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

**DOI: http://dx.doi.org/10.4025/notandum.48.6**

# Panorama històric de la constitució de corpus: orígens, consolidació i expansió[1]

Elena Sánchez López
Universitat d'Alacant / ISIC/IVITRA

**Resum:** La constitució de corpus ha experimentat una gran evolució des dels seus inicis fins a l'actualitat. A fi d'entendre millor les bases de la disciplina, farem un breu recorregut per la seua història a través d'algunes fites importants. El nostre camí partirà dels orígens, amb el corpus *BROWN*. Després recorrerà l'etapa de consolidació, període que representa una època de reflexió i de creació de nombrosos corpus, com ara el *British National Corpus*. Fins a arribar a l'actualitat, on s'ha produït una gran expansió de l'activitat de compilació de corpus, que ha provocat, al seu torn, noves reflexions i propostes de millora en la disciplina. En el nostre article, palesarem l'estreta relació entre corpus i TIC. Posteriorment, presentarem la definició actual de corpus lingüístic, que inclou els trets d'autenticitat, representativitat i gran extensió. Dedicarem l'epígraf següent al disseny de corpus, on revisarem els conceptes de representativitat estadística i equilibri, així com l'aplicació real de la noció de representativitat als corpus (tant de llengua actual, com de llengua antiga). Per últim, proposarem la diversitat i l'extensió com a aproximacions acceptables per a assolir el major grau de representativitat possible.
**Paraules Clau**: Lingüística de corpus, TIC, corpus lingüístics, representativitat.

**Abstract**: The constitution of corpus linguistics has experienced a big transformation since its origins. In order to better understand the basis of this discipline, I will present a brief history of it. I will start with the corpus BROWN. I will also explain the consolidation of the discipline, period in which a number of corpus was created, such as the *British National Corpus*. I will also describe the situation of this discipline nowadays, in which we have seen plenty of new corpora being created, along with new theoretical insights. My next step will be to present the close relationship between ICT and corpus and to define what is understood by corpus linguistic in today's world, in which concepts such as authenticity, representativity, etc., are particularly relevant. After this, I will explain how corpus are designed and I will apply concepts such as statistical representativity and equilibrium, along with the real application of the notion of representativity to corpus (both in current and ancient language). Finally, I propose diversity and extension as acceptable approximations to get the biggest level or representativity possible.
**Keywords**: corpus linguistics and Corpora, ICT, representativity.

## 1. Introducció

### *1.1. Evolució de la lingüística de corpus*

La lingüística de corpus ha experimentat una gran evolució des dels seus inicis fins a l'actualitat. A fi d'entendre millor les bases de la disciplina, farem un breu recorregut per la seua història a través d'algunes fites importants. Per a poder resseguir aquest camí, hem de

[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]).

tenir en compte que la lingüística de corpus està configurada per tres grans elements indissolublement lligats: l'objecte, que són els corpus; la metodologia, basada en les Tecnologies de la Informació i la Comunicació (TIC) i els objectius, que estan marcats per la recerca lingüística. Aquests tres constituents han evolucionat de forma paral·lela i interrelacionada, esperonant-se mútuament, fins a arribar l'aliança relativament estable que formen en l'actualitat.

El punt de partida de la compilació de corpus se situa en els anys 60, amb la creació del primer corpus informatitzat, el Brown corpus. La llengua anglesa encetava un camí, que després seguirien altres llengües, entre d'altres les llengües romàniques. Durant els anys 80 i 90, es va promoure molt la creació de corpus lingüístics, especialment amb fins lexicogràfics. Entre alguns corpus que es van constituir durant aquesta època, podem destacar el *Collins Corpus* (1980-85) dins del projecte COBUILD (Collins Birmingham University International Language Database), que havia de servir com a base per a la redacció del *COBUILD Dictionary* (1987); el *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC) (1985-1997), com a referència per a la compilació del *Diccionari de la Llengua Catalana* de l'Institut d'Estudis Catalans (DIEC) i el *Corpus de Referencia del Español Actual* (CREA) (1998), constituït per la Real Academia Española, amb l'objectiu de servir com a documentació per al *Diccionario de la Lengua Española*. Malgrat aquests projectes destacats, amb gran suport institucional, John Sinclair (1996), director del projecte COBUILD i membre del projecte EAGLES, considerava que en aquella època els "corpus electrònics eren una cosa nova".

Posteriorment, amb el canvi de segle i de mil·lenni, es va generalitzar la creació de corpus, no tan sols amb fins lexicogràfics, sinó també per a dur a terme estudis lingüístics. A aquest període pertanyen el *Corpus Informatitzat del Català Antic* (CICA), el *Corpus Informatitzat de la Gramàtica del Català Antic* (CIGCA), el *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCMod), el *Corpus del Nuevo Diccionario Histórico del Español* (CNDHE), la versió anotada del *Corpus de Referencia del Español Actual* (CREA), el *Corpus del español del siglo XXI* (CORPES XXI), *O corpus do português*, *Basa Textuala per la lenga d'Òc* (BaTelÒc) o el *PAISÀ* (corpus de l'italià contemporani), així com els diversos corpus creats per Mark Davies[2] de la Brigham Young University.

De manera sintètica, podríem considerar que els anys 60-70 van comportar l'inici de la compilació de corpus, els 80-90 van suposar la consolidació de la metodologia i, a partir del 2000, es va produir l'expansió de la lingüística de corpus en un doble vessant, el constituïdor i l'investigador. Tot i que no es tracta d'una relació exhaustiva, aquesta breu línia temporal resulta indicativa pel que fa a l'expansió dels corpus en relació amb la lingüística, ja que ens

[2] Des del 2002, Mark Davies ha posat a disposició dels usuaris el *Corpus del Español*, el *British National Corpus*, el *Corpus do Português*, el *TIME Corpus of American English*, el *Corpus of Contemporary American English* (COCA), *Corpus of Historical American English* (COHA), *Google Books Corpus*, *Corpus of American Soap Operas*, *Strathy Corpus* (Canadian English), *Corpus of Global Web-Based English* (GloWbE), *Wikipedia corpus*, *Hansard corpus*, *CORE corpus*, *NOW corpus*, *US Supreme Court corpus*, *Early English Books Online* (EEBO) corpus. Podeu trobar més informació sobre aquests corpus en https://corpus.byu.edu/corpus.asp.

mostra uns inicis amb voluntat compiladora, que es van obrir a una fase lexicogràfica i que van desembocar en una utilitat lingüística en general.

*1.2. Corpus i TIC*

L'eclosió de la lingüística de corpus no ha estat casual i ha anat sempre de la mà de les Tecnologies de la Informació i la Comunicació (TIC). L'evolució de les TIC durant les últimes dècades, especialment pel que fa a l'emmagatzematge de dades, el processament del llenguatge natural i Internet, ha tingut un gran impacte en la lingüística de corpus. En primer lloc, l'increment i la facilitat que s'ha assolit pel que fa a l'emmagatzematge de dades, ha permès fer uns corpus cada vegada més grans. En segon lloc, les novetats procedents del processament del llenguatge natural, han facilitat l'etiquetatge dels corpus, tant automàtic com manual, fet que ha incrementat la versatilitat dels corpus. En aquest cas, s'ha produït una relació de doble direcció, ja que les aportacions del camp de la lingüística computacional han aprofitat per a augmentar les capacitats dels corpus i els corpus han servit com a base per a testar les eines de processament de la llengua i han contribuït de manera decisiva a la seua millora.

La Comunitat Europea, conscient de la importància creixent del processament informàtic del llenguatge natural, va finançar en els anys 90 (1993-95) un projecte de recerca, EAGLES (Expert Advisory Group on Language Engineering Standards), amb l'objectiu de fixar uns estàndards internacionals pel que fa als recursos lingüístics "a gran escala", entre els quals es troben els corpus lingüístics. John Sinclair va liderar el grup que s'hi dedicava i va assentar les bases actuals per a la constitució de corpus lingüístics.

En tercer lloc, l'aparició i universalització d'Internet ha tingut una gran influència en els corpus des d'una doble perspectiva. D'una banda, ha condicionat la manera en què accedim als corpus, ja que molts d'ells estan disponibles a través d'Internet, amb unes eines de consulta bastant potents. Aquest fet ha comportat la generalització de l'accés, que abans estava molt més limitat a causa del suport físic en què es desaven els corpus. D'una altra, ha posat a la nostra disposició una gran quantitat de mostres de llengua, tant en formats més tradicionals com ara els llibres o els diaris, com en nous formats, que inclouen els textos publicats en xarxes socials, els missatges d'audio o de vídeo.

Com hem vist, tots els avanços que s'han produït en l'àmbit de les TIC han tingut una gran influència en la constitució i utilització dels corpus. En una última fase, la que podríem considerar d'expansió, fins i tot han "democratitzat" la creació de corpus, ja que han posat al nostre abast les eines (el software) i la matèria primera (els textos) per a dur-la a terme de manera individual. Aquest fet ha provocat que qualsevol persona tinga l'oportunitat de generar el seu corpus ad-hoc, ajustat a uns interessos de recerca concrets, i explotar-lo de la manera més productiva segons el seu criteri. De manera que uns projectes que en els seus inicis necessitaven d'un gran suport institucional i, fins i tot empresarial (si pensem en el cas del COBUILD), en l'actualitat es poden emprendre a petita escala o a títol individual. Tot i

que som conscients d'aquest canvi i de la seua influència en el conjunt de la disciplina, en aquest article ens centrarem en els corpus a gran escala amb suport institucional.

## 2. Concepte de corpus

### *2.1. Definició general*

El concepte de corpus també ha anat evolucionant al llarg dels anys i ha pres matisos diversos segons la disciplina en què s'inserira. Mentre que en literatura es concebia tradicionalment com un conjunt d'obres o bé d'un autor particular, o bé d'un gènere i període concrets, en la lingüística aplicada s'entenia com un recull de dades creat per a la recerca lingüística. De manera que un dels primers reptes de la lingüística de corpus va ser definir l'abast del terme a partir de dues qüestions bàsiques: les característiques essencials que havia de tenir un recull de mostres de llengua per a ser considerat un corpus i la diferenciació entre corpus dissenyats per a reflectir l'ús general de la llengua i corpus que recullen comportaments lingüístics específics (SINCLAIR, 1996).

En l'àmbit de la lingüística de corpus s'ha arribat al consens que un *corpus* és "un recull de textos (o fragments de textos) en format electrònic que han estat seleccionats seguint criteris externs a fi de representar, en la mesura del possible, una llengua o una varietat de llengua, amb l'objectiu de fornir de dades per a la recerca lingüística" (SINCLAIR, 2005). Per tant, dins d'aquesta disciplina, textual i informatitzat són trets bàsics del concepte de corpus. Stefanowitsch (2017, 24) perfila la definició afirmant que es tracta d'un recull de "mostres de llengua en ús", que han de complir els criteris d'autenticitat, gran extensió i representativitat.

### *2.2. Trets distintius des de la visió actual: autenticitat, gran extensió i representativitat*

Un corpus lingüístic es distingeix d'altres reculls electrònics de textos en què les mostres que recull són autèntiques i han estat escollides per a ser representatives de la parcel·la de la llengua que es vol estudiar, de manera que la seua composició ha estat dissenyada seguint uns criteris determinats, que depenen d'uns objectius de recerca concrets. A més, el fet que es tracte d'un conjunt extens de textos contribueix a la potencial representativitat.

Si volem estudiar la llengua real, la primera característica que ha de complir un corpus és l'autenticitat. Compartim totalment l'opinió de Stefanowitsch (2017, 24) quant a la importància que les mostres de llengua que s'inclouen en un corpus siguen autèntiques. En aquest sentit, autèntiques significa que hagen estat creades amb el fi de la comunicació. Unes dades creades directament per a la recerca lingüística no reflectirien l'ús real de la llengua, sinó la intuïció lingüística del compilador, amb la qual cosa el corpus perdria el seu valor empíric. A més, com menys intervenció presenten els textos per part del creador del corpus, més mantindran la seua autenticitat.

L'extensió recomanada per a un corpus ha anat variant al llarg del temps i ha augmentat amb l'increment de la capacitat d'emmagatzematge informàtic. El *Corpus BROWN* (KUČERA; FRANCIS, 1967) aspirava a tenir 1 milió de paraules i en aquella època es

percebia com un miracle que es pogueren recuperar tantes paraules amb una sola ordre informàtica. Hem de recordar que es treballava amb ordinadors molt lents, que estaven orientats al processament de xifres i tenien moltes dificultats amb els caràcters alfabètics. A més, els textos s'editaven en forma de targetes perforades que es guardaven en fitxers i la informació es processava en blocs durant la nit. Tota la informació havia de picar-se a mà en uns aparells poc ergonòmics.

La irrupció dels ordinadors personals, així com dels escàners, va facilitar en gran mesura la constitució de corpus i va provocar una primera fornada de reculls informatitzats de textos, com ara la *Birmingham Collection of English Text* (1985), que ja comptava amb 20 milions de paraules o el *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC1) (1985-1997), amb 52.372.058 mots. La facilitat per a aconseguir textos informatitzats i la capacitat per a emmagatzemar textos no ha deixat de créixer, fet que ha tingut un gran impacte en la grandària dels corpus. A meitat dels 90, el *Bank of English* contenia 200 milions de mots i, en l'actualitat, n'aplega uns 4.500 milions. El *Corpus de Referencia del Español Actual* (CREA), per posar un altre exemple, en tenia més de 170 milions en la seua última versió (3.2, juny de 2008). Si parlem del català, la *continuació del Corpus Textual Informatitzat de la Llengua Catalana* (CTILC2), tot i que no tenim constància del nombre total de mots que inclourà, sabem que en va incorporar més de 3 milions tan sols en el període 2015-2016 (RAFEL I FONTANALS, 2018).

La tendència, com ens demostren les dades, és crear corpus cada vegada més extensos. Sinclair (1996) ja ho avançava quan afirmava que els corpus han de ser tan extensos com permeta la tecnologia del moment. En la dècada dels 90, la doctrina sobre el tema assumeix que la grandària, com la diversitat, contribueix a la representativitat del corpus (vegeu, per exemple, MCENERY; WILSON, 1996). La premissa bàsica és senzilla, com més paraules incorpore un corpus, més probabilitats hi haurà que la que volem estudiar es trobe entre elles.

El fet que les tecnologies actuals permeten una adquisició i emmagatzematge de dades pràcticament il·limitats, ha obert la possibilitat de constituir corpus infinits. Els corpus que no es tanquen i continuen incloent noves dades són coneguts com a corpus de referència. Un exemple n'és el *Bank of English*, que incorpora paraules noves totes els mesos. Aquest tipus de corpus resulten molt útils, per exemple, per a la identificació de neologismes. En el cas de corpus destinats a la recerca lingüística, és important gaudir de conjunts de textos tancats i amb un disseny concret i documentat, ja que serà l'única manera de poder fer una correcta interpretació dels resultats obtinguts i de poder replicar els estudis fets. De manera més gràfica, tan sols podem fer descripcions acurades d'un objecte fix, en el moment comence a canviar la seua composició, l'objecte variarà i la nostra descripció ja no serà vàlida.

La tercera característica que ha de tenir un corpus és la representativitat. En la definició de Sinclair (2005), ja es palesava que els textos d'un corpus han de ser representatius de la llengua o varietat lingüística que volem investigar. Sembla evident que, si volem estudiar el llenguatge jurídic, haurem d'introduir mostres de llenguatge jurídic. Aquesta premissa, però,

s'ha vist sovint contradita en la pràctica, quan s'ha intentat descriure el llenguatge general a partir de llenguatge purament literari (i escrit) o una llengua global a partir d'una varietat diatòpica concreta. Per a evitar aquestes temptacions, la representativitat ha esdevingut una de les nocions clau en el disseny de corpus. Però ens hem de preguntar fins a quin punt és possible representar fidelment un constructe social tan complex com és la llengua. Com que es tracta d'un tema d'importància cabdal, dedicarem l'epígraf següent a revisar la noció de representativitat i l'operacionalització que se n'ha fet al llarg dels anys en la constitució de corpus.

## 3. Disseny de corpus: representativitat, equilibri, extensió i diversitat

El disseny de corpus ha evolucionat de manera paral·lela a les eines de creació i gestió, així com dels objectius de recerca. En aquest epígraf farem atenció als avanços que s'han produït en aquest àmbit, tenint en compte que les nocions de representativitat, equilibri, extensió i diversitat han resultat centrals en aquest procés. El nostre objectiu serà veure com s'han aplicat al llarg del temps des d'una perspectiva crítica.

### *3.1. Representativitat estadística i equilibri*

La representativitat s'ha considerat tradicionalment un dels trets essencials del concepte de corpus, que el diferencia d'una mera col·lecció de textos en format electrònic (vegeu Atkins 1991). Molt lligada a aquesta noció hi ha la d'equilibri, que la complementa per a poder aplicar-la en la constitució de corpus. Podríem considerar que, mentre que la representativitat fa referència a les categories de textos que s'hi han d'incloure, l'equilibri determina la proporció en què han de ser inclosos.

Durant els anys 90, la dècada de consolidació de la lingüística de corpus, es van dedicar moltes reflexions a establir la manera correcta d'optimitzar el disseny de corpus i d'aplicar-hi la noció de representativitat (vegeu ATKINS, 1991; LEECH, 1991; CLEAR, 1992; BIBER, 1993; MANNING; SCHÜTZE, 1999). Aquests autors partien d'una perspectiva estadística, segons la qual "una mostra és considerada representativa si els resultats que ens proporciona són aplicables a la població en general" (vegeu MANNING; SCHÜTZE, 1999, 119), per tant, les mostres són "versions a escala d'una població més àmplia" (vegeu VÁRADI, 2000).

Si adoptem aquesta perspectiva pel que fa als corpus lingüístics, un corpus hauria de ser una versió reduïda de la llengua (o segment de llengua) a la qual vol representar. Es tracta d'una noció aplicable si definim una població molt concreta, com ara totes les publicacions d'un país determinat en un any concret. Per exemple el *corpus BROWN* era representatiu de tota la població de textos publicats en anglés als Estats Units l'any 1961. També resulta operacionalitzable si volem representar un segment de la llengua ben definit, que comparteix un criteris externs (situacionals) i interns (lingüístics) concrets, com ara el llenguatge jurídic escrit català actual que apareix en les pòlisses d'assegurança. Les dificultats es presenten quan

volem constituir un corpus per a la descripció de la llengua general, per exemple per a la redacció de diccionaris, gramàtiques o etiquetadors automàtics.

En les especificacions de disseny del *British National Corpus*, Atkins (1991, 7) —citant Clear (1992, 21)— palesava que seria ideal poder seguir els principis teòrics del "statistic sampling and inference", és a dir, poder fer el mostratge (la selecció de textos) seguint criteris estadístics perquè, posteriorment, es pogueren dur a terme generalitzacions vàlides a partir dels resultats obtinguts. A pesar de ser conscient de la conveniència del mètode, la mateixa investigadora reconeixia que aquests criteris no eren aplicables a l'hora de constituir corpus lingüístics, a causa de tres causes ben clares: no és possible definir la població de "llengua general", no hi ha una unitat clara de llengua per a fer el mostratge i la població de llengua és tan immensa que és impossible representar tots els seus trets "a escala".

En primer lloc, perquè els mètodes estadístics sempre se centren en poblacions (conjunts d'elements amb característiques comunes) clarament definides, mentre que en l'àmbit de la llengua és molt difícil delimitar una població total de manera rigorosa. No tan sols perquè s'haurien de definir tots els factors externs que la condicionen, sinó també els factors interns i la relació entre ells. Stefanowitsch (2017, 30) concreta encara més aquesta limitació quan afirma que, a pesar que poguèrem disposar de dades sobre les variables demogràfiques més rellevants del grup sota estudi, no podríem saber quina és la distribució dels textos en la població global, ja que no podem establir un percentatge que determine la representació de la llengua escrita versus la llengua oral, del llenguatge periodístic front al llenguatge literari o que especifique els temes més tractats en les converses. Encara més, tot i que poguérem establir quina és la distribució d'aquests subconjunts en la població de textos, seria impossible determinar si tots influeixen en el sistema lingüístic en la mateixa mesura, perquè no es pot saber la recepció que tenen. Per exemple, els e-mails són probablement una de les fonts escrites més produïda, però tan sols arriben a un grapat de persones. En canvi, les notícies tenen una audiència molt més àmplia, a pesar que se'n redacten menys. Un altre inconvenient a l'hora de definir la població de "llengua general" és que les comunitats de parlants no són homogènies, de manera que definir l'equilibri basant-nos en la proporció de tipologies textuals que presenta una comunitat, encara que fóra possible, no ens donaria una representació realista de la llengua, ja que cada membre de la comunitat participa en diferents situacions comunicatives que inclouen diverses tipologies textuals.

En segon lloc, no hi ha una unitat de la llengua clara que es puga utilitzar per a fer el mostratge i definir la població. Podem considerar que la unitat mínima de llengua són els mots, les oracions o els textos sencers, fins i tot podem defensar que es tracta de sintagmes o altres tipus d'elements. En tercer lloc, la immensitat de la població, i més encara si tenim en compte les fonts actuals i les futures, provoca que siga impossible representar tots els seus trets adequadament en la mostra. Ens trobem davant d'una població que creix dia a dia i a la que no tenim accés en la seua totalitat. De fet, la gran majoria de manifestacions de "llengua" són efímeres, ja que es tracta de produccions orals que no són recollides enlloc.

Una altra característica que haurien de complir els corpus, segons els postulats tradicionals, és la d'equilibri, molt lligada a la representativitat. Es considera que un corpus és equilibrat quan inclou una gran selecció de categories de textos que aparentment representen la llengua o varietat de la llengua que és objecte d'estudi de forma proporcional a "la realitat". L'objectiu de l'equilibri és que el corpus faça la funció d'"un model exacte a petita escala, i per tant més manejable, del material lingüístic que els creadors del corpus volen estudiar" (ATKINS et al. 1992, 6). Com deien adés, la representativitat fa referència a les categories de textos i l'equilibri a la proporció en què han de ser inclosos en el corpus.

La determinació de les proporcions per a aconseguir l'equilibri s'ha fet de maneres molt diverses. En molts casos, tan sols s'ha atés al fet que totes les categories estiguen igualment representades, per exemple, 50% de llengua escrita i 50% de llengua oral. En altres casos, que han estat precedits d'una reflexió més profunda, es tractava d'incloure les categories atenent a la seua importància en el conjunt de la llengua, tant des de la perspectiva de la producció com de la recepció. La metodologia que seguien era fer una estimació, a partir de les dades de difusió de determinats tipus de text, de les categories que influïen més en la configuració del constructe "llengua". Malgrat la bona voluntat i la validesa del procés de reflexió, és evident que, si no és possible definir la població "llengua general", no serà possible representar-la des d'una perspectiva estadística i tan poc hi haurà una manera científica de determinar quines són les proporcions per a aconseguir un corpus "equilibrat". McEnery, Xiao i Tono (2006) palesen aquesta idea d'una manera força contundent quan expressen que declarar que un corpus és equilibrat és "més un acte de fe que una exposició de fets, ja que en l'actualitat no hi ha cap instrument científic fiable que mesure l'equilibri dels corpus". A banda de la viabilitat d'assolir l'equilibri, Biber (1993, 247) se'n planteja la conveniència, ja que té la impressió que els lingüistes no estan interessats per les expressions de llengua més freqüents, sinó d'una representació que incloga tota la variació possible.

Al llarg del present punt hem argumentat que definir la població "llengua" i la seua estratificació a fi de poder fer-ne un mostratge representatiu i equilibrat és un ideal impossible d'atènyer en la realitat. Aquest fet ens porta a plantejar-nos si té sentit fer atenció al disseny d'un corpus en el procés de constitució. La resposta intuïtiva és que sí, ja que els corpus que tenen en compte la variació i les diverses tipologies textuals resulten més adients per a la recerca lingüística que aquells que no ho fan. A fi de concretar aquesta intuïció en pautes reals d'actuació farem una revisió d'alguns corpus constituïts, parant atenció a la documentació que hi ha sobre ells i que recull anys de debat i d'experiència compiladora.

### *3.2. Aplicació del concepte de representativitat en corpus*

L'aplicació del concepte de representativitat i equilibri en els corpus ha estat desigual. Els creadors dels primers, van haver de determinar mètodes per a descriure la població total i fer-ne la selecció de mostres. Com a exemple d'aquests pioners, farem una ullada a la composició del *Corpus BROWN* (1967). Posteriorment farem atenció a dos corpus de l'època de

consolidació, el *British National Corpus* (1991-1994) i l'*International Corpus of English* (1990-1998, pel que fa al segment britànic). Podrem observar que en aquesta època hi havia molt d'interés per l'argumentació i la documentació de les decisions preses en el processos de constitució, fet que ens han deixat valuoses contribucions al debat sobre la representativitat. Hem seleccionat dos corpus d'anglés coetanis per la gran diferència entre els seus objectius, ja que mentre que el primer tenia vocació de representar l'anglés britànic per a tots els usos, el segon se centrava en la comparació de la gramàtica entre les diverses varietats de l'anglés. Els tres corpus inicials són molt significatius perquè, posteriorment, com identifiquen McEnery et al (2006), molts constituïdors es van dedicar simplement a adoptar un model de corpus preexistent i a aplicar-lo al seu, assumint que el model adoptat dotaria d'equilibri a la seua obra. Com a conseqüència, el disseny del BROWN, el BNC i el ICE es poden veure reflectits en la composició d'altres corpus. Com a colofó de la descripció dels procediments que segueixen els corpus per a assolir la representativitat, farem atenció a dos corpus en llengües romàniques, el CREA i el CTILC.

### 3.2.1. Corpus BROWN (BROWN)

El *Brown University Standard Corpus of Present-Day American English* (Corpus BROWN) (1967) va ser el primer recull de textos informatitzats que es va dissenyar i constituir. A pesar del seu nom, si atenem a la composició, la població que pretenia representar aquest corpus no era l'anglés nord-americà dels anys 60 (present-day), sinó l'anglés de les fonts escrites publicades als Estats Units el 1961. Aquest corpus aplega 500 mostres de llengua d'unes 2000 paraules cadascuna, que han estat seleccionades a partir de diferents tipus de text. El primer criteri de selecció que s'hi va aplicar era de no-ficció (286 mostres), ficció (126 mostres) i premsa (88 mostres), que després se subdividia en subgèneres/temàtica i tipus de document concret. En el cas de la no-ficció, es definien temes com ara la religió, les aficions, la cultura popular, literatura, biografies i memòries, miscel·lània i ciència, amb mostres procedents de llibres, tractats, publicacions periòdiques i altres tipus de document més concrets. Pel que fa a la ficció, hi havia una primera classificació segons la temàtica, que incloïa general, de misteri i detectius, de ciència ficció, d'aventures i de l'oest, romàntica i d'humor, subdividides entre novel·les i històries curtes, afegint assajos en el cas de l'humor. En el cas de la premsa, es recollien reportatges, editorials i ressenyes, subdividits entre les seccions clàssiques d'un diari. Totes aquestes mostres estaven extretes de fonts escrites publicades als Estats Units durant el 1961.

El procés de selecció de les obres es va produir en dues fases: una classificació inicial subjectiva amb la determinació de quantes mostres s'inclourien per cada categoria i una selecció aleatòria de les mostres. La majoria de les mostres es van seleccionar a partir dels fons de la Biblioteca de la Universitat de Brown i de l'Ateneu de Providence. Tot i que, per a algunes categories, es va haver de recórrer a altres fonts (KUČERA; FRANCIS, 1967).

Durant el procés de selecció, hi havia una aspiració d'equilibri. La llista de categories principals i les subdivisions van ser consensuades a una conferència que va tenir lloc a la Universitat de Brown el 1963, on tots els participants van aportar la seua opinió sobre el nombre de mostres que hauria d'incloure cada categoria. A partir d'aquestes opinions es va calcular la mitjana i es va aplicar. Posteriorment, es van fer algunes correccions durant el procés de selecció i la subdivisió més concreta es va basar en la quantitat de publicacions reals durant 1961 (Kučera & Francis, 1967). De manera que la representativitat es volia assolir mitjançant la intuïció d'experts sobre la rellevància de subgèneres o temàtiques determinades i la proporció real de textos publicats.

Stefanowitsch (2017, 35) apunta que, a pesar de les limitacions que presenta aquest disseny, el *corpus BROWN* va gaudir de bastant èxit i va ser replicat per altres experiments com ara el *Lancaster-Oslo/Bergen Corpus* (LOB), que recollia mostres d'anglés britànic de 1961, el *Freiburg Brown* (FROWN) i el *Freiburg LOB* (FLOB) corpus d'anglés nord-americà i d'anglés britànic de 1991, el *Wellington Corpus of Written New Zealand English*, d'anglés de Nova Zelanda, i el *Kolhapur Corpus*, que recollia l'anglés de l'Índia. Aquest autor atribueix la popularitat del corpus i del seu disseny al fet que és molt útil poder estudiar corpus comparables, independentment del disseny que presenten. Entre línies, podem llegir que el fet de documentar correctament el procés de disseny i constitució d'un corpus és clau per a determinar la seua utilitat.

### 3.2.2. British National Corpus (BNC)

Els constituïdors del *British National Corpus* van tenir en compte aquesta premissa. Per a ells era molt important treballar amb una metodologia ben definida i documentar-la bé, a fi que altres investigadors i recopiladors de corpus pogueren revisar-ne, reproduir-ne o adaptar-ne el disseny (BURNARD, 2007). L'11 d'abril de 1991 van publicar un document amb els usos previstos per al corpus que estaven dissenyant (BNCW02), on es declaraven les següents àrees d'aplicació: publicació de llibres de referència, recerca lingüística acadèmica, ensenyament de llengües, intel·ligència artificial, processament del llenguatge natural, processament de veu, recuperació de la informació. En el mateix document, s'identificava que del corpus es podia extraure informació lèxica, semàntica i pragmàtica, sintàctica, morfològica i sobre ortografia. Tot i que aquestes eren les aplicacions previstes de partida, als 15 anys es va observar que el corpus i els mètodes seguits havien tingut un impacte molt més ampli (BURNARD, 2007). Es tractava d'un corpus amb objectius generals, que incloïa mostres (no textos complets) d'un màxim de 45.000 mots; de tipus sincrònic (textos de ficció de 1960 i de no ficció de 1975); general, monolingüe (d'anglés britànic) i mixt (ja que contenia mostres de llengua oral i escrita). Ateses les limitacions de temps i pressupost (BURNARD, 2007), la llengua oral (uns 10 milions de mots) constituïa un 10% del corpus i la llengua escrita el 90% restant (uns 90 milions de mots), a pesar del consens que per a un

corpus amb voluntat de representar la llengua general, la proporció de llengua oral hauria de ser molt més gran que la de llengua escrita.

Els responsables d'aquest corpus documentaven tots els passos seguits i les decisions preses durant el procés (vegeu ATKINS, 1991 i BURNARD, 2007), així com els resultats en forma d'explicitació de la composició del corpus. De manera coherent amb la impossibilitat de determinar la unitat de llengua que s'havia de mostrejar (ATKINS, 1991), totes les mostres introduïdes es mesuren atenent a 3 unitats de llengua: textos (mostres que no superaven els 45.000 mots), unitats S (equivalents a oracions, segons el sistema CLAWS) i unitats W (equivalents a mots, segons el sistema CLAWS). El primer filtre que utilitzaven per a la selecció era el tipus de text: textos orals (segons criteris demogràfics) amb un 4,3% dels mots; textos orals (amb restriccions relacionades amb la situació comunicativa, determinats pel camp) amb un 6, 27% dels mots; textos escrits publicats (llibres i publicacions periòdiques) amb un 80,55% dels mots; textos escrits per a ser parlats, amb un 1,29% dels mots; i textos escrits miscel·lanis amb un 7,56% dels mots.

Dins dels textos escrits, el segon filtre era la temàtica. A banda de les dades utilitzades per a la selecció, s'hi afegien moltes dades descriptives, com ara el tipus d'autor (múltiple, únic, etc.), el sexe, l'edat, el domicili (62,8 % desconegut, 36,05% Irlanda i Regne Unit, 0,51% país de la Commonwealth, 0,24% Europa continental, 0,34% EUA, 0,05% altres llocs), el lector ideal per edat i per sexe, el lloc de publicació (on destacava el sud del Regne Unit amb un 67,36%) i el tipus de mostra (desconegut, text sencer, mostra del començament, del mig, del final o composta). En els criteris de selecció dels títols, tenien en compte factors tant de publicació com de recepció. Per aquesta raó, la meitat dels llibres provenien de la llista *Whitaker's Books in Print* (BIP) de 1992 i la resta eren best-sellers, havien rebuts premis literaris o havien estat molt prestats per les biblioteques, indicis que assenyalen una major influència en la configuració de la "llengua".

Pel que fa als textos orals, el mostratge es va fer partint de la producció de llengua de la població de parlants d'anglés britànic del Regne Unit. La meitat dels textos es va seleccionar amb criteris demogràfics dels parlants i la resta a partir de tipologies textuals determinades pel camp. Aquesta distinció ens sembla molt pertinent ja que, com postula Stefanowitsch (2017, 30), cada membre de la comunitat participa en diferents situacions comunicatives que inclouen diverses tipologies textuals, és a dir, un mateix membre pot produir tipus de llengua diversos segons el context. Entre els factors demogràfics es tenien en compte l'edat, la classe social i el sexe. Pel que fa a les tipologies textuals determinades pel camp, s'establien diversos tipus de situacions: educatives i informatives (conferències, comentaris de notícies, interacció a classe), negocis (xarrades i entrevistes, reunions de sindicats, demostracions de vendes, reunions de negocis i consultes), públiques o institucionals (mítings, sermons, xarrades institucionals, plens de governs i ajuntaments, trobades religioses, actes parlamentàries, actes de judicis), lleure (xarrades, comentaris esportius, xarrades a clubs, intervencions telefòniques en programes de televisió i de ràdio, reunions de clubs).

### 3.2.3. International Corpus of English (ICE)

L'*International Corpus of English* va començar a constituir-se el 1990, en la mateixa època que el BNC, però amb uns objectius ben diferents. La seua finalitat era recopilar material per als estudis contrastius de l'anglés arreu del món. És important tenir en compte que el focus eren els estudis de la gramàtica. La idea va partir de Sidney Greenbaum (1988), que en una nota breu en *World Englishes* reconeixia el valor per a l'estudi de la gramàtica de tenir dos corpus d'anglés escrit com ara el BROWN (d'anglés nord-americà) i el LOB (d'anglés britànic). Segons aquest investigador, havien d'augmentar les possibilitats dels estudis comparatius informatitzats de tres maneres: recopilant les varietats d'altres països on l'anglés fora la llengua materna, com ara el Canadà i Austràlia; recopilant varietats nacionals de la llengua de països on l'anglés fora llengua cooficial i incloent anglés oral i anglés escrit no publicat, a banda de l'anglés escrit publicat (GREENBAUM, 1988).

En l'actualitat, aquest projecte aplega equips de recerca d'Austràlia, Camerun, Canadà, països de l'est d'Àfrica (Kènia, Malawi i Tanzània), Fiji, Gran Bretanya, Hong Kong, l'Índia, Irlanda, Jamaica, Kènia, Malta, Malàisia, Nova Zelanda, Nigèria, Pakistan, Filipines, Sierra Leone, Singapur, Sud-Àfrica, Sri Lanka, Trinitat i Tobago i els Estats Units. Cadascun dels equips segueix un disseny comú de corpus i el mateix sistema d'anotació, per a garantir la màxima comparabilitat entre els diversos components (NELSON et al, 2002, 3). En un futur, quan tots els mòduls estiguen acabats, el corpus contindrà més de 21 milions de mots, estarà anotat sintàcticament i podrà consultar-se mitjançant un software comú, l'ICECUP.

El disseny de cadascun dels mòduls inclou 500 textos d'uns 2.000 mots. Podríem considerar que el primer filtre que s'hi aplica és el de procedència geogràfica, ja que cada país tan sols recull textos del seu propi país. En segon lloc, el filtre del mitjà, segons el qual per cada país es recullen 300 textos pertanyents a la llengua oral i 200 a la llengua escrita. El tercer filtre varia segons s'aplique a la llengua oral o a l'escrita. Les categories de textos representatius de la llengua oral coincideixen a grans trets amb els descrits pel BNC pel que fa a la llengua oral determinada pel camp. Les categories en què es divideix la llengua escrita comencen per no-publicada (50) i publicada (150). En la categoria de no-publicada trobem escrits no professionals (relacionats amb textos produïts per estudiants) i correspondència (cartes generals i de negocis). Dins de les categories publicades hi ha textos acadèmics (de temàtiques diverses), textos no-acadèmics (de les mateixes temàtiques), reportatges, textos instructius (administratius/reguladors o de hobbies), textos apel·latius (editorials de premsa) i escriptura creativa (novel·les/històries). Si fem una abstracció, en la part publicada trobem textos amb funció informativa (que varien segons el grau d'expertesa-divulgació), textos instructius, textos apel·latius i textos on predomina la funció estètica.

### 3.2.4. Altres corpus sincrònics de l'època de consolidació

El *Corpus de Referencia del Español Actual* (*CREA*) va sorgir en els anys 90 amb l'objectiu d' "oferir als investigadors d'aquesta llengua [l'espanyola] i als que s'hi interessen

una mostra representativa i equilibrada de l'espanyol estàndard que s'utilitza en l'actualitat en el món" (CREA). Segons els responsables, el disseny estava basat en una estructura complexa on s'encreuen criteris diversos de caire cronològic (el CREA recull textos de 1975-1999), geogràfic (50% d'Espanya, 50% d'Amèrica), del mitjà (textos publicats en llibres, revistes, diaris, transcripció oral, etc.) i temàtics (ciència, política, vida quotidiana, economia, ficció, etc.). No hem pogut trobar la manera com s'interrelacionen aquests criteris, tot i que hi ha dades sobre la distribució dels textos per criteri. El corpus presenta un 90% de textos escrits i un 10% de textos orals. Un 50% dels textos procedeixen d'Espanya i un 50% d'Amèrica, distribuïts per zones (40% de la zona mexicana, 3% de la zona central, 17% de la zona caribenya, 20% de la zona andina, 6% de la zona xilena i 14% de la zona d'Argentina, Paraguai i Uruguai). Les dades cronològiques estan distribuïdes per franges de 5 anys, els primers cinc anys tenen una representació del 10%, els cinc següents (1980-1984) d'un 15%, els del 1985-1989 un 20%, 1990-1994 un 25% i la franja del 1995-1999 d'un 30%. Les grans àrees temàtiques ("hipercampos") estan distribuïdes de la manera següent: ciència i tecnologia (10,125%), ciències socials, creences i pensament (13,5%), política i economia (13,5%), arts (10,125%), oci i vida quotidiana (10,125%), salut (10,125%) i ficció (22,5%). Seria molt interessant trobar la interrelació entre els diversos criteris de selecció, així com la motivació per a l'assignació dels percentatges. En l'última versió (3.2, juny de 2008), el nombre de documents s'ha incrementat amb textos produïts entre 1975 i 2004 , de manera que el nombre de mots del corpus ha arribat als 160 milions de formes. Segons la pàgina web de referència, "els materials que integren el CREA han estat seleccionant d'acord amb els paràmetres habituals, intentant mantenir l'equilibri establert en el disseny". Aquest "paràmetres habituals" remeten a la classificació que presentàvem anteriorment, amb una petita variació, ja que ara ficció i no-ficció constitueixen dos grans grups, de manera que queden 6 "hipercampos" que es distribueixen entres 20 àrees temàtiques més concretes.

El *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC) va ser constituït en els anys 90 i presentat el 1997. Tenia com a objectiu servir com a base per al *Diccionari Descriptiu de la Llengua Catalana* (DDLC). Conté 52,3 milions de mots, que estan distribuïdes de la manera següent: un 56% formen part de llengua no literària i un 44% de la llengua literària. La llengua literària està dividida en quatre gèneres: narrativa (59.92%), poesia (10,91%), teatre (15,95%) i assaig (13,22%). Per altra banda, la llengua no literària està estructurada en 10 grups de base temàtica: filosofia (6,05%) , religió i teologia (10,22%), ciències socials (19,2%), premsa (12,06%), ciències pures i naturals (7,66%), ciències aplicades (15,47%), belles arts/oci/esports/jocs (9,57%), llengua i literatura (7,63%), història i geografia/biografia (11,69%) i correspondència (0,45%).

Des d'una perspectiva cronològica, el CTILC abraça un període de 150 anys. Les seccions I (1833-1873) i II (1874-1913) estan dividides en 4 grups cronològics de 10 anys. La I inclou 2.260.083 mots i la II 8.457.482. La secció III (1914-1988) abraça 15 grups cronològics de 5

anys i està formada per 41.654.379 mots. Aquestes divisions no són aleatòries i atenen a tres períodes de l'ús social de la llengua.

*3.3. Pautes per a l'aplicació del concepte de representativitat: extensió i diversitat*

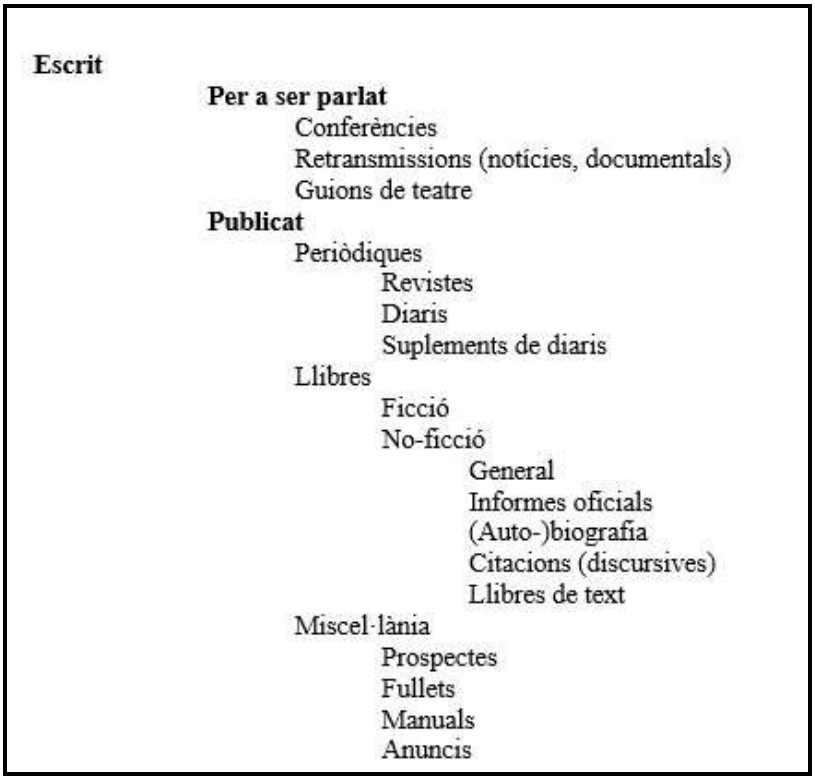
En l'apartat 3.1 hem vist la impossibilitat d'assolir la representativitat estadística d'una llengua en un corpus i, per tant, l'equilibri. Stefanowitsch (2017, 38) ens fa una proposta que ens resulta convincent: si no podem atènyer la representativitat i l'equilibri, ens haurem d'orientar cap a la diversitat. El seu raonament es basa en una premissa molt senzilla, tot i que un conjunt de textos no reflectisca fidelment la distribució lingüística d'una societat, com més diversa siga la seua composició, més possibilitats hi haurà que el fenomen lingüístic que volem investigar hi estiga representat.

Si ens fixem en els corpus precedents, aquesta és la perspectiva que realment apliquen. Creen una bresca de cel·les configurades a partir d'etiquetes encreuades rellevants i les omplin amb els textos pertinents. Aquesta pràctica implica que el corpus siga extens, ja que ha de recollir mostres de llengua de cadascuna de les cel·les. De manera que la diversitat i l'extensió són les maneres d'aconseguir un determinat grau de representativitat. En la pràctica, podem concloure que aquesta noció s'ha operacionalitzat en forma de criteris de disseny explícits, relacionats amb els objectius de recerca per als quals es crea el corpus i que reflecteixen la variació en el sentit més ampli. En aquest context, els termes gènere, tipus de text i tipologia textual no s'entenen amb el sentit propi que tenen en altres disciplines, sinó com a una categoria resultat de l'encreuament que comentàvem adés.

Els experts en l'àrea (vegeu, per ex. ATKINS, 1991, BIBER, 1993) postulen que la selecció inicial de textos s'haurà de fer basant-se en criteris externs, mentre que, posteriorment, es podrà millorar el corpus atenent a alguns factors lingüístics, que formen part de la caracterització interna. Atkins (1991, 15-19) identifica 25 atributs que poden aplicar-se als textos d'un corpus: mode (oral/escrit), participació (nombre de persones que participen en la interacció), preparació (associat al nivell d'espontaneïtat), mitjà (llibre, publicació periòdica, diari, televisió, ràdio, etc.), estil (prosa, poesia, etc.), gènere, constitució (simple o compost), realitat (ficció/no-ficció), context (de la situació comunicativa), funció (sense, narrativa, informativa, exhortativa, instructiva, reflexiva, creativa, etc.), tema, grau d'especialització, data, estatus (primera publicació, reedició, etc.), llengua, relacions amb altres llengües (en el cas de traduccions), estatus de la llengua (text origen/text traduït), metodologia seguida (en el cas de corpus terminològics), autoria (nom), sexe de l'autor, edat de l'autor, regió de l'autor, nacionalitat de l'autor, llengua materna de l'autor, autoritat de l'autor. Es tracta d'una caracterització genèrica que permetrà als compiladors fer una bona selecció de textos tenint en compte el focus d'interés.

Posteriorment (1991, 20-21) proposa un disseny més concret aplicat a una taxonomia de textos d'anglés britànic actual, amb l'objectiu de crear un corpus per als lexicògrafs i gramàtics de la Oxford University Press.


Figura 1: Taxonomia de tipologies textuals (ATKINS, 1991, 20-21), llengua oral



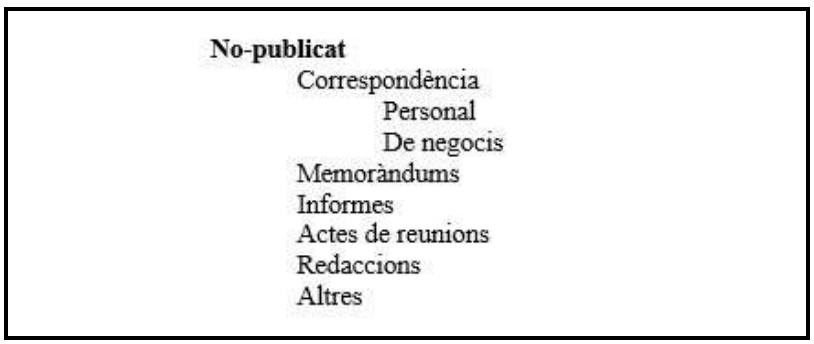




Figura 2: Taxonomia de tipologies textuals (ATKINS, 1991, 20-21), llengua escrita

Aquesta taxonomia comença pels dos grans blocs, llengua oral/llengua escrita, i posteriorment es detalla amb tipus més concrets d'interacció lingüística. La qüestió de la temàtica es deixa per a un filtre posterior. Atkins (1991, 22) assenyala que és molt important que consignem la temàtica per una recuperació i anàlisi posteriors, tot i que reconeix que és una etiqueta controvertida a l'hora d'establir taxonomies, per una banda perquè no està clar si es tracta d'un criteri extern o intern; per l'altra perquè l'assignació d'un tema depèn de la classificació que fem del món i; per últim perquè és una etiqueta que conflueix amb les etiquetes prèvies referides a la interacció. Independentment de la dificultat d'assignar una temàtica, l'autora aconsella que es definisquen algunes macrocategories (ciències i enginyeria, ciències socials, lleure-esports, lleure-art, etc.), semblants als "hipercampos" del CREA, per a controlar l'aparició de textos especialitzats en el corpus. Aquestes macrocategories temàtiques serviran com a pedra de toc per a evitar que hi haja una sobrerpresentació de determinades tipologies textuals determinades pel camp.

Biber (1993), per la seua banda, proposa una metodologia cíclica per a aconseguir la representativitat en un corpus. Parteix de la definició de la població, que inclou els seus límits i l'organització jeràrquica. Aquesta població compta amb criteris externs, que són els paràmetres situacionals que distingeixen els diferents textos dins d'una comunitat lingüística, i amb paràmetres interns, que són els trets lingüístics que s'analitzaran en el corpus. La selecció inicial de mostres haurà de fer-se seguint paràmetres externs, després es procedirà a la recopilació de textos, que portaran a investigacions empíriques de la variació lingüística i a la revisió del disseny inicial. Aquest procés de compilació, anàlisi de resultats i revisió del disseny es farà de forma cíclica fins que s'aconseguisca una mostra de llengua adequada atenent a criteris interns. Segons Biber (1993) un dels efectes positius de la compilació de corpus seguint aquesta metodologia serà aconseguir una descripció de la relació entre paràmetres externs i paràmetres interns de la llengua.

A més de la metodologia proposada, Biber (1993, 247) aporta determinades pautes molt relacionades amb la diversitat i l'extensió. Segons aquest autor, la llengua requereix una noció diferent de representativitat, on un mostratge proporcional a la realitat no és adequat. Si férem un corpus organitzat demogràficament, contindria els registres que la gent sol utilitzar. Aquesta mena de corpus permetria extraure estadístiques descriptives del conjunt de la llengua representada. Però aquestes generalitzacions basades en la llengua més emprada no solen ser interessants per als lingüistes, que prefereixen mostres que siguen representatives, en el sentit que incloguen tota la variació present en la llengua.

Així doncs, la representativitat fa referència al grau d'inclusió de tota la variabilitat d'una població, que en el cas de la llengua inclou tant les tipologies textuals com les distribucions lingüístiques. La forma d'operacionalitzar aquesta noció és mitjançant "sampling frames", és a dir mitjançant la determinació d'una llista de tots els membres de la població elegibles. Dins d'aquests "ventalls" de mostratge, ens aconsella que extraguem mostres estratificades, que són sempre més representatives que les no-estratificades. Amb aquesta "estratificació" ens

remet a la creació de categories dins de tot el conjunt de la població, és a dir, a les cel·les de bresca i, per tant, a la diversitat. De nou, el fet d'haver d'omplir les cel·les de bresca implica gran extensió.

Les reflexions anteriors ens revelen que és important tenir en compte la grandària d'un corpus a l'hora de fer-ne el disseny. Des d'una perspectiva aplicada, Stefanowitsch (2017, 38) apunta dues recomanacions: una de mínims i una de màxims. La més modesta proposa que un corpus ha de ser "suficientment extens per a contenir una mostra d'ocurrències del fenomen que es vol investigar". La més ambiciosa diu que "ha de ser suficientment extens per a contenir mostres suficientment grans de cada estructura gramatical, mot, etc.". Aquestes afirmacions ens remeten a un dels principis bàsics de la constitució de corpus: el disseny dependrà fonamentalment dels objectius de la nostra investigació.

Una vegada hem definit els objectius de la nostra investigació, hem de definir la mostra de població, el ventall de mostratge i fer-ne l'estratificació. A partir de l'estratificació, hem d'aconseguir les mostres pertinents. Si aconseguim un nombre de paraules acceptable per cada estrat, estarem en el bon camí. Però, com deien adés, es tracta d'un criteri extern que no garanteix que els microfenòmens de la llengua hi estiguen representats. L'única manera de comprovar aquest punt és dur a terme anàlisis de caire lingüístic, com ara les proves a què ha estat sotmés el contingut del *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCatMod) (vegeu Antolí Martínez, en aquest mateix volum).

La representativitat dels corpus especialitzats, almenys pel que fa al lèxic, pot mesurar-se pel grau de *closure* (clausura) (MCENERY; WILSON, 2001, 166) o *saturation* (saturació) (BELICA, 1996, 61-74) del corpus. La *clausura* o *saturació* d'un element lingüístic determinat (per exemple, el lèxic) en un segment de la llengua (com ara, les pòlisses d'assegurança de viatges) significa que aquest element sembla ser finit o està subjecte a una variació molt limitada a partir d'un punt determinat. Per a determinar la saturació d'un corpus, s'ha de dividir en segments que continguen els mateixos *tokens* (mots). Es considera que un corpus està saturat a nivell lèxic en el moment que, quan s'afig un nou segment, l'increment d'elements lèxics nous és aproximadament igual que el del segment anterior, és a dir, que la "corba de creixement lèxic esdevé asimptòtica" (TEUBERT 2000). En aquest mateixa línia, Corpas i Seghiri (2007) han desenvolupat l'algoritme N-Cor per a calcular el llindar de representativitat dels corpus a partir de la saturació lèxica. La noció de saturació es considera més adequada que altres com ara la d'equilibri, ja que es pot mesurar.

Els protocols d'avaluació dels corpus lingüístics varien segons l'objectiu que presenten. En el cas dels corpus destinats a millorar el processament del llenguatge natural, han de seguir uns protocols molts estrictes a fi que els resultats siguen aplicables. Quant a la recerca lingüística, si es volen fer estudis quantitatius, s'ha de parar especial atenció a la grandària del corpus i a la caracterització del contingut, especialment si es volen analitzar resultats procedents de diferents corpus. Per a estudiar freqüències d'elements en corpus amb diferent nombre de paraules és convenient fer atenció a freqüències relatives, més que no absolutes.

Conèixer bé la composició d'un corpus és fonamental també per a fer estudis qualitatius. Com més informació (en forma d'atributs) tinguem, millor podrem analitzar els resultats obtinguts, així com la seua validesa per a fer generalitzacions. El fet de disposar de corpus ben documentats ens obri les portes per a crear subcorpus comparables per a estudis concrets.

*3.4. Aplicació referida a corpus diacrònics*

Els corpus diacrònics presenten algunes característiques diferencials en relació amb els corpus sincrònics. La primera i principal és que els materials dels quals poden nodrir-se són finits, ja que es limiten als textos conservats de l'època de referència. Els responsables del *Representative Corpus of Historical English Registers* (ARCHER) reconeixen que "hi ha algunes limitacions naturals pel que fa als materials d'alguns gèneres en els períodes primerencs", que en el seu cas és el s. XVII. Aquest fet condiciona clarament la definició de la població, que haurà de cenyir-se als textos disponibles de l'època que es vol estudiar.

L'estratificació de la població està molt condicionada per la disponibilitat de textos i per les nocions culturals que hi intervenen. Hem de tenir en compte que moltes de les classificacions que fem per a la llengua actual resulten anacròniques si les apliquem a la llengua antiga. Per altra banda, si tenim com a objectiu fer estudis contrastius entre èpoques, serà molt convenient establir unes categories comparables.

En primer lloc, quant a la divisió llengua oral/llengua escrita, haurem de tenir en compte que la llengua oral no podrà ser representada, si no és en transcripcions conservades de l'època. L'opció d'enregistrar i transcriure aquest tipus de llengua està descartada per motius obvis. Fet que tindrà com a conseqüència que la presència de llengua oral en el corpus siga molt marginal.

En segon lloc, en aquesta mena de corpus, la cronologia té una especial importància i s'ha de tenir en compte no tan sols com a criteri descriptiu, sinó com a criteri de selecció. L'acurada descripció de les franges temporals i la voluntat d'omplir-les amb textos diversos serà la clau per a aconseguir un corpus diacrònic representatiu. La datació d'originals i edicions serà un altre dels factors que gaudirà de molta importància en el nostre procés.

En tercer lloc, el criteri geogràfic és més controvertit, ja que, com a mínim en el procés de selecció inicial, ens obliga a projectar la variació diatòpica actual en textos d'una altra època. En el cas que en comprovacions posteriors detectem que la distribució dels fenòmens interns era diferent a l'esperada, haurem de decidir si mirar de fer una nova classificació o conservar l'antiga. Des del punt de vista contrastiu, és interessant conservar les mateixes categories a fi de poder fer comparacions.

En quart lloc, quant al criteri de la classificació temàtica, Atkins (1991, 21) es plantejava que "posar una etiqueta a un tema és una classificació del món i el món és una estructura molt complexa que no remet fàcilment a una classificació evident". Aquesta reflexió augmenta la seua rellevància si parlem d'un món diferent al nostre. Sempre que etiquetem amb criteris actuals estem de nou projectant el nostre coneixement del món sobre una cultura diferent.

Crec que hem de ser conscients d'aquesta qüestió, però que no ens pot condicionar en el procés inicial de constitució. Aquesta premissa és vàlida també per a la variació diafàsica, diastràtica i per a les funcions dels textos, ja que es tracta de convencions que poden variar al llarg del temps. La idea és fer el disseny amb els nostres criteris externs actuals i després comprovar la seua validesa amb criteris interns. En qualsevol cas, si volem fer un corpus que siga comparable, haurem de mantenir les classificacions.

### 3.4.1. Alguns exemples de corpus diacrònics

El *Representative Corpus of Historical English Registers* (ARCHER) és un corpus multigènere que inclou anglés britànic i americà del període comprés entre 1600 i 1999. L'ARCHER 3.1 conté 1.789.309 mots dividits en vuit gèneres: teatre, ficció, sermons, diaris (de viatges/polítics) o dietaris, medicina, notícies, ciència i correspondència. Quant al paràmetre cronològic, el corpus inclou set períodes: 1650-99, 1700-49, 1750-99, 1800-49, 1850-99, 1900-40, 1950-99. L'ARCHER 3.2 conté 3.298.080 mots dividits en 12 gèneres: publicitat, teatre, ficció, sermons, dietaris, dret, medicina, notícies, prosa antiga, ciència, epistolaris i diaris (de viatges o polítics) i s'hi afegirà el període de 1600-49. L'ARCHER 3.3 haurà de completar tots els gèneres britànics del 1600-49, llevat del teatre, la prosa antiga i la medicina, també haurà d'incorporar més textos publicitaris britànics anteriors al 1750. Pel que fa a la varietat americana, haurà de completar tots els gèneres anteriors al 1750, textos mèdics de les franges 1800-49 i 1900-49, i sermons de les franges 1800-49 i 1900-49.

El *Corpus Diacrónico del Español* (CORDE) inclou 125.000.000 mots des dels inicis de la llengua espanyola fins al 1975, on comença el CREA. De manera anàloga al sincrònic, el CORDE ha estat estructurat seguint paràmetres cronològics, geogràfics i de modalitat i gènere. Des de la perspectiva cronològica, el corpus abraça tres etapes: Edat Mitjana (21%), Siglos de Oro (28%) i Època Contemporània (51%), que estan subdividides en períodes més breus. "Atesa la seua perspectiva diacrònica", recull un 74% d'espanyol peninsular i un 26% de l'espanyol de la resta del món. La distribució genèrica correspon a un 44% de ficció i un 56% de no ficció (didàctica – 9%, ciència i tècnica – 15%, societat – 8%, religió – 6%, història – 13%, dret i ciències jurídiques – 5%).

El *Corpus Informatitzat del Català Antic* (CICA) és un corpus que conté 414 obres en català des d'un període que abraça des del segle XI fins al segle XVIII, tot que se centra en la franja que va dels primers textos (s. XI-XII) fins al s. XVI. Aquests segles estan dividits en franges de 50 anys. Recull obres de procedències diverses classificades en dos macrogrups: oriental i occidental. Aquest grups es divideixen, al seu torn en: nord-occidental, valencià, septentrional, balear, central i alguerès. Quant al paràmetre de gènere, inclou 11 tipologies: prosa de ficció, cròniques i obres historiogràfiques, obres religioses i morals, prosa cancelleresca, textos administratius, textos jurídics, llibres de cort, textos científics i tècnics, epistolaris i dietaris, poesia i obres gramaticals i lexicogràfiques.

El *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCatMod) es troba en procés de constitució. En l'actualitat compta amb 240 textos i cinc milions i mig de mots d'un període que abraça des del s. XVII fins a 1832. Els segles estan dividits en franges de 50 anys, més els 33 del període final. Recull obres de procedències diverses de català oriental (rossellonés o septentrional, central, balear i alguerès) i de català occidental (català nord-occidental i valencià). Quant al paràmetre del gènere, inclou 7 tipologies: textos cientificotècnics, literaris, pedagògics, juridicoadministratius, historiogràfics/epistolaris i dietaris, obres gramaticals/lexicogràfiques i textos filosòfics, religiosos i morals. Podeu trobar més informació quant al procés de constitució d'aquest corpus en Antolí Martínez (en aquest mateix volum).

**4. Conclusions**

El recorregut per la història de la constitució de corpus ens ha servit per a revisar l'evolució de la metodologia referida al disseny i compilació d'aquestes eines imprescindibles en la recerca lingüística. Aquesta perspectiva ens forneix la informació necessària per a fer una mirada crítica a les pautes que s'utilitzen en l'actualitat per a la creació de corpus. En primer lloc, ens permet actualitzar el concepte de corpus, que podem definir com "un recull de mostres reals de llengua en format electrònic que s'ha recopilat amb l'objectiu de fer recerca lingüística". Perquè aquesta recerca siga vàlida, el conjunt de textos seleccionats ha de ser, en alguna mesura, representatiu de la llengua que volem descriure. En segon lloc, ens indueix a redefinir el concepte de representativitat en termes de disseny (estratificació), diversitat i extensió, ja que la millor manera d'assolir aquesta representativitat és mitjançant la inclusió de tota la diversitat possible, cosa que implica fer un corpus extens. La noció d'equilibri, que tradicionalment s'utilitzava per a aconseguir la representativitat i que s'aplicava mitjançant la fragmentació de textos i la imposició d'uns percentatges intuïtius, ha deixat de tenir sentit. Hem de tenir en compte que aquesta restricció tenia el seu origen en un moment on la capacitat d'emmagatzematge era limitada i era necessari recórrer a la fragmentació de textos per a assolir la diversitat. En tercer lloc, ens permet extraure una sèrie de consells aplicables a la constitució de nous corpus lingüístics. Els corpus lingüístics han de presentar un disseny d'acord amb els seus objectius de recerca. Aquest disseny haurà de ser explícit, abraçar de manera conceptual tota la població que vol representar i presentar una estratificació adient. L'estratificació, plantejada a partir de categories encreuades, servirà com a base per a la recopilació de les mostres de llengua i contribuirà a la diversitat i a l'extensió. Un corpus institucional, que sol tenir una motivació primera, però que sol aprofitar-se per a usos diversos, hauria de tendir a la màxima diversitat i a una extensió considerable. A pesar d'aquesta premissa, el corpus haurà de ser finit perquè els investigadors puguen utilitzar-lo com a base per a estudis comparables o replicables. A banda de ser finit, haurà d'estar ben documentat, ja que és imprescindible conèixer bé les mostres de llengua amb què treballem per a poder fer-ne una anàlisi acurada. La documentació, presentada en forma d'informació

addicional, permetrà la creació de subcorpus adients en la fase de cerca, així com una correcta interpretació dels resultats en la fase d'anàlisi.

## 5. Bibliografia


Antolí, Martínez, J. "El procés de constitució del Corpus Informatitzat de la Gramàtica del Català Modern (CIGCMod). Objectius, criteris i avaluació." En: *Revista Notandum*, 42, 2018.

Atkins, S. "Corpus Design Criteria – British National Corpus". En: *British National Corpus*, 1991. Consultable en línia: http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf [22/01/2018]

Atkins, S., Clear, J., & Ostler, N. "Corpus design criteria". En: *Literary and Linguistic Computing* 7(1), 1992, pàg. 1-16.

Belica, C. "Analysis of temporal change in corpora". En: *International Journal of Corpus Linguistics* 1(1), 1996, pàg. 61-74.

Biber, D. Representativeness in Corpus Design. En: *Literary and Linguistic Computing*, Vol. 8, N. 4, 1993, pàg. 243–257, https://doi.org/10.1093/llc/8.4.243.

Burnard, L. *Reference Guide for the British National Corpus*. Oxford: British National Corpus Consortium & Oxford University, 2007. Consultable en línia: http://www.natcorp.ox.ac.uk/docs/URG.xml [22/01/2018]

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

Clear, J. "Corpus Sampling". En: Leitner, G. *New Directions in English Language Corpora: Methodology, Results, Software Developments*. Berlín/New York: De Gruyter, 1992, pàg. 21-31.

DDAA. *ARCHER: A Representative Corpus of Historical English Registers*. Manchester: The University of Manchester. Consultable en línia: http://www.projects.alc.manchester.ac.uk/archer// [24/02/2018]

DDAA. *British National Corpus (BNC).* Consultable en línia: http://www.natcorp.ox.ac.uk/ [24/02/2018]

DDAA. *International Corpus of English (ICE).* Consultable en línia: http://www.ucl.ac.uk/english-usage/projects/ice.htm [24/02/2018]

DDAA. *Collins Wordbanks Online*. New York: Collins. Consultable en línia: https://wordbanks.harpercollins.co.uk/ [24/02/2018]

Firth, J.R. “Modes of Meaning”. En: *Papers in Linguistics 1934-1951*. London: Oxford University Press, 1957, pàg. 190-215.

Greenbaum, S. “A Proposal for an International Computerized Corpus of English”. En: *World Englishes*, Vol. 7, Issue 3, 1988.

Gross, J. (Ed.). *Information Society Multi-Conference Proceedings Language Technologies*. Ljubljana, 2000, pàg. 1-5.

Halliday, M.A.K. “Categories of the Theory of Grammar”. En: *Word*, Vol. 17, No.3, 1961.

---. “Lexis as a Linguistic Level”. En: Bazell C. et al. (Ed.). *In Memory of J.R. Firth*. London: Longman, 1966, pàg. 148-62.

IEC. *Corpus Textual Informatitzat de la Llengua Catalana (CTILC)*. Consultable en línia: https://ctilc.iec.cat/ [24/02/2018]

IEC. *Diccionari de la llengua catalana de l'IEC*. Barcelona: Institut d'Estudis Catalans. Consultable en línia: https://mdlc.iec.cat/ [24/02/2018]

Indurkhya, N.; Damerau, F.J. *Handbook of Natural Language Processing*. London/New York: CRC Press, 2010.

Kučera, H.; Francis, N. *Computational Analysis of Present-day American English*. Providence: Brown University Press, 1967.

Leech, G. *The State of Art in Corpus Linguistics*. En: Aijmer K; and Altenberg, B. (Ed.), *English Corpus Linguistics*, pàg. 8-29. London: Longman, 1991.

Manning, C.; Schütze, H. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 1999.

Martines, J.; Martines, V. (dir.). *Corpus Informatitzat de la Gramàtica del Català Antic* (CIGCA).

Martines, J.; Martines, V. (dir.). *Corpus Informatitzat de la Gramàtica del Català Modern* (CIGCatMod).

McEnery, T.; Wilson, A. *Corpus Linguistics*. Edinburgh: Edinburgh University Press, 1996.

---. *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press, 2001.

McEnery, T.; Xiao, R.; Tono, Y. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge, 2006.

Nelson, G., Wallis, S., Aarts, B. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2002.

RAE. *Corpus Diacrónico del Español (CORDE)*. Consultable en línia: http://www.rae.es/recursos/banco-de-datos/corde [24/02/2018]

RAE. *Corpus de Referencia del Español Actual (CREA)*. Consultable en línia: http://www.rae.es/recursos/banco-de-datos/crea [24/02/2018]

RAE. *Diccionario de la lengua española* (22.a ed.), 2001. Consultable en línia: http://www.rae.es/rae.html [24/02/2018]

Rafel i Fontanals, J. *Diccionari de freqüències: 3 dades globals*. Barcelona: Institut d'Estudis Catalans, 1998.

---. "Continuació del Corpus Textual Informatitzat de la Llengua Catalana". En: IEC. *Memòria 2015-2016*. Barcelona: IEC, 2018, p. 292.

---. Continuació del Corpus Textual Informatitzat de la Llengua Catalana (CTILC2), fase inicial. Barcelona: IEC. Consultable en línia: https://www.iec.cat/recerca/projecte1.asp?codi=PR2015-S04-RAFEL [24/02/2018]

Sinclair, J. "Beginning the study of lexis". En: Bazell C. et al. (Ed.). *In Memory of J.R. Firth*. London: Longman, 1966, pàg. 410-30.

---. *Looking up*. London: COBUILD, 1987.

---. *COBUILD Dictionary*. New York: Collins, 1987. Consultable en línia: https://www.collinsdictionary.com/dictionary/english [24/02/2018]

---. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

---. *EAGLES Preliminary Recommendations on Corpus Typology*. Pisa: Expert Advisory Group on Language Engineering Standards, 1996.

---. Corpus and Text - Basic Principles. En: Wynne, M. (Ed.). *Developing Linguistic Corpus: a Guide to Good Practice*. Oxford: Oxbow Books, 2005, pàg. 1-16. Consultable en línia: http://ota.ox.ac.uk/documents/creating/dlc/ [22/01/2018].

Stefanowitsch, A. *Corpus Linguistics: A Guide to the Methodology*, 2017. Inèdit.

Teubert, W. "Corpus Linguistics—A Partisan View". En: *International Journal of Corpus Linguistics* 4(1), 2000, pàg. 1-16.

Torruella, J. (dir.); Pérez Saldanya, M.; Martines, J. *Corpus Informatitzat del Català Antic (CICA)*, 2009. Consultable en línia: http://lexicon.uab.cat/cica [24/02/2018]

Váradi, T. "Corpus Linguistics—Linguistics or Language Engineering?" En: Erjavec, T.; Woods, A.; Fletcher, P.; Hughes, A. *Statistics in Language Studies*. Cambridge: Cambridge University Press, 1986.

Wynne, M (editor). *Developing Linguistic Corpus: a Guide to Good Practice*. Oxford: Oxbow Books, 2005. Consultable en línia: http://ota.ox.ac.uk/documents/creating/dlc/ [22/01/2018].

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