. BRUNET, 1988: «La structure lexicale dans l'oeuvre de Hugo», dins P. THOIRON; D. LABBÉ; D. SERANT (ed.), p. 23-42.

E. BRUNET, 2009: *Écrits choisis: comptes d'auteurs. Études statistiques. De Rabelais à Gracq,* vol. I, edició de Damon Mayaffre, prefaci de Henri Béhar, París: Champion.

E. BRUNET, 2011: *Ce qui compte, écrits choisis*, vol. II, edició de Céline Poudat, París: Champion.

CICA = Corpus Informatitzat del Català Antic [CD-ROM], coordinat per Joan Torruella, Bellaterra: Universitat Autònoma de Barcelona, 2005. [Corpus en formació. Actualment s'hi pot accedir telemàticament: <http://www.cica.cat/index.php.>]

CTILC = Corpus Textual Informatitzat de la Llengua Catalana. [Consultes feta del juny a l'octubre del 2005.] Adreça actual: <http://ctilc.iec.cat/>.

A. COSSETTE, 1994: *La richesse lexicale et sa mesure*, París: Editions Honoré Champion.

J. GINEBRA, 2006: «Sobre la riquesa lèxica de Verdaguer: una primera prospecció quantitativa», *Anuari Verdaguer*, núm. 14, p. 311-324.

J. GINEBRA, 2010: «Consideracions sobre la llengua de Narcís Oller», dins J. GINEBRA i altres, *Narcís Oller i Vilaniu. Primeres Jornades Narcís Oller*, Valls: Cossetània Edicions, p. 9-24.

P. GUIRAUD, 1960: *Problèmes et méthodes de la statistique linguistique*, París: Presses Universitaires de France.

D. I. HOLMES, 1988: «The Analysis of Literary Style. A Review», dins P. THOIRON; D. LABBÉ; D. SERANT (ed.), p. 67-76.

D. L. HOOVER, 2003: «Another Perspective on Vocabulary Richness», *Computers and the Humanities*, vol. 37, núm. 2, maig, p. 151-178.

D. LABBÉ; P. HUBERT, 1997: «Vocabulary richness», *Lexicometrica*, núm. 0 [Revista electrònica: consulta a l'agost del 2006]. <http://www.cavi.univ-paris3.fr/lexicometrica/article/numero0/DLVocRich.html>.

D. MALVERN; B. RICHARDS; N. CHIPERE; P. DURÁN, 2004: *Lexical Diversity and Language Development. Quantification and Assessment*, Houndmills: Palgrave Macmillan.

N. MENARD, 1983: *Mesure de la richesse lexicale. Théorie et vérifications expérimentales. Études stylométriques et sociolinguistiques*, Ginebra-París: Champion-Slatkine.

C. MULLER, 1968: *Initiation à la statistique linguistique*, París: Larousse. Versió espanyola (ed. u.): *Estadística lingüística*, Madrid: Gredos, 1973.

C. MULLER, 1970: «Sur la mesure de la richesse lexicale», *Études de Linguistique Apliquée*, nova etapa, núm. 1, p. 20-46. Reproduït dins Muller (1979), p. 281-307.

C. MULLER, 1979: *Langue française et linguistique quantitative*, Ginebra: Editions Slatkine.

J. RAFEL, 1996: «Introducció» a J. Rafel (dir.), *Diccionari de freqüències. 1 Llengua no literària*, Barcelona: Institut d'Estudis Catalans, p. VII-LXIII.

J. RAFEL, 1998*a*: «Introducció» a J. Rafel (dir.), *Diccionari de freqüències. 2 Llengua literària*, Barcelona: Institut d'Estudis Catalans, p. VII-XIII.

J. RAFEL, 1998*b*: «Introducció» a J. Rafel (dir.), *Diccionari de freqüències. 3 Dades globals*, Barcelona: Institut d'Estudis Catalans, p. VII-XIV.

À. RIBA; J. GINEBRA, 2000: «Riquesa de vocabulari i homogeneïtat d'estil en el *Tirant lo Blanc*», *Revista de Catalunya*, núm. 152, juny, p. 99-118.

D. SERANT; P. THOIRON, 1988: «Richesse lexicale et topographie des formes répétées», dins P. THOIRON; D. LABBÉ; D. SERANT (ed.), p. 125-139.

P. THOIRON; D. LABBÉ; D. SERANT (ed.), 1988: *Études sur la richesse et la structure lexicales / Vocabulary structure and lexical richness*, prefaci de Charles Muller, París-Ginebra: Champion-Slatkine.

J. TORRUELLA, 2005: «Taula de fitxes. Corpus del Català Antic (9-5-05)» [document electrònic de Word], Bellaterra: Universitat Autònoma de Barcelona.

F. J. TWEEDIE; R. H. BAAYEN, 1998: «How Variable May a Constant be? Measures of Lexical Richness in Perspective», *Computers and the Humanities*, vol. 32, núm. 5, setembre, p. 323-352.

Apèndix. Obres del subcorpus literari del CTILC que tenen entre 20.001 i 30.000 paraules d'extensió, jerarquitzades segons l'IRLR

Tipus

A (assaig)
N (narrativa)
P (poesia)
T (teatre)

Classe
Or (original)
Tr (traducció)

PDA (procedència dialectal de l'autor o traductor)
NR (nord-occidental)
V (valencià)
S (septentrional)
C (central)
B (balear)
AL (alguerès)
E (no especificat)

N = extensió (nombre d'ocurrències)
IRLR = índex de riquesa lèxica relativa: (L/N) x 100 [*lemma-token ratio*]

autor	*obra*	*any*	*N*	*V (formes)*	*lemes*	*IRLR*	*tipus*	*classe*	*PDA*
Jaume Bofill i Mates	Sàtires	1929	23.169	6.419	4.746	20,48	P	Or	C
Prudenci Bertrana	L'ós benemèrit i altres bèsties	1932	21.981	6.210	4.430	20,15	N	Or	C
Jaume Bofill i Mates	La montanya d'amethystes	1908	21.296	6.042	4.283	20,11	P	Or	C
Joan Perucho	Monstruari fantàstic	1976	20.786	5.787	3.855	18,54	N	Or	C
Llorenç Riber	La minyonia d'un infant orat	1935	21.523	5.613	3.828	17,78	N	Or	B
Lluís Nicolau d'Olwer	El pont de la mar blava	1928	24.158	6.393	4.203	17,39	A	Or	C
Ignasi Ribera i Rovira	Iberisme	1907	20.919	5.339	3.496	16,71	A	Or	C
Llorenç Riber	Els IV llibres de les Geòrgiques	1917	23.647	6.216	3.938	16,65	P	Tr	B
Lluís Nicolau d'Olwer	Comentaris (1915-1917)	1920	22.404	5.716	3.703	16,52	A	Or	C
Aurora Bertrana	Paradisos oceànics	1930	23.007	5.617	3.723	16,18	N	Or	C
Agustí Calvet	Hores viatgeres	1926	22.627	5.431	3.637	16,07	A	Or	C
Joaquim Ruyra	Les coses benignes	1925	24.775	6.024	3.959	15,97	N	Or	C
Narcís Oller	L'escanya-pobres	1884	22.548	5.664	3.600	15,96	N	Or	C
Víctor Català	La Mare Balena	1920	24.138	5.416	3.772	15,62	N	Or	C
Llorenç Villalonga	Mort de dama [1931]	1931	23.776	5.835	3.693	15,53	N	Or	B
Joan Fuster	El descrèdit de la realitat	1955	22.908	5.029	3.476	15,17	A	Or	V
Francesc Pineda i Verdaguer	L'espasa trencada	1936	23.741	5.490	3.576	15,06	N	Or	C
Prudenci Bertrana	El fantasma de Montcorb	1930	20.240	4.668	3.049	15,06	T	Or	C
Josep M. Barberà i Canturri	Lo prodigi del sigle	1868	28.959	6.872	4.359	15,05	P	Or	C
Francesc Trabal	Quo vadis, Sànchez?	1931	21.993	4.712	3.262	14,83	N	Or	C
Enric Moreu-Rey	El pro i el contra dels Borja	1958	20.287	4.584	3.006	14,81	A	Or	C
Bernat Morales	Idilis llevantins	1910	24.971	5.761	3.698	14,80	N	Or	V
Narcís Oller	Cróquis del natural	1879	21.295	5.218	3.125	14,67	N	Or	C
Jacint Verdaguer	L'Atlàntida	1877	23.229	5.691	3.401	14,64	P	Or	C
Josep Iglésies	Segarrenques	1917	29.439	6.722	4.305	14,62	P	Or	C
Delfí Abella	Tòtems actuals	1960	22.783	4.659	3.321	14,57	A	Or	C

autor	*obra*	*any*	*N*	*V (formes)*	*lemes*	*IRLR*	*tipus*	*classe*	*PDA*
Joaquim Ruyra	L'idil·li d'en Temme	1920	26.505	5.916	3.844	14,50	N	Or	C
Pere Calders	Gent de l'alta Vall seguit de tres reportatges especials	1957	27.453	5.941	3.925	14,29	N	Or	C
Jaume Bofill i Mates	L'altra concòrdia	1930	25.605	5.296	3.654	14,27	A	Or	C
Antoni Careta i Vidal	Euras	1882	20.177	4.830	2.831	14,03	P	Or	C
Gabriel Maura	Aygo-forts	1892	28.819	6.477	4.008	13,90	N	Or	B
Blai Bonet	Has vist, algun cop, Jordi Bonet, Ca N'Amat a l'ombra?	1976	20.276	4.176	2.810	13,85	P	Or	B
Xavier Casp	Proses en carn	1952	23.124	4.729	3.138	13,57	N	Or	V
Carme Riera	Jo pos per testimoni les gavines	1977	20.185	4.407	2.728	13,51	N	Or	B
Ofèlia Dracs	Deu pometes té el pomer	1980	29.586	6.070	3.984	13,46	N	Or	C
Carles Salvador	Les festes de Benassal	1952	26.312	5.389	3.538	13,44	N	Or	V
Josep Pla	Madrid	1933	28.632	5.875	3.862	13,36	N	Or	C
Josep Carner	La inútil ofrena	1924	24.522	5.173	3.273	13,34	P	Or	C
Josep Maria Poblet	Retorn	1942	29.866	6.031	3.964	13,27	N	Or	C
Alfons Maseras	Interpretacions i motius	1919	29.768	6.072	3.940	13,23	A	Or	C
Carles Riba	Sis Joans	1928	26.812	5.331	3.488	13,00	N	Or	C
Josep Pijoan	El meu don Joan Maragall	1927	20.007	4.051	2.590	12,94	N	Or	C
Francesc Puig-Espert	Nits d'hivern	1919	26.773	5.451	3.451	12,88	N	Or	V
Josep Carner	El Ben Cofat i l'altre	1951	20.143	4.122	2.572	12,76	T	Or	C
Gaietà Vidal i Valenciano	La familia del Mas dels Salzers	1880	23.036	5.010	2.912	12,64	N	Or	C
Rossend Arús	Cartas a la dona	1877	27.711	6.164	3.494	12,60	P	Or	C
Josep M. de Sagarra	El mal caçador	1916	20.226	4.150	2.545	12,58	P	Or	C
Prudenci Bertrana	Les ales d'Ernestina	1921	20.665	4.069	2.600	12,58	T	Or	C
Carles Riba	L'ingenu amor	1924	26.482	5.073	3.323	12,54	N	Or	C
Francesc Fayos Antony	Plansons	1882	23.874	5.250	2.978	12,47	N	Or	V
Joan Barceló i Cullerés	Ulls de gat mesquer	1979	23.997	4.790	2.989	12,45	N	Or	NR
Jordi Coca	El detectiu, el soldat i la negra	1980	22.392	4.099	2.777	12,40	N	Or	C
Jacint Verdaguer	Canigó	1886	28.351	5.908	3.508	12,37	P	Or	C
Josep Martí i Folguera	Versos catalans	1893	21.531	4.417	2.602	12,08	P	Or	C
Agustí Bartra	L'arbre de foc	1946	21.435	4.314	2.585	12,05	P	Or	C
Tomàs Junoy	Compendi de la historia de Espanya	1839	25.210	5.336	2.984	11,83	P	Or	C
Antoni Cayrol	Contalles de Cerdanya	1961	28.190	5.284	3.311	11,74	N	Or	S
Josep Pous i Pagès	Al marge de la revolució i de la guerra	1937	24.309	4.150	2.845	11,70	A	Or	C
Joan Oliver	L'òpera de tres rals	1963	22.187	4.212	2.597	11,70	T	Tr	C
Antoni Ferrer i Codina	Lo pagés de l'Ampurdá ó flors trasplantadas	1875	21.605	4.443	2.479	11,47	T	Or	C
Agustí Esclasans	La filosofia de Josep Torras i Bages	1950	20.562	3.490	2.357	11,46	A	Or	C
Jordi Pàmias	Camí de mort	1979	20.270	3.629	2.301	11,35	T	Or	NR
Josep Pin i Soler	La Sirena	1891	23.307	4.788	2.641	11,33	T	Or	C
Baltasar Porcel	Solnegre	1961	28.376	5.310	3.214	11,32	N	Or	B
Pep Subirós	Mites i raons de la modernitat	1984	26.447	4.421	2.994	11,32	A	Or	C
Josep Ferrater Mora	Els mots i els homes	1970	21.807	3.611	2.450	11,23	A	Or	C
Eduard Valentí i Fiol	Dels deures, I	1938	21.888	3.905	2.424	11,07	N	Tr	C
Maria del Pilar Maspons i Labrós	Narracions y llegendas	1874	28.106	5.365	3.099	11,02	N	Or	C

autor	*obra*	*any*	*N*	*V (formes)*	*lemes*	*IRLR*	*tipus*	*classe*	*PDA*
Carles Capdevila	L'home i les armes	1934	23.537	4.142	2.585	10,98	T	Tr	C
Mateu Obrador Bennàssar	Devers dels homens	1877	21.874	4.130	2.400	10,97	A	Tr	B
Antoni Careta i Vidal	Narracions estranyes	1905	29.913	5.530	3.284	10,97	N	Or	C
Josep Ferrater Mora	El llibre del sentit	1948	22.784	3.666	2.499	10,96	A	Or	C
Joaquim Torres-García	Diálegs	1915	21.316	3.844	2.306	10,81	A	Or	E
Adrià Gual	La pobra Berta	1909	21.273	4.014	2.297	10,79	T	Or	C
Gaietà Vidal i Valenciano	La pubilla del Mas de Dalt	1866	22.323	4.306	2.406	10,77	N	Or	C
Ventura Gassol & Joan Puig i Ferreter	Peer Gynt	1936	23.429	4.215	2.515	10,73	T	Tr	C
Francesc Trabal	Judita	1930	26.379	4.700	2.828	10,72	N	Or	C
Conrad Roure	Lo castell y la masia	1891	21.511	4.356	2.307	10,72	T	Or	C
Josep M. de Sagarra	Els ocells amics	1922	20.760	3.492	2.216	10,67	N	Or	C
Magí Morera Galícia	Coriolà	1915	29.877	5.711	3.184	10,65	T	Tr	NR
Manuel de Pedrolo	D'ara a demà	1982	25.559	4.366	2.715	10,62	T	Or	NR
Gabriel Bas	Diàlegs de Carmelites	1964	29.401	4.928	3.120	10,61	T	Tr	E
Carles Pi i Sunyer	La corda greu	1937	29.954	4.940	3.174	10,59	A	Or	C
Manuel de Pedrolo	Les mosques	1965	23.113	4.165	2.439	10,55	T	Tr	NR
Josep Ferrater Mora	Les formes de la vida catalana	1944	21.439	3.221	2.254	10,51	A	Or	C
Manuel Figuerola i Aldrofeu	L'esca del pecat	1890	25.747	4.915	2.684	10,42	N	Or	C
Maria Barbal	Pedra de tartera	1985	23.589	3.993	2.454	10,40	N	Or	NR
Marçal Olivar	El militar fanfarró	1949	20.340	3.514	2.112	10,38	T	Tr	C
Carles Cardó	Qüestions naturals, III	1959	28.874	5.090	3.000	10,38	N	Tr	C
Frederic Soler	La banda de bastardía	1882	22.451	4.220	2.330	10,37	T	Or	C
Víctor Balaguer	Don Joan de Serrallonga	1868	22.648	4.038	2.338	10,32	T	Or	C
Frederic Soler	Lo comte l'Arnau	1900	25.380	4.694	2.616	10,30	T	Or	C
Pere Coromines	Jardins de Sant Pol	1927	27.582	4.437	2.817	10,21	A	Or	C
Joan Oliver	Les tres germanes	1972	21.149	3.608	2.139	10,11	T	Tr	C
Josep M. Benet i Jornet	Fantasia per a un auxiliar administratiu	1970	23.679	4.045	2.378	10,04	T	Or	C
Frederic Soler	Batalla de reynas	1887	20.564	3.719	2.025	9,84	T	Or	C
Isa Tròlec	Ramona Rosbif	1976	21.720	3.602	2.132	9,81	N	Or	V
Josep M. de Sagarra	L'hostal de la Glòria	1931	22.321	3.860	2.167	9,70	T	Or	C
Josep M. Galdácano	Evangeli de sant Lluc	1933	24.142	4.094	2.338	9,68	N	Tr	C
Josep Vallverdú	En Roc drapaire	1971	26.857	4.371	2.557	9,52	N	Or	NR
Anònim	Evangeli segons Mateu	1979	22.147	3.738	2.104	9,50	N	Tr	E
Alexandre Ballester	Siau benvingut	1967	21.196	3.315	2.011	9,48	T	Or	C
Josep M. Folch i Torres	La ventafocs	1920	20.763	3.282	1.864	8,97	T	Or	C
Anònim	La Xinxa	1869	29.890	4.791	2.668	8,92	N	Or	E
Santiago Rusiñol	L'auca del senyor Esteve	1917	23.424	3.681	2.076	8,86	T	Or	C
Santiago Rusiñol	La bona gent	1906	24.534	3.899	2.120	8,64	T	Or	C
Josep Maria Millàs-Raurell	Anna Christie	1930	26.801	3.835	2.280	8,50	T	Tr	C
Josep Carner	Alícia en Terra de Meravelles	1927	29.215	4.090	2.454	8,39	N	Tr	C
Josep M. de Sagarra	La filla del Carmesí	1929	21.267	3.154	1.759	8,27	T	Or	C
Frederic Soler	La dida	1872	25.057	4.135	2.022	8,06	T	Or	C
Àngel Guimerà	Jesús que torna	1917	25.923	3.778	2.038	7,86	T	Or	C

autor	*obra*	*any*	*N*	*V (formes)*	*lemes*	*IRLR*	*tipus*	*clas-se*	*PDA*
Josep-Melcior Prat & Ramon Busanyà	Los fets dels apóstols	1832	23.564	3.743	1.846	7,83	N	Tr	C
Jaume Olives Canals	Diàlegs, VII. Fedó	1962	29.689	3.964	2.206	7,43	A	Tr	C
Àngel Guimerà	La filla del mar	1900	23.644	3.136	1.536	6,49	T	Or	C