Notandum **48 set-dez 2018 CEMOrOC-Feusp / IJI-Univ. do Porto**

DOI: http://dx.doi.org/10.4025/notandum.48.1

El procés de constitució del Corpus Informatitzat de la Gramàtica del Català Modern (CIGCMod). Objectius, criteris i avaluació[1]

Jordi M. Antolí Martínez
Universitat d'Alacant, ISIC/IVITRA

Resum: Aquest article descriu i avalua el procés de constitució del *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCatMod), un corpus històric, diacrònic i general, que reuneix textos en llengua catalana datats entre els anys 1601-1832 i constituït per a fornir de dades als redactors de la futura *Gramàtica del Català Modern*. En concret, en les pàgines que segueixen es contextualitza el corpus atenent els objectius amb què ha estat creat i es reflexiona sobre els conceptes de representativitat, equilibri i diversitat, aplicats a un corpus textual històric i, en concret, de català modern. Partint d'aquesta reflexió, es descriu el procés de constitució que s'ha seguit i els criteris que s'ha aplicat en la selecció dels textos. Finalment, partint de l'estat actual del corpus, s'aplica un instrument de control a fi d'avaluar la diversitat de la mostra i de fer una projecció que permeta saber el nombre de mots i tipus de textos necessaris per a culminar el procés de constitució.
Mots clau: Català modern, corpus diacrònic, Lingüística de Corpus, representativitat

Abstract: In this essay I describe and evaluate the constitution process of the *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCatMod). This is a historical, general and diachronic corpus that contains Catalan texts from 1601 to 1832. This *Corpus* was created to proportionate data to the researchers and authors of the future *Gramàtica del Català Modern*. More concretely, in this essay I present a contextualization of this corpus, along with some thoughts about concepts such as representitivity, equilibrium, and diversity, all of them applied to a historical corpus and, more particularly, to modern Catalan. I also describe the criteria used to select the texts that have been incorporated on this *Corpu*s. Last but not least, I apply a control tool in order evaluate the diversity of the sample presented and discussed here, and to evaluate what still needs to be done in order to complete the constitution of this *Corpus*.
Keywords: modern Catalan, diachronic corpus, corpus linguistics, representitivity

1. Introducció

1.1. El projecte del Corpus Informatitzat de la Gramàtica del Català Modern

En l'actualitat, al si de l'Institut Superior d'Investigació Cooperativa IVITRA (ISIC-IVITRA) de la Universitat d'Alacant (UA) i amb finançament de l'Institut d'Estudis Catalans (IEC) i del Ministeri de Cultura del Govern espanyol es prepara un nou corpus textual informatitzat sota la direcció del prof. Josep Martines (UA) que vol ser representatiu del català de l'edat moderna (un corpus, doncs, històric, diacrònic i general). El *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCatMod), que des d'una perspectiva cronològica és la continuació del *Corpus Informatitzat de la Gramàtica del Català Antic* (CIGCA), reuneix textos datats entre el 1601 i fins a la data simbòlica del 1833, any en què es convé que s'inicia la Renaixença catalana i any també que es pren com a punt de partida del *Corpus Textual*

[1] Aquest estudi ha estat desenvolupat al si de l'Institut Superior d'Investigació Cooperativa IVITRA (GVA, ref. ISIC/012/042), i en el marc dels projectes de recerca "Continuación de la Gramática del Catalán Moderno (1601-1834)" (MINECO-FEDER, ref. FFI2015-69694-P), PROMETEO/2009/042 i PROMETEOII/2014/018 (Programa PROMETEU per a Grups d'Investigació en I+D d'Excel·lència, Generalitat Valenciana), PT 2012-S04-MARTINES, IEC1-15X, PR2015-S04-MARTINES, VIGROB-125, i del Grup d'Investigació en Tecnologia Educativa en Història de la Cultura, Diacronia lingüística i Traducció (Universitat dAlacant [Ref. GITE-09009-UA]).
Vull agrair de tot cor a Lydia Moragón la seua col·laboració en el tractament estadístic de les dades d'aquest estudi.

Informatitzat de la Llengua Catalana (CTILC) de l'IEC. Una vegada constituït, l'objectiu primer d'aquest corpus serà fornir de dades lingüístiques als investigadors que redactaran una futura *Gramàtica del Català Modern*. Això no obstant, atès el fet que el corpus serà d'accés obert a la comunitat científica, esdevindrà el principal referent per als investigadors que vulguen aconseguir dades sobre el català de l'edat moderna.

L'any 2013 començà la tasca de constitució del corpus, un procés que ha anat acompanyat d'una reflexió sobre els processos i els criteris aplicats (Martines Peres & Sánchez López, 2014; Sánchez López, 2013; Sánchez López & Antolí Martínez, 2014a; Sánchez López & Antolí Martínez, 2014b). En aquest punt, després de superar els 5,5 milions de mots i els 240 textos introduïts, el volum de dades de què es disposa fa possible l'avaluació del procés que s'ha seguit amb l'objectiu d'encarar la fase final del projecte. Així, doncs, ha calgut fer una reflexió sobre els conceptes de representativitat, equilibri i diversitat aplicats a corpus (vegeu Sánchez López, en aquest mateix volum), així com dissenyar i implementar un instrument per a avaluar, a posteriori, la diversitat de la mostra acumulada. Aquesta informació ha estat valuosa a fi d'estimar el volum que haurà de tenir el corpus perquè oferisca resultats significatius als investigadors i d'aconseguir criteris objectius per a determinar les característiques dels textos necessaris.

1.2. Representativitat, equilibri i diversitat en els corpus lingüístics

D'acord amb Stefanowitsch (2017, 30) «to be representative of a particular language, the distribution of linguistic phenomena (words, grammatical structures, etc.) would have to be identical to their distribution in the language as a whole (or in the variety under investigation [...]).» Segons això, la mostra que integre el CIGCMod hauria de representar exactament la distribució en el català de l'edat moderna els fenòmens lingüístics objecte d'estudi. En els darrers anys, comença a haver-hi consens en la comunitat científica que aquest principi de representativitat és inabastable (vegeu, per al cas concret dels corpus històrics, Kabatek 2013 o Torruella 2016). Stefanowitsch hi oposa quatre raons: *a)* per a aconseguir un corpus representatiu, hauríem de saber a priori la distribució dels fenòmens a estudiar en la llengua real; *b)* hauríem de saber també en quina mesura s'usa cada varietat de la llengua (territorial, cronològica, de registre i social) i la representen; *c)* consegüentment, hauríem de poder determinar el pes que correspon a cadascuna de les varietats segons l'ús que en fa la comunitat; *d)* hauríem de tenir una mostra de totes les varietats lingüístiques. Aconseguir un corpus representatiu no equival, doncs, a aportar una mostra percentualment similar de textos de tipologies textuals diferents o de varietats territorials diferents.

Als arguments anteriors, aplicables als corpus lingüístics en conjunt, cal afegir un seguit de variables que introdueixen un biaix en la mostra i que cal controlar per a constituir un corpus de llengua antiga i, específicament, de català antic i modern: *a)* no coneixem amb detall quines eren i quina extensió tenien les varietats de la llengua; *b)* conservem unes mostres textuals que no han estat generades *ad hoc* per a la constitució del corpus; això implica, per exemple, una infrarepresentació de determinades varietats (dels textos de registres col·loquials, per exemple); *c)* l'expansió de l'espanyol en els àmbits formals (científic, literari...), literaris en el XVII i en la resta des del XVIII, limita, específicament en el cas del català, la disponibilitat de mostres de

totes les varietats; *d)* sols disposem d'edicions de part dels materials conservats, moltes vegades d'aquells que han interessat més als historiadors i als filòlegs –fet que introdueix un nou biaix a la mostra–; els textos d'edat moderna han merescut una atenció molt menor respecte dels medievals; *e)* dels textos editats, no tots acompleixen els criteris de transcripció que els fan aptes per a l'estudi lingüístic. I, finalment, cal tenir també en compte *f)* que, com apunta J. Kabatek (2013, 9): «la lengua, aunque solo se manifieste en textos, no es la suma de textos sino algo distinto»; segons això, els corpus històrics no poden aspirar més que a conèixer indirectament la llengua antiga (Kabatek, 2013, 15), la qual cosa és important si tenim en compte que el fet més freqüent és que la innovació lingüística aparega en la llengua oral i que, sols posteriorment, arribe a manifestar-se de forma escrita.

Si acceptem que l'equilibri o la representativitat són inabastables en la constitució d'un corpus, podem concloure, com fa Stefanowitsch, que la solució realista és aplicar un criteri diferent a aquests: el de la diversitat, pel qual el corpus ha d'aspirar a ser el més divers possible i, doncs, ha de representar el major nombre de varietats possible de la llengua a estudiar. En el cas dels corpus històrics com el CIGCMod, caldrà afegir, al criteri anterior, que el corpus no podrà ser representatiu del català modern en abstracte, sinó que ho serà de la llengua dels textos (escrits) conservats.

Els conceptes anteriors, representativitat, equilibri i diversitat, són clau a l'hora de definir el llindar a partir del qual la mostra acumulada és suficient. Segons Stefanowitsch (2017), un corpus «must be large enough to contain sufficiently large samples of every grammatical structure, vocabulary item, etc.»; d'això es desprèn que el volum del corpus oscil·larà segons *a)* la llengua que es vulga representar; no és el mateix un corpus de llengua general en diacronia com el CIGCMod, que un corpus d'un llenguatge d'especialitat; i *b)* la finalitat del corpus; el propòsit pel qual es constitueix el corpus és un factor determinant a l'hora de preveure'n la grandària (Nelson, 2010, 54): un corpus constituït amb la finalitat d'estudiar fenòmens freqüents, com són els que tracta una gramàtica, serà necessàriament menys voluminós que aquell constituït amb una finalitat lexicogràfica.

1.3. L'estudi

En el marc del projecte definit (§1.1) i partint dels plantejaments teòrics anteriors (§1.2), aquest estudi pretén de descriure i d'avaluar el procés de constitució del CIGCMod a fi d'orientar-ne la fase final. En concret, s'avaluaran els resultats assolits en el moment actual de desenvolupament del corpus (definit a §2.1), partint d'una previsió dissenyada a priori (una tipologia de textos que atén la variació prevista per al cat. modern; §2.2) amb un instrument de control definit a posteriori (descrit a §2.3). Després de mostrar els resultats d'aplicar aquest instrument, se'n farà una anàlisi i s'obtindran unes projeccions (§3) per establir el volum de dades i tipus de textos necessaris per cloure la constitució del CIGCMod.

2. Metodologia

2.1. El procés de constitució

El procés de constitució del corpus del CIGCMod (2013-2017) ha superat ja tres de les cinc fases previstes:

FASE 1. L'any 2013 s'elaborà el llistat inicial de les obres (50 textos del període 1601-1833) que havia de constituir el corpus essencial del CIGCMod. Aquest procés de selecció fou dirigit pel prof. Vicent Martines (UA, IEC) i comptà amb la col·laboració d'un comitè científic integrat pels profs. Vicent J. Escartí (Universitat de València, UV), Antoni Ferrando (UV, IEC), Joan Miralles (IEC, Universitat de les Illes Balears), Joan R. Ramos (Universitat de València), Joan Peytaví (IEC, Universitat de Perpinyà) o Manuel Pérez Saldanya (IEC, UV), entre d'altres. A l'hora d'establir aquesta selecció inicial, es dissenyà una tipologia de textos que atenia tota la variació prevista per al cat. modern (vegeu §2.2) i que responia al principi segons el qual «representativeness and balance can be attempted by carefully stratifying the corpus beforehand» (Nelson, 2010, 60).

FASE 2. En un segon moment, aquest llistat inicial fou ampliat amb nous textos escollits per l'equip tècnic, a fi de completar la mostra. Amb aquesta ampliació s'ha arribat a superar els 240 textos i els 5,5 milions de paraules. D'aquest volum de mots, més de 4.750.000 són en català i els restants són en llatí o castellà, especialment.

FASE 3. En el moment actual, el procés de constitució es troba en una fase d'avaluació en la qual s'ha valorat la validesa de la mostra que integra actualment el corpus –en aquest article se'n mostren els resultats– i s'ha demanat, novament, a un comitè d'experts que emeten uns informes a fi de poder considerar l'ampliació del corpus. Alguns d'aquests informes són accessibles en aquest mateix monogràfic.

De l'avaluació del corpus en l'estat actual i dels informes dels experts derivaran les dues fases darreres:

FASE 4. Tancament de la mostra de textos del corpus essencial.

FASE 5. Una vegada s'haurà tancat la constitució del corpus essencial, es procedirà a constituir el corpus de control del CIGCMod, aquell que permetrà contrastar els resultats del corpus essencial i aconseguir més dades en cas que la mostra no siga suficient.

2.2. Criteris de constitució del corpus

A l'hora de seleccionar la mostra que integra el CIGCMod, en primer lloc es van fixar uns requisits formals, vinculats amb la qualitat de la mostra i determinats pel tipus d'edició que s'haja fet modernament del text. Així, doncs, són seleccionats textos preferiblement editats complets i no fragmentàriament, i en edicions que necessàriament han de ser fidels a l'original. Els criteris de transcripció dels textos determinen la validesa de la mostra en els corpus històrics en tant que les intervencions dels editors en l'original poden conduir «a la circularidad del estudio de la lengua de una época creada en parte por los investigadores mismos» (Kabatek, 2013, 10). En la mesura que el CIGCMod haurà de servir per a fornir de dades els redactors de

la futura *Gramàtica del Català Modern*, no es podran fer servir textos regularitzats ortogràficament, fet que els inhabilita per a estudis d'ortografia o de fonètica, i, menys encara, amb una modernització lèxica, morfològica o sintàctica. Els criteris de transcripció del català que s'ajusten més bé a aquestes exigències són els de la col·lecció «Els Nostres Clàssics» de l'editorial Barcino (Martines Peres, 1999, 90-98).

Partint del requisit anterior, i atesa la màxima de diversitat a què ha d'aspirar el CIGCMod, segons s'ha dit anteriorment (§1.2), en la FASE 1 del procés es definí, a priori, un primer model amb les obres que hauria de tenir el corpus atenent als tipus de variació prevista per al català modern. D'acord amb la premissa que la llengua es manifesta mitjançant les seues varietats, per representar una llengua en tota la seua diversitat, s'ha d'incloure una mostra de les diferents varietats lingüístiques que integra, i no sols de la varietat més culta d'aquesta. D'aquesta manera, a l'hora de seleccionar els textos que havien de constituir el CIGCMod, es va partir d'aquells factors que, a priori, podien haver generat variació en català modern: el canvi a què es troba sotmesa qualsevol llengua natural amb el pas del temps (variació diacrònica), la diversitat geogràfica (variació diatòpica), la variació dels usos lingüístics que es dona entre grups o classes socials (variació diastràtica) i la variació segons àmbits d'ús (variació diafàsica). Aquests factors de variació van permetre dissenyar unes tipologies textuals, cadascun dels tipus de les quals hauria d'estar suficientment representat en el corpus. Tot seguit aprofundirem en aquestes tipologies.

2.2.1. Variació diacrònica

Es va establir, seguit el model del CIGCA, que convindria tenir una mostra suficient per segments de 50 anys. Això equival a 4 segments complets (segles XVIIa, XVIIb, XVIIIa i XVIIIb) i un segment incomplet, de 1800-1832, que clou cronològicament la mostra.

2.2.2. Variació diatòpica

El segon criteri de selecció dels textos ha estat el territorial, amb l'objectiu que la mostra represente tots els parlars catalans. Amb aquest fi, hem pres com a referència la tipologia dels parlars catalans, àmpliament acceptada, de Joan Veny publicada en *Els parlars* (1978):

Català oriental	Rossellonès o català septentrional
	Català central
	Balear
	Alguerès
Català occidental	Català nord-occidental
	Valencià

Taula 1. Distribució dels dialectes dels dos grups dialectals del català, segons Veny (2002, 23)

A més de la tipologia anterior, s'ha tingut en compte durant el procés de selecció de textos diversos subdialectes amb una gran entitat, a fi d'evitar sobrerepresentar parlars com el valencià central o el català central (barceloní i gironí, en concret), afavorits pel gran volum de textos

conservats i editats. És el cas del valencià septentrional: els parlars de la Plana de Castelló, del Baix Maestrat i, ja de transició cap al tortosí, els parlars dels Ports, l'Alt Maestrat i l'Alcalatén, segons els descriu Colomina (1999); i també del valencià meridional, que, segons Colomina, comprèn els parlars catalans del sud del Xúquer: d'una banda, els parlars de la Costera, la Vall d'Albaida, la Safor, l'Alcoià-Comtat i les Marines; de l'altra, els parlars més meridionals, de l'Alacantí, les comarques del Vinalopó i, en aquest període encara també, el Baix Segura, que està representada al corpus amb textos editats per B. Montoya (Montoya, 1986; Montoya, 2013; Cremades & Montoya, 2012).

Quant al català nord-occidental, s'ha volgut representar la transició amb l'aragonès que constitueix el ribagorçà, i no s'inclou –com és criteri general– el benasquès com a parlar català, tot i que diversos estudis en consideren la catalanitat (Cahner & Lloret, 1971; Babia, 1997). També ha estat atès el tortosí (que abraça, seguint la classificació de Veny 2002, el Baix Ebre, el Montsià, la Terra Alta i la Ribera d'Ebre).

Pel que fa al català oriental, s'ha considerat la necessitat de representar el català septentrional de transició cap al rossellonès, inclòs en la tipologia de Veny (1978) i que abraça el sector més septentrional de l'Alt Empordà, la Garrotxa, el Ripollès i tota la Cerdanya. Aquests parlars de transició, tot i que actualment han viscut un anivellament amb el català oriental general, tenien plena vigència en el període estudiat. També s'ha tingut la cura de representar els altres parlars balears diferents al mallorquí (menorquí i eivissenc i formenterer).

2.2.3. Variació diafàsica i diastràtica

Una mostra diversa quant a la tipologia textual permet, en els corpus de llengua antiga, reflectir la variació diafàsica i diastràtica de la llengua del passat (vegeu, com a mostra, la relació que estableix Montoya 2009 entre tipologia textual i registres en català antic). La inclusió de processos judicials, com ja argumentà Montoya (1989), o de dietaris personals i obres de literatura popular garantirà que, en la mesura del possible, el corpus reflectisca la llengua col·loquial del moment. D'altra banda, la incorporació de tractats cientificotècnics, per exemple, permet conèixer els llenguatges d'especialitat i, en particular, la terminologia de l'època. És per això que cal aconseguir una mostra representativa dels tipus de textos existents en el període de referència.

A l'hora de definir aquesta tipologia, s'han tingut en compte les tipologies dels altres corpus diacrònics disponibles en el moment de constitució (el CICA en català i el CORDE en espanyol), així com també la tipologia elaborada per J. Miralles a l'*Antologia de textos de les Illes Balears* (Vols. I i II 2006).[2] D'això resulta una tipologia que pretén de ser significativa en el procés de constitució. En la Taula 2 sistematitzem la tipologia textual dissenyada per a la constitució del CIGCMod. A l'hora d'oferir les dades als usuaris, a més d'aquestes categories, es pretén posar a l'abast una informació més detallada sobre la tipologia textual de cada document d'acord amb la proposta de Miralles (2006), a fi que els investigadors puguen

[2] Excloem, de la tipologia de Miralles, els textos crítics, etnogràfics i polítics. L'*Antologia* no en recull exemples dels segles XVII i XVIII i són absents també de la nostra mostra.

constituir subcorpus segons els seus interessos concrets i comparar els resultats amb els d'altres corpus.

1. Textos cientificotècnics
2. Textos literaris
3. Textos pedagògics
4. Textos juridicoadministratius
5. Textos historiogràfics, epistolaris i dietaris
6. Obres gramaticals i lexicogràfiques
7. Textos filosòfics, religiosos i morals

Taula 2. Tipologies textuals del CIGCatMod

2.2.4. Disseny inicial

Partint dels criteris anteriors, es definí una classificació a priori de les obres que haurien de constituir el corpus. Aquesta classificació establia 83 tipus de documents amb una caracterització específica segons el tipus de varietat lingüística. De cadascun dels tipus, s'establí el criteri d'aconseguir una mostra mínima (però no màxima) de 20.000 mots, la qual equival a un text d'extensió mitjana i ha estat considerada la quantitat mínima de mots per a representar un tipus textual; segons Oostdijk (1991, 50): "A sample size of 20.000 words would yield samples that are large enought to be representative of a given variety". El disseny tenia en compte que hi havia tipologies textuals o varietats geogràfiques per a les quals seria difícil aconseguir reunir una mostra suficient, de manera que se'n rebaixava la previsió. D'altra banda, es va fer un tractament particularitzat de l'alguerès, atesa la seua singularitat com a illa lingüística de dimensions reduïdes.

<table>
<tr><th rowspan="2">Tipologia textual</th><th colspan="5">Varietat territorial</th><th rowspan="2">Període</th></tr>
<tr><th>NO</th><th>V</th><th>S</th><th>C</th><th>B</th></tr>
<tr><td>1. Textos cientificotècnics</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>XVIIa-1832</td></tr>
<tr><td rowspan="6">2. Textos literaris</td><td rowspan="2">6</td><td>7</td><td rowspan="2">9</td><td>10</td><td rowspan="2">12</td><td>XVIIa</td></tr>
<tr><td>8</td><td>11</td><td>XVIIb</td></tr>
<tr><td rowspan="2">13</td><td>14</td><td rowspan="2">16</td><td>17</td><td rowspan="2">19</td><td>XVIIIa</td></tr>
<tr><td>15</td><td>18</td><td>XVIIIb</td></tr>
<tr><td>20</td><td>21</td><td>22</td><td>23</td><td>24</td><td>1800-1832</td></tr>
<tr><td>3. Textos pedagògics</td><td colspan="5">25</td><td>XVIIa-1832</td></tr>
<tr><td rowspan="6">4. Textos juridicoadministratius</td><td rowspan="2">26</td><td>27</td><td rowspan="2">29</td><td>30</td><td rowspan="2">32</td><td>XVIIa</td></tr>
<tr><td>28</td><td>31</td><td>XVIIb</td></tr>
<tr><td rowspan="2">33</td><td>34</td><td rowspan="2">36</td><td>37</td><td rowspan="2">39</td><td>XVIIIa</td></tr>
<tr><td>35</td><td>38</td><td>XVIIIb</td></tr>
<tr><td>40</td><td>41</td><td>42</td><td>43</td><td>44</td><td>1800-1832</td></tr>
<tr><td rowspan="6">5. Textos historiogràfics, epistolaris i dietaris</td><td rowspan="2">45</td><td>46</td><td rowspan="2">48</td><td>49</td><td rowspan="2">51</td><td>XVIIa</td></tr>
<tr><td>47</td><td>50</td><td>XVIIb</td></tr>
<tr><td rowspan="2">52</td><td>53</td><td rowspan="2">55</td><td>56</td><td rowspan="2">58</td><td>XVIIIa</td></tr>
<tr><td>54</td><td>57</td><td>XVIIIb</td></tr>
<tr><td>59</td><td>60</td><td>61</td><td>62</td><td>63</td><td>1800-1832</td></tr>
<tr><td>6. Obres gramaticals i lexicogràfiques</td><td>64</td><td>65</td><td>66</td><td>67</td><td>68</td><td>XVIIa-1832</td></tr>
<tr><td rowspan="6">7. Textos filosòfics, religiosos i morals</td><td rowspan="2">69</td><td rowspan="2">70</td><td rowspan="2">71</td><td rowspan="2">72</td><td rowspan="2">73</td><td>XVIIa</td></tr>
<tr><td>XVIIb</td></tr>
<tr><td rowspan="2">74</td><td rowspan="2">75</td><td rowspan="2">76</td><td rowspan="2">77</td><td rowspan="2">78</td><td>XVIIIa</td></tr>
<tr><td>XVIIIb</td></tr>
<tr><td>79</td><td>80</td><td>81</td><td>82</td><td>83</td><td>1800-1832</td></tr>
</table>

Taula 3. Tipus textuals previstos per a la constitució del CIGCMod

Aquest disseny inicial volia ser un instrument orientatiu per a la constitució del corpus, que establia una mostra mínima a fi de garantir el criteri de diversitat, però que no establia una mostra màxima per tipus. Si s'assumeix (segons les reflexions de l'apartat 1.2) que no és possible ni potser significatiu assolir una mostra equivalent entre els tipus (amb percentatges similars en volum de mots per a cadascun dels tipus) ni proporcional (amb una mostra percentualment proporcional a la població que es vol representar), es conclou que la mostra màxima dependrà de la quantitat de textos conservats i editats i, en darrer terme, del pes que li atorgue el comitè d'experts. Se segueix, doncs, el model de corpus com el CIGCA, el CORDE (vegeu el manual de consulta, al web del corpus) o el CTILC (Rafel, 1998, VII-IX; Soler i Bou, 2002).

2.3. Instrument d'avaluació del corpus

A l'hora de dur a terme l'avaluació prevista en la FASE 3 (§2.1), s'ha definit un instrument amb què examinar, a posteriori, la mostra de dades acumulada. Tenint en compte que es tracta

d'un corpus sense paral·lels en llengua catalana, no s'hi ha pogut utilitzar un instrument d'avaluació i comparació de caràcter relatiu, en què es contrasten les dades del CIGCMod amb les d'un corpus similar (vegeu els experiments que, en aquest sentit, proposen Schäfer & Bildhauer 2013). Partint d'aquest condicionament, s'ha dut a terme una anàlisi de l'evolució, per obra introduïda, del percentatge de novetat i de repetició de les formes (*types*) respecte del volum de dades acumulat.

En aquest mateixa línia, Corpas i Seghiri (2007) van desenvolupar l'algoritme N-Cor per a calcular a posteriori el llindar de representativitat dels corpus, partint del concepte de densitat lèxica. En concret, la seua proposta es fixa en l'increment de la densitat lèxica del corpus amb l'observació, text a text, dels *types* acumulats en relació amb els *tokens* acumulats. L'objectiu és veure fins a quin punt l'afegiment de nous *tokens* implica un augment significatiu de *types*.

L'instrument que s'ha implementat en aquest estudi és més senzill i no estudia la densitat lèxica del corpus, sinó que se centra a analitzar les formes (*types*): es contrasta, de cada text, els *types* que presenta amb els *types* acumulats al corpus per a veure el percentatge de *types* no registrats que hi havia al text (no el percentatge de *types* nous que aporta al conjunt del corpus). Aquests *types* s'afigen a l'acumulat i es repeteix el procés amb les dades del text següent. D'aquesta manera, els percentatges de novetat o repetició de *types* no són relatius al conjunt del corpus i en relació amb el volum global de *tokens*, sinó relatius als textos individuals. En altres paraules, no estudiem l'evolució de la densitat lèxica del corpus, sinó simplement el nombre de *types* registrats i no registrats de cada text que introduïm al corpus.

Segons aquest instrument, el corpus haurà assolit la màxima diversitat possible com més s'aproximarà a zero el percentatge de novetat i a cent el de repetició. Parteix, doncs, de les premisses que: *a)* els tipus de variació definits en el punt 2.2 es manifesten gràficament. El fet de treballar amb una llengua sense una ortografia estandarditzada i en part sense un model de llengua culta, com és el cas del català modern, fa viable aquesta premissa per a molts dels fenòmens susceptibles d'estudi. I *b)* sempre quedarà un percentatge de formes no registrades, especialment antropònims i topònims, o mots amb problemes de digitalització que cal esmenar posteriorment.

Aquest instrument té, però, limitacions evidents: *a)* Té en compte solament aquells fenòmens que impliquen variació formal, i aquest fet exclou la variació semàntica. *b)* En tant que es pren com a unitat d'estudi la forma, no es té en compte la variació sintàctica. *c)* L'aparició d'un sol registre d'una forma no és, sempre, suficient per a l'investigador.

Amb tot, a fi de superar el biaix que introdueix l'instrument, una vegada tancat el procés de constitució es duran a terme cerques de prova per part d'experts i, com ja s'ha avançat, es constituirà un corpus de control en el qual els investigadors podran, en cas que ho necessiten, augmentar el nombre de resultats.

3. Anàlisi de resultats

3.1. Distribució de les dades d'acord amb el disseny inicial

A finals de l'any 2017, CIGCMod ha superat els 5,5 milions de mots i els 240 textos. La distribució de les dades (nombre de mots) d'acord amb el disseny inicialment previst (Taula 3) és la següent:

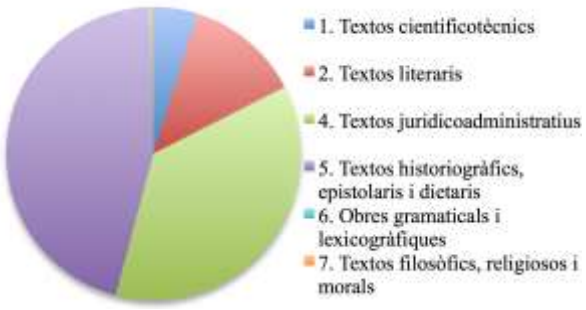
Tipologia textual	Total	Varietat territorial					Període
		NO	*V*	*S*	*C*	*B*	
1. Textos cientificotècnics	269.146	-	6.107	-	145.338	117.701	XVIIa-1832
2. Textos literaris	696.505	21.077	72.532	-	324.408	-	XVIIa
			6.003				XVIIb
		28.702	56.023	17.232	80.971	80.449	XVIIIa
			-		3.600		XVIIIb
		-	1.280	-	4.228	-	1800-1832
3. Textos pedagògics	-	-					XVIIa-1832
4. Textos juridicoadministratius	2.019.524	269.362	179.554	17.728	403.403	88.431	XVIIa
			84.215		57.782		XVIIb
		94.839	121.736	-	350.832	-	XVIIIa
			26.698		291.977		XVIIIb
		-	17.770	-	15.197	-	1800-1832
5. Textos historiogràfics, epistolaris i dietaris	2.493.923	3.820	407.585	104.745	385.504	-	XVIIa
			141.363		51.841		XVIIb
		19.216	417.334	14.848	373.047	33.069	XVIIIa
			0		243.338		XVIIIb
		-	11.200	-	287.013	-	1800-1832
6. Obres gramaticals i lexicogràfiques	16.862	-	16.862	-	-	-	XVIIa-1832
7. Textos filosòfics, religiosos i morals	17.035	-	9.067	-	-	-	XVII
		-	7.675	-	-	293	XVIII
		-	-	-	-	-	1800-1832
Total	5.512.995	437.016	1.583.004	154.553	3.018.479	319.943	Total

Taula 4. Nombre de mots incorporats al CIGCMod per tipus textual

A banda, cal tenir en compte que el corpus conté 93.856 mots corresponents a l'alguerès.

Si s'observa la distribució dels mots ara en correlació solament amb un dels criteris de variació s'obté la distribució següent:

Tipus de text	Xifres absolutes	%
1. Textos cientificotècnics	269.146	4,88
2. Textos literaris	696.505	12,63
3. Textos pedagògics	-	-
4. Textos juridicoadministratius	2.019.524	36,63
5. Textos historiogràfics, epistolaris i dietaris	2.493.923	45,24
6. Obres gramaticals i lexicogràfiques	16.862	0,31
7. Textos filosòfics, religiosos i morals	17.035	0,31
Total	5.512.995	100

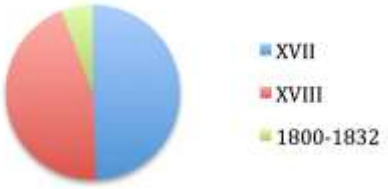


Taula 5 i Figura 1. Distribució percentual del nombre de mots segons la tipologia textual dels documents

Varietat geogràfica	Xifres absolutes	%
NO	437.016	7,93
V	1.583.004	28,71
S	154.553	2,80
C	3.018.479	54,75
B	319.943	5,80
Total	5.512.995	100

Taula 6 i Figura 2. Distribució percentual del nombre de mots segons la varietat territorial del català a la qual s'adscriuen els textos

Període	%
XVII	49,31
XVIII	44,75
1800-1832	5,94
Total	100



Taula 7 i Figura 3. Distribució percentual del nombre de mots segons la varietat territorial del català a la qual s'adscriuen els textos

Segons les dades anteriors, per cloure la constitució del CIGCMod caldria completar 44 dels tipus definits (10 amb una mostra insuficient i 34 sense mostra) amb nous textos que responguen als criteris següents:

a) Des de la perspectiva de la tipologia textual, caldria ampliar la representació dels tipus següents: *1.* Textos cientificotècnics; *6.* Obres gramaticals i lexicogràfiques; i *7.* Textos filosòfics, religiosos i morals. D'altra banda, caldria aconseguir una mostra de *3.* Textos pedagògics.
b) Si es té en compte la variació diatòpica, caldria ampliar la mostra del català nord-occidental i balear, i especialment del rossellonès.
c) Des de la perspectiva cronològica, caldria ampliar solament la mostra del període 1800-1832.

D'acord amb aquesta anàlisi, si tenim en compte el nombre mínim de paraules proposades per tipus (20.000) i el nombre de tipus a completar, caldria un mínim de 880.000 mots més per a tancar el corpus essencial del CIGCMod.

3.2. Avaluació de les dades amb l'instrument de control

De l'aplicació de l'instrument de control presentat abans s'obtenen els resultats següents. En la Taula 8 es mostren els resultats per trams de 250.000 mots:

Trams	**Mitjana de mots**		**Total acumulat (*tokens*)**	**Total acumulat (*types*)**
	Nous	*Coincidents*		
1	69,24	30,76	250.000	30.084
2	56,54	43,46	500.000	61.258
3	39,37	60,63	750.000	100.792
4	37,41	62,59	1.000.000	105.100
5	31,44	68,56	1.250.000	122.248
6	39,03	60,97	1.500.000	143.738
7	25,69	74,31	1.750.000	172.576
8	27,9	72,1	2.000.000	199.675
9	20,48	79,52	2.250.000	228.758
10	23,38	76,62	2.500.000	247.237
11	32,66	67,34	2.750.000	258.511
12	28,98	71,02	3.000.000	276.246
13	33,69	66,31	3.250.000	302.142
14	26,7	73,3	3.500.000	305.643
15	20,07	79,93	3.750.000	331.521
16	22,79	77,21	4.000.000	356.075
17	18,15	81,85	4.250.000	381.959
18	16,65	83,35	4.500.000	415.749
19	15,67	84,33	4.750.000	456.185
20*	16,25	83,75	4.960.693	469.945

*El tram núm. 20 conté una mostra incompleta, d'acord amb l'estat del corpus al mes de desembre de 2017.

Taula 8. Evolució del percentatge de novetat i de repetició per trams de 250.000 mots

Segons les dades anteriors, i com il·lustra la Figura 4 (per trams de 250.000 mots) i la Figura 5 (per document incorporat), amb tan sols mig milió de paraules el percentatge de formes repetides en els textos superà el de formes noves. Aquesta primera fase d'increment molt ràpid del percentatge de formes repetides acabarà en assolir els 1.250.000 mots; des de llavors i fins als 4 milions de mots observem una segona fase en què la tendència de decreixement percentual de les formes no registrades es veu interrompuda puntualment. Finalment, a partir dels quatre milions de mots s'obre una tercera fase en què s'observa una tendència de decreixement del percentatge de novetat més estable i constant. Aquestes tres fases es podrien relacionar amb *a)* la caiguda ràpida del percentatge de novetat s'ha d'associar amb la introducció dels mots

gramaticals, formes verbals i substantius més freqüents. *b)* en una segona fase es manifesta la incorporació, molt més lenta, del paradigma gràfic i flexiu dels mots no tan freqüents; la introducció de textos corresponents a tipus diferents (Taula 3) explica la oscil·lació; *c)* en una tercera fase, el diccionari del corpus ja conté la major part de les formes de la llengua i en tota la variació gràfica i, doncs, s'estabilitza amb un índex de novetat al voltant del 15%, que correspon fonamentalment a formes gràfiques poc freqüents, termes d'especialitat o noms propis.

Les figures 4 i 5 representen el diagrama de dispersió que s'obté de les dades de repetició i novetat. La Figura 4 recull gràficament les dades de la Taula 8: a l'eix de les abscisses trobem el percentatge de novetat o de repetició i a l'eix de les ordenades trobem el nombre de *tokens* acumulats, organitzats en blocs de quart de milió de mots. Els punts del diagrama representen cadascun dels valors que hem obtingut en l'anàlisi de les dades del corpus: els blaus representen el percentatge de repetició dels *types* per blocs de quart de milió de paraules, i els rojos el percentatge de novetat dels *types*. Per la seua banda, la Figura 5 representa aquest percentatge ara per text introduït i no per volum acumulat de *tokens*. Es pot avançar que, en tots els casos, el la distribució de les dades apunta cap a una funció hiperbòlica.

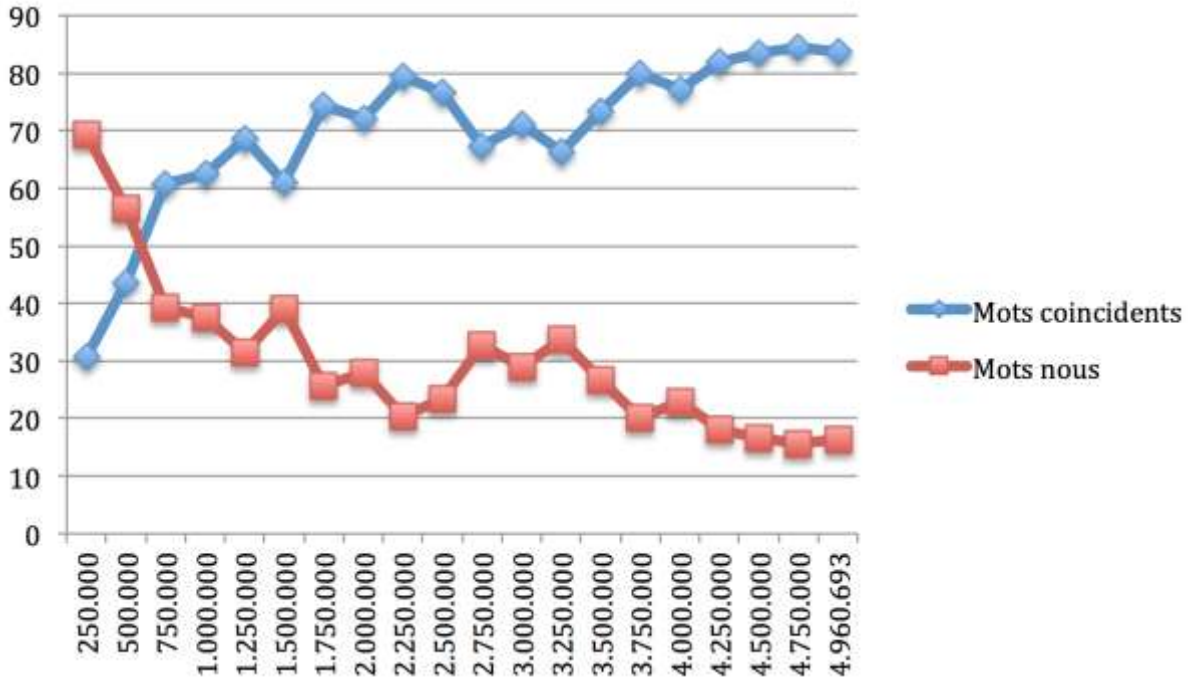


Figura 4. Representació gràfica de l'evolució del percentatge de novetat i de repetició per trams de 250.000 mots

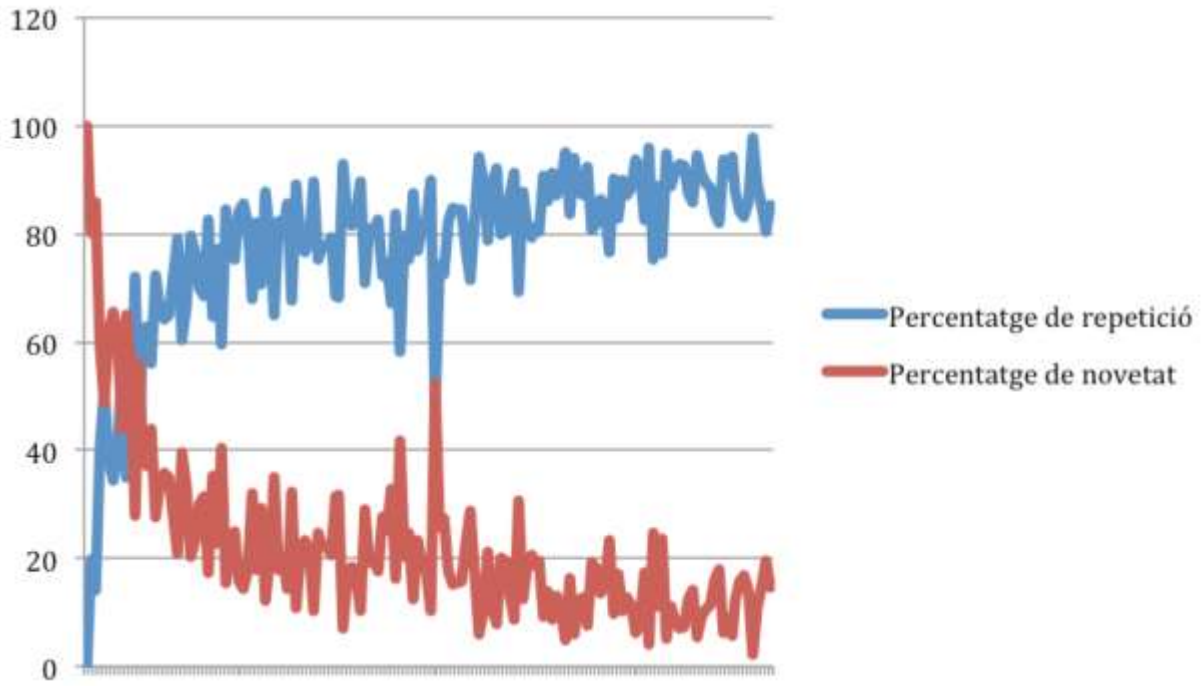


Figura 5. Representació gràfica de l'evolució del percentatge de novetat i de repetició per obra introduïda

El volum de dades acumulat permet fer una estimació puntual de l'increment necessari per a considerar que el corpus representa suficientment el català modern. Entenem que aquest punt s'assoleix quan el diccionari del corpus ja conté la major part dels mots (inclosos els paradigmes gràfic i flexiu d'aquests) del català modern (amb tota la variació funcional, cronològica, territorial i social que integra) conservats en els textos. D'entrada, cal assumir que no s'assolirà mai un percentatge de repetició mitjà del 100%; aquest objectiu, que no és realitzable en corpus de llengua contemporània (Corpas i Seghiri 2007), ho és encara menys en els corpus històrics. Això s'explica per la presència de formes poc freqüents del paradigma gràfic o flexiu d'un mot, d'hàpaxs, de noms propis poc freqüents (com els exotopònims i antropònims d'altres llengües) o d'errors del manuscrit o de digitalització. Sols com a mostra, segons una aproximació que s'ha realitzat a partir de les dades del CIGCMod, els noms propis i errors del manuscrit o de digitalització representen aproximadament l'11% del total de formes no reconegudes dels darrers 10 textos introduïts; això equival aproximadament al 2% dels *types* d'aquest bloc de textos.

Per a fer aquesta estimació, ha calgut realitzar una modelització de les dades de repetició anteriors (preses en relació amb el nombre de *types* acumulats), de la qual resulta una primera hipèrbola (Figura 6, en roig). Amb tot, els càlculs realitzats per a avaluar la fiabilitat del model han aconsellat d'introduir un paràmetre de control per a mantenir la tendència a 1 de la hipèrbola, amb un creixement $c<1$. D'acord amb les estimacions, el model que millor s'ajusta a les dades és $c=0'7$ (Figura 6, en negre), en el qual el creixement és més lent; això indica que en percentatges alts de repetició, per a incrementar mínimament el percentatge caldrà incrementar molt el nombre de *types*.

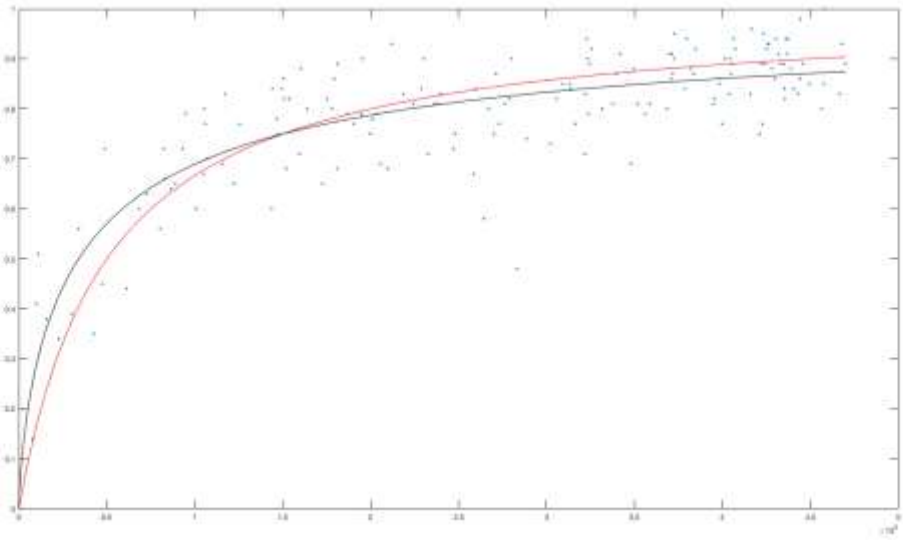


Figura 6. Hipèrbola inicial (en roig) i hipèrbola controlada (en negre) que s'obté de modelitzar les dades de repetició de les obres del CIGCMod

Segons el model, amb els 469.945 *types* acumulats el percentatge de repetició mitjà és del 88'56% i amb aquesta mateixa quantitat de *types* podem assegurar que el percentatge de repetició mínim és del 76'18%, amb una confiança del 95%. L'objectiu que s'ha proposat és determinat el volum de *types* necessaris perquè el CIGCMod assolisca el llindar desitjable d'un percentatge, mínim i/o mitjà, del 90% de repetició. Segons el model, d'acord amb les dades actuals, per a assolir un percentatge mitjà del 90% de repetició caldria un total de 572.240 formes acumulades (de 469.945; un increment del 17,87% respecte del volum actual del CIGCMod); aquest volum garantiria un mínim de repetició del 78'9%. En segon lloc, el percentatge mínim del 90% s'assoliria amb un total de 1.847.700 formes acumulades (la qual cosa representa, a més, un percentatge mitjà del 95'67%). Aquest darrer objectiu, amb tot, no es considera viable, atès el fet que quadruplicar el volum del CIGCMod no és factible per la limitació de textos editats en català del període, més encara si es vol respectar el model de distribució definit en la Taula 3.

4. Conclusions

En aquest estudi s'ha volgut descriure i avaluar el procés de constitució del Corpus Informatitzat de la Gramàtica del Català Modern (CIGMod). Els principis teòrics i metodològics, així com els principals resultats que s'hi ha mostrat són els següents:

a) Els conceptes de representativitat o equilibri són difícils d'aplicar a la constitució de corpus lingüístics, més encara si són històrics. Per això s'ha optat, en el cas del CIGCMod, d'acord amb Stefanowitsch (2017), per aplicar el principi de la diversitat: el corpus haurà de contenir una mostra com més diversa millor, a fi de representar tota la variació, en aquest cas, del català modern.

b) La constitució d'un corpus ha d'integrar també l'avaluació de la mostra. El procés de constitució del CIGCMod ha superat dues de les cinc fases previstes: després de fer-ne un disseny a priori inicial i de processar els textos seleccionats, s'ha assolit la tercera fase (d'avaluació), que anirà seguida per una ampliació de la mostra i la constitució d'un corpus de control que complemente el corpus essencial.
c) Un corpus general ha d'atendre tota la variació de la llengua. Per això, per dur a terme la selecció dels textos del CIGCMod, es va elaborar un model que integrava 83 tipus de documents amb una caracterització diferenciada, en la qual es combinen com a criteris la variació diacrònica, diatòpica, i, per mitjà de les tipologies textuals, la variació diastràtica i diafàsica. Per a cadascun dels tipus de document a representar es va establir una mostra mínima –no màxima– de 20.000 mots.
d) D'acord amb l'instrument de control que s'ha aplicat, el CIGCMod es troba ja en la fase final del procés de constitució. En aquest moment el CIGCMod compta amb 5,5 milions de mots i més de 240 textos. Per a avaluar la validesa de la mostra, s'ha definit un instrument de control; en concret, s'ha observat, per text i per volum de mots acumulat (*token*), l'evolució percentual de formes (*types*) noves i ja registrades al corpus. Segons aquest instrument, el corpus haurà assolit la màxima diversitat possible com més s'aproximarà a zero el percentatge de novetat i a cent el de repetició. Segons els darrers registres, amb un volum de 4.750.000 mots (*tokens*), el percentatge de novetat és tan sols del 15,67%, mentre que el de repetició és del 84,33%. Partint de la tendència constatada, amb la modelització de les dades i la determinació de la fórmula hiperbòlica que les relaciona, s'ha pogut determinar que, amb un total de 469.945 formes acumulades, el corpus haurà superat el llindar (un percentatge de repetició mitjà del 90% de repetició) a partir del qual es podrà considerar que ofereix resultats significatius. Això equival a un increment del 17,87% respecte del volum actual del CIGCMod).

Bibliografia

Babia, A. *La Franja de la Franja. La parla de la Vall de Benasc, on el català és patuès*. Barcelona: Empúries, 1997.

Cahner, M.; Lloret, T. "Benasquès". En: *Gran Enciclopèdia Catalana*. Barcelona: Edicions 62, 1971. Pàg. 408.

Colomina i Castanyer, J. *Dialectologia catalana: introducció i guia bibliogràfica*. Alacant: Departmament de Filologia Catalana, Universitat d'Alacant, 1999.

Corpas Pastor, G.; Seghiri Domínguez, M. "Determinación del umbral de representatividad de un corpus mediante el algoritmo N-Cor". *Procesamiento del lenguaje natural*, n. 39, pàg. 165-172, 2007.

Kabatek, J. "¿Es posible una lingüística histórica basada en un corpus representativo?" *Iberoromania*, n. 77, pàg. 8-28, 2013.

Martines Peres, V. *L'edició filològica de textos*. València: Publicacions de la Universitat de València, 1999.

Martines Peres, V.; Sánchez López, E. "L'ISIC-IVITRA i el metacorpus CIMTAC. Noves aportacions a la lingüística de corpus". *Estudis Romànics*, n. 36, pàg. 423-436, 2014.

Miralles i Montserrat, J. *Antologia de textos de les Illes Balears*. Vol. I. Segles XIII-XVI. Vol. II. Segles XVII-XVIII. Barcelona: Institut d'Estudis Baleàrics/Publicacions de l'Abadia de Montserrat. 2006.

Montoya Abat, B. *Variació i desplaçament de llengües a Elda i a Oriola durant l'edat moderna*. Alacant: Diputació d'Alacant, 1986.

---. "Un repte per a la lingüística històrica: copsar la llengua parlada del passat". *Caplletra: revista internacional de filologia*, n. 6, pàg. 71-88, 1989.

---. "Tipologia textual i de registres en el català antic". En: PEREZ SALDANYA, M.; MARTINES PERES, J. (Ed.). *Per a una gramàtica del català antic*. Alacant: Institut Interuniversitari de Filologia Valenciana, 2009. Pàg. 73-85

---. "L'ús comercial de la costa marítima d'Oriola: edició i notes onomàstiques i lingüístiques sobre el text d'un plet amb la ciutat d'Alacant (1643)". *Scripta: revista internacional de literatura i cultura medieval i moderna*, n. 2, pàg. 139-169, 2013.

Montoya Abat, B.; Cremades Rodríguez, F. "Rerefons lingüístic i mèdic d'un text sobre l'epidèmia de 1678 a Oriola (1: Rerefons lingüístic)". *Estudis Romànics*, vol. 34, pàg. 165-207, 2012.

Nelson, M. "Building a written corpus. What are the basis?" En: O'KEEFFE, A., & McCARTHY, M. (Eds.). *The Routledge Handbook of Corpus Linguistics*. New York: Routledge Handbooks, 2010. Pàg. 53-65.

Oostdijk, N. *Corpus Linguistics and the Automatic Analysis of Englis*h. Amsterdam/Atlanta: Editions Rodopi, 1991.

Rafel, J. *Diccionari de freqüències*. Barcelona: Institut d'Estudis Catalans, 1998. Pàg. VII-IX.

Sánchez López, E. *Estudi de la llengua d'Ausiàs March a través de les col·locacions. Una aproximación semiautomàtica*. Boston/Berlín: De Gruyter, 2013.

Sánchez López, E.; Antolí Martínez, J. "M. El Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis (CIMTAC)". *Estudios Hispánicos*, n. XXII, Pàg. 117-121, 2014a.

---. "L'exonímia en el Corpus Informatitzat Multilingüe de Textos Antics i Contemporanis (CIMTAC)". En: Casanova, E.; Payà, E. (Eds.). *Topònims entre dos llengües. L'exonímia, una manifestació de la globalitat*. València: Denes, 2014b. Pàg. 157-165.

Schäfer, R.; Bildhauer, F. *Web Corpus Construction*. Morgan and Claypool, 2013.

Soler i Bou, J. "El Corpus Textual Informatitzat de la Llengua Catalana". *Hizkuntza-corpusak. Oraina eta geroa*, n. 24/25, Pàg. 1-12, 2002.

Stefanowitsch A. *Corpus Linguistics: A Guide to the Methodology* [Esborrany]. 2017. Disponible a Internet: http//:goo.gl/Ysihgk

Torruella Casañas, J. "Tres propuestas en el ámbito de la lingüística de corpus". En: Kabatek, J. (Ed.). *Lingüística de corpus y lingüística histórica iberorrománica*. Berlin/Boston: Walter de Gruyter, 2016. Pàg. 90-113

Veny i Clar, J. *Els parlars: síntesi de dialectologia catalana*. Barcelona: Dopesa 2, D.L. 1978.
---. *Els parlars catalans: (síntesi de dialectologia)*. Palma de Mallorca: Moll, 2002.

Recebido para publicação em 08-02-18; aceito em 09-03-18