El processament de corpus I: la lingüística empírica

Joaquim Rafel i Fontanals
Joan Soler i Bou

PID_00233513

Temps de lectura i comprensió: **3 hores**

UOC
Universitat Oberta de Catalunya

Índex

Introducció

Totes les ciències que tenen per objectiu la formulació d'hipòtesis per a l'explicació d'algun aspecte de la realitat han de basar les seves conclusions en les dades que la mateixa realitat els ofereix. Les ciències naturals solen anomenar *dades empíriques* aquesta mena d'informació. Les ciències humanes també han de recórrer a dades empíriques que serveixin de referència de la realitat que pretenen descriure, i que contrastin o donin validesa a les seves hipòtesis. Quines són, però, les dades empíriques que la lingüística ha de prendre com a referent de la seva descripció? La llengua (la *langue*, oposada a la *parole* en la terminologia de Saussure) com a estructura és allò que el lingüista vol arribar a descriure, però no té altre camí per fer-ho que els indicis i les proves que li forneixen els enunciats lingüístics. Dit d'una altra manera, la lingüística també ha de cercar l'element de referència que confirmi les seves hipòtesis sobre la descripció del llenguatge i que aporti dades sobre el comportament general de la llengua.

Els corrents generativistes primerencs (Chomsky, 1957), que partien de la idea que el nombre d'enunciats d'una llengua és infinit, mantenien que no hi pot haver cap repertori finit de dades prou adequat per a l'explicitació dels mecanismes de producció lingüística. Segons això, calia cercar l'objecte de descripció en la figura d'un parlant ideal que posseeix la *competència* lingüística que li permet de produir enunciats en la seva llengua.

La lingüística passava, així, a preferir un enfocament *racionalista*, de base introspectiva, a un enfocament empirista. Aquesta reorientació comportà implícitament la crítica a la utilització de corpus com a base de descripció de la llengua. L'evolució posterior de la lingüística, i l'especial atenció prestada darrerament a la lingüística aplicada, ha fet veure, però, que l'objectiu d'un corpus no és, però, donar una visió total de la llengua, sinó oferir-ne una mostra representativa, que permeti al lingüista de fonamentar la seva recerca en dades objectives. Un corpus no es pot identificar amb la llengua sinó que és un conjunt de dades que la representa d'una manera més o menys fiable.

Així, sense donar encara una definició més precisa de què és un corpus, podem dir que es tracta d'un conjunt de dades sobre la llengua. Per a un lingüista interessat en la morfologia del lèxic, el corpus pot ser el conjunt de paraules derivades d'una llengua; per a un lingüista interessat en la sintaxi, un conjunt variat de frases de la llengua.

Però la noció de corpus s'utilitza d'una manera una mica més restrictiva i en una clara relació amb allò que anomenem processament del llenguatge. Les tècniques de constitució de corpus, els sistemes per a la seva explotació, els criteris i les eines que determinen el tipus d'informacions que s'hi afegeixen,

fan que la constitució de corpus ocupi un lloc destacat dins de les diferents aplicacions del processament del llenguatge, el qual s'inscriu en el marc més general dels anomenats **recursos lingüístics**.

Objectius

L'objectiu general d'aquest mòdul és que l'estudiant adquireixi un cert grau de familiaritat amb les qüestions relacionades amb la constitució i utilització dels corpus com a eina de presentació i estructuració de les dades lingüístiques. Això s'aconsegueix a partir de tres objectius específics:

1. Analitzar els aspectes que caracteritzen els diferents dissenys de corpus i avaluar com podem abordar la qüestió de la representativitat dels corpus lingüístics.

2. Conèixer els diferents tipus de corpus (orals i escrits) i assenyalar les característiques principals de les informacions lingüístiques que contenen.

3. Conèixer les diferents aplicacions que poden tenir els corpus en diferents àmbits d'estudi que tenen el llenguatge com a element comú.

A partir de la presentació conjunta de tots aquests aspectes, il·lustrats amb casos pràctics que permetin sempre una comprensió directa de la matèria d'estudi, pretenem que l'estudiant reflexioni, en un pla més general, sobre les grans possibilitats que la tecnologia ha obert en el camp de les ciències humanes i molt particularment de la lingüística.

1. Què és un corpus

Un corpus és un conjunt de dades organitzades. El lingüista basa les seves descripcions en les dades que li aporta el corpus, i comprova que les seves descripcions són correctes a partir de la contrastació d'aquestes amb el corpus.

1.1. Conceptes fonamentals

Un corpus (Sinclair) és una col·lecció d'elements lingüístics seleccionats i ordenats d'acord amb criteris lingüístics explícits amb la finalitat de ser usat com a mostra de la llengua. Complementàriament, un corpus automatitzat (o informatitzat) és aquell que s'ha codificat de manera estàndard i homogènia per a diferents tasques de recuperació de la informació. Es diu que un corpus és de referència quan ha estat dissenyat per a proporcionar informació sobre els diversos aspectes d'una llengua, de manera que representa totes les seves varietats de registre, de tipus de discurs, de vocabulari, etc.

Lectura complementària

J. Sinclair (1996). *Preliminary Recommendations on Corpus Typology*. S. l.: EAGLES Document EAG-TCWG-CTYP/P.

Com ja hem dit anteriorment d'una manera implícita, la lingüística ha trobat en l'anàlisi de corpus un camí per a la seva projecció descriptiva. Els motius d'aquest nou interès en l'anàlisi de corpus, que ha donat lloc al naixement de l'anomenada *lingüística de corpus*, cal buscar-los en les necessitats de descripció empírica i en el bandejament de la introspecció. Cal assenyalar, però, que la recuperació del mètode empíric no s'hauria pogut produir sense el desenvolupament tecnològic que ha permès la constitució i l'explotació de corpus cada vegada més extensos i cada vegada més complexos. Aquest desenvolupament ha possibilitat que els corpus es converteixin en un punt de referència cada vegada més generalment utilitzat per a diferents tipus de recerca lingüística (Biber *et al*.).

Lectura complementària

D. Biber; S. Conrad; R. Reppen (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Tots els tipus de corpus tenen una finalitat que d'una manera o d'una altra justifica l'estructura i el procediment que s'ha utilitzat per a formar-los. Entre les aplicacions primàries d'un corpus de referència hi ha la de servir com a base per a l'elaboració de diferents tipus de productes sobre la llengua, principalment diccionaris de diversa mena i gramàtiques. Un exemple més o menys recent d'aquestes aplicacions és el projecte COBUILD, que ha produït, a partir del corpus de la Universitat de Birmingham, un bon nombre de productes sobre l'anglès.

Per a acomplir la finalitat per a la qual ha estat dissenyat, el corpus ha d'estar estructurat en una base de dades dotada d'un sistema d'interrogació que permeti la recuperació de la informació textual. En general, un corpus de referència recull un volum de textos important (de desenes de milions de mots,

almenys) pertanyents a mitjans, tipologies i temàtica diversos, equilibrats en funció d'una hipòtesi de distribució entre aquests diversos paràmetres de classificació.

La *representativitat* d'un corpus respecte de la llengua que té com a referent està en funció d'una tria equilibrada entre els diferents tipus de textos que són susceptibles de formar-ne part.

Les possibilitats d'explotació de la informació continguda en un corpus depenen en gran mesura de l'**etiquetatge** o de l'**anotació** de què s'ha dotat. Aquest etiquetatge explicita, en forma de categories lingüístiques, característiques del text o dels mots que en formen part, que d'altra manera romandrien implícites tal com es produeixen en els enunciats lingüístics corrents. A través de l'etiquetatge d'un text podem, per exemple, determinar la categoria gramatical de cada element lèxic, i especificar les propietats flexionals de les seves distintes aparicions al llarg del text. Una operació d'etiquetatge es basa en un conjunt d'**etiquetes** (o **etiquetari**[1]) que representen les distintes categories lingüístiques utilitzades.

(1)En anglès *tagset*.

A diferència d'un simple arxiu de textos o d'una biblioteca electrònica (com ara l' *Oxford Text Archive*), els textos d'un corpus, doncs, solen incorporar informacions de diversa naturalesa que *classifiquen, anoten* o *etiqueten* les diferents unitats del text amb què s'han constituït i que permeten d'explotar-los o de processar-los per a finalitats específiques. Aquestes informacions poden referir-se a:

- aspectes bibliogràfics del text: autor, títol, any de publicació, tema, gènere, etc.
- aspectes d'estructura del text: marcatge tipogràfic, divisions textuals, paràgrafs, citacions, títols, fragments no analitzables, etc.
- caracterització de les unitats lèxiques: lèxic general, estrangerismes, noms propis, abreviatures, sigles, etc.
- caracterització morfosintàctica de les unitats lèxiques: categoria gramatical, forma gràfica del lema, categories morfològiques, etc.
- constituents sintàctics: representació dels arbres sintàctics, funcions sintàctiques de cada constituent, etc.
- caracterització semàntica: desambiguació d'unitats homògrafes, caracterització d'usos polisèmics, etc.

Algunes d'aquestes informacions poden afegir-se als corpus mitjançant procediments automatitzats, basats en el processament de la informació textual. Els més corrents entre aquests procediments son dels anomenats *etiquetadors* (*taggers*) morfosintàctics que, a partir d'anàlisis estadístiques o de l'aplicació de regles gramaticals, caracteritzen els elements lèxics en termes de categoria morfològica i en termes de part de l'oració (POS[2]). Alguns d'aquests etiqueta-

(2)POS, de l'anglès *part of speech*.

dors assignen també el *lema* (entitat abstracta que agrupa totes les variants flexionals d'un mot) a què pertany cada una de les ocurrències del text. Malgrat l'alt grau d'automatització d'aquests aspectes de l'etiquetatge textual, cal fer esment del fet que, per més baixos que siguin els nivells d'error amb què treballen aquestes aplicacions, la validació manual de les dades que en resulten esdevé, en molts casos, indispensable per a determinats usos.

Per a l'organització del treball lexicogràfic la fiabilitat de la informació que proporciona el corpus és de gran importància: el lexicògraf que redacta un article sol treballar amb segments seleccionats del corpus que corresponen a una entrada (o a una sèrie d'entrades relacionades) de diccionari; la presència en el corpus d'errors en l'assignació de la informació morfosintàctica pot ser un factor de desviació important en les dades que analitza el lexicògraf.

Documents de referència com ara les *Guidelines* de TEI (Sperberg-McQueen i Burnard) i el *Corpus Encoding Standard* (CES) d'EAGLES (Ide i Véronis) proporcionen informació sobre les possibilitats de tractament i d'integració de tota aquesta informació dins del corpus. Tots dos esquemes de marcatge es basen en definicions d'elements i relacions expressades en DTD[3] de SGML[4]. Per ser exactes, el CES d'EAGLES[5] és una interpretació més restrictiva i adequada específicament als corpus textuals de l'estàndard de TEI. Els canvis són mínims, de manera que el nivell de similitud entre tots dos esquemes de marcatge és molt alt.

[3] De l'anglès *Document Type Definitions*.

[4] De l'anglès *Standard Generalised Markup Language*.

[5] De l'anglès *Expert Advisory Group on Language Engineering Standards*.

SGML és un esquema de marcatge autodeclaratiu que està basat en una norma ISO (8879), i permet la definició de marques i etiquetes per a tot tipus de textos i d'estructures. És àmpliament utilitzat en l'etiquetació i en l'anotació de corpus per la seva ductilitat i les capacitats d'intercanvi de la informació.

Lectura complementària

C. M. Sperberg-McQueen i L. Burnard (Eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange*. Chicago-Oxford: Text Encoding Initiative.

La intenció de les propostes d'estàndard és la definició d'un format d'intercanvi entre els corpus, que en permeti la utilització multidisciplinar i per part de diferents grups de treball, i possibiliti també el desenvolupament d'eines de tractament i explotació de corpus no lligades a cap corpus concret. La idea d'intercanvi o de reutilització de la informació textual implica que totes les informacions que més amunt hem assenyalat com a caracteritzadores d'un corpus han de ser presents al text en el format estàndard, independentment de quina sigui l'estructura física d'emmagatzemament de la informació en una base de dades.

Els corpus juguen un paper fonamental com a **recursos lingüístics** que, a més de ser utilitzats en la recerca pròpiament lingüística, possibiliten el desenvolupament d'aplicacions computacionals de diversa naturalesa, o d'altres recursos lingüístics de segon nivell com ara lèxics electrònics (diccionaris per a usos relacionats amb el processament del llenguatge natural) de diversa mena.

En relació als corpus com a recursos lingüístics, s'ha posat molt d'èmfasi en els darrers anys en la idea de **reutilització** d'aquests recursos amb finalitats i orientacions diverses; amb la perspectiva d'ampliar les possibilitats d'utilització d'aquests recursos s'ha treballat en la línia de l'estandardització de les eines i estratègies de constitució de corpus.

1.2. Els corpus i l'ordinador

Els corpus constituïts amb anterioritat a l'aparició de l'ordinador estaven limitats per una sèrie de circumstàncies lligades al tractament manual de la informació. Prenguem com a exemple el *Diccionari Català-Valencià-Balear* (DCVB) d'Antoni Maria Alcover i Francesc de Borja Moll; aquest diccionari es va elaborar a partir d'un repertori de textos del qual s'extragueren les citacions utilitzades en el diccionari; aquest cedulari, que reuní obres de molt diversa naturalesa, constitueix un dels exemples més importants, en català, d'utilització de dades textuals en l'elaboració d'un diccionari. Donem un cop d'ull a com es va elaborar aquesta informació:

> Per a la replega de la llengua escrita s'han aplicat dos sistemes principals: primer, despullar i buidar en cèdules tot el vocabulari de certes obres (feina comanada a col·laboradors no especialitzats); segon, lectura, pels redactors i col·laboradors especialment preparats, de certes obres o textos, subratllant-ne els mots que interessava de registrar (i que després eren transcrits en cèdules, sia pels mateixos especialistes, sia per copistes no especialitzats).
>
> Introducció al *DCVB*, p. XXIV.

D'aquest fragment de text podem deduir dues característiques molt comunes en els corpus textuals desenvolupats sense el concurs de mitjans informàtics:

- La **falta de sistematicitat** en el tractament de la informació (es recull "tot el vocabulari" d'algunes obres, mentre que en d'altres obres es tracta selectivament el vocabulari que s'inventaria).

- La **participació de no especialistes** en les tasques de constitució del corpus.

- La intervenció del **criteris intuïtius** difícils de sistematitzar en la recol·lecció de la informació.

Els corpus constituïts i explotats manualment evidencien, doncs, el subjectivisme inherent al treball manual i a la falta de sistematicitat en el tractament de la informació textual, especialment en les operacions de selecció de la informació que cal recollir i de la que cal desestimar.

La utilització de l'ordinador permet una major **sistematicitat** en el tractament de grans volums de dades, de manera que no cal recórrer a certs **tractaments manuals** ni a l'aplicació de **criteris intuïtius** que poden desviar o emmascarar les dades que es volen estudiar.

La introducció dels mitjans informàtics en la constitució, processament i explotació de corpus ha permès de tractar grans volums d'informació d'una manera sistemàtica i objectiva, sense la intervenció de factors externs que no han de formar part de la recerca, com ara la intuïció del lingüista.

1.3. Tipologia dels corpus. Organització de la informació textual

Hi ha molts tipus de corpus, que es poden establir en funció del seu disseny, de les característiques formals o dels mètodes utilitzats per a la seva constitució.

Els aspectes més generals dels diferents tipus de corpus es determinen en funció de les hipòtesis sobre els nivells de representativitat i també pel tipus de funció concreta que es vol donar al corpus. Els criteris del primer tipus determinen aspectes força generals com, per exemple, el fet que un corpus estigui format per textos complets o bé per un conjunt de fragments de major o menor extensió, la selecció dels textos que n'han de formar part, etc.

Pel que fa al segon criteri de caracterització tipològica, dos dels principals grups que podem establir són el dels corpus monolingües o el dels corpus multilingües. Un tipus especial de corpus multilingües són els corpus paral·lels, que estan constituïts per versions dels mateixos textos en diferents llengües que han estat *alineats* (és a dir, relacionats entre sí els paràgrafs, frases o mots de les diverses versions) a partir d'un procediment d'etiquetatge.

Els criteris per al disseny de l'estructura de corpus fan referència a les característiques dels textos que els constitueixen. Cadascuna d'aquestes característiques es constitueix en un paràmetre de classificació, i a partir d'una combinació determinada d'aquests paràmetres es pot establir el disseny d'un corpus. Per a la determinació dels valors de cada paràmetre, hi ha **criteris externs** (que fan referència als tipus de text que estableixen les classificacions més usuals,

o a les característiques del context social en què ocorren els texts) i **criteris interns** (que fan referència a les característiques diferenciadores del llenguatge del text).

Els paràmetres de classificació tradicional fan referència, en general, a la **cronologia**, al **tema**, a l'**estil**, al **gènere literari**, i al **mitjà de publicació** d'un text. Alguns d'aquests paràmetres es poden determinar a partir de criteris interns o de criteris externs. Els criteris interns han estat sovint criticats pel fet que no estan avalats per cap tipus d'evidència científica ja que, en la pràctica, conviuen diferents classificacions sense criteris objectius que facin preferir-ne una sobre l'altra; a més, els nivells de detall de les descripcions poden ser molt variables. Els criteris externs són més objectivables i per aquest motiu han estat adoptats darrerament com a principal factor de disseny de corpus actuals.

Classificacions tipològiques més modernes, com les orientacions sobre tipologia textual elaborades per a EAGLES per John Sinclair, fan referència a la repartició de paràmetres en funció dels criteris externs i interns. Entre els paràmetres associables a criteris externs tenim:

- E.1. **origen**: aspectes de l'origen del text que poden afectar-ne l'estructura o el contingut; dades diverses sobre l'autor o autors, editor, etc.

- E.2. **estat**: qüestions relatives a l'aspecte físic del text i al seu suport en el moment en què és seleccionat per al corpus; mode de transmissió (oral, escrit o electrònic), relació amb el mitjà (característiques del suport de transmissió).

- E.3. **objectius**: qüestions relatives a la motivació del text i a les finalitats que persegueix; el tipus d'audiència o de públic a qui s'adreça, resultats que espera obtenir de la seva difusió, etc.

Seguint la mateixa classificació, els paràmetres associables a criteris interns són els següents:

- I.1. **tema**: el domini de coneixement a què pertany el text, els temes (en relació a les classificacions de la realitat) que tracta.
- I.2. **estil**: el tipus de model de llengua que segueix, sobre alguna de les classificacions externes existens.

A la pràctica, les estratègies de disseny de corpus solen fer referència a un nombre determinat de paràmetres. Seguidament analitzarem alguns d'aquests paràmetres.

1.3.1. Paràmetres cronològics

Una de les dades més objectives que podem donar sobre un text és la seva data d'elaboració. Per a la gran majoria de textos podem determinar fàcilment amb molta precisió quin és l'any (o el període cronològic) que correspon a la seva datació. Aquesta facilitat fa que la data sigui una de les informacions primàries amb què es caracteritzen els textos que formen part d'un corpus.

A l'hora d'establir el disseny d'un corpus, o de presentar-ne els seus materials, un dels criteris més fonamentals que en determinen l'estructura és la cobertura cronològica que abracen els seus textos.

Abast cronològic d'un corpus

En funció d'aquest abast, podem classificar els corpus en sincrònics, semisincrònics o diacrònics.

L'abast cronològic d'un corpus en determina el caràcter sincrònic o diacrònic. Per avaluar les diferents possibilitats que s'obren en aquest sentit, podem adduir un exemple. Suposem tres corpus que reuneixen les característiques següents:

a) Corpus 1: conté un nombre determinat de textos publicats en català entre el segle XII i el segle XVIII (ambdós inclusivament). Aquest corpus té una perspectiva històrica: vol reflectir l'evolució de la llengua en un període llarg. Es recullen variacions en la sintaxi, grans variacions en el lèxic. Aquest és un corpus bàsicament diacrònic.

b) Corpus 2: conté obres publicades entre 1830 i 1988 (uns 150 anys). Es detecta certa perspectiva evolutiva. Per exemple, aparició i desaparició de determinats significats, mots puntuals, etc. que s'associen a períodes concrets del corpus. Es tracta d'un corpus semisincrònic.

c) Corpus 3: conté obres publicades entre 1988 i 1999. Perspectiva evolutiva escassa. Intent de reflectir al màxim la llengua en la seva perspectiva sincrònica. Aquest és un corpus sincrònic.

En els corpus que es constitueixen actualment, el caràcter sincrònic sol prevaldre sobre el diacrònic, de manera que cada cop són més freqüents els corpus d'extensió considerable que estan formats per textos publicats en un espai de temps breu. Aquesta circumstància es veu propiciada per la major facilitat que ofereix la replega de materials textuals recents, que en molts casos es poden trobar directament en suport electrònic, la qual cosa facilita el seu processament amb vistes a la incorporació a un corpus.

Tot i això, cal destacar el fet que existeixen grans corpus semisincrònics o diacrònics que poden ser considerats entre les grans realitzacions de la lingüística aplicada: el **Frantext**, per exemple, és un gran corpus diacrònic que conté 150 milions de mots procedents de textos francesos publicats entre el segle XIV i l'actualitat.

1.3.2. Paràmetres mediàtics

Els paràmetres centrats en el mitjà de publicació del text són un dels criteris externs més utilitzats en l'actualitat per al disseny tipològic dels corpus. Això obeeix a la tendència que manifesta preferència pels criteris externs, atès que no presenten el subjectivisme inherent als criteris interns.

El caràcter limitat (malgrat les possibilitats de varietat actuals) del nombre de mitjans de difusió textual, i el fet que es tracta de característiques no inherents al text, facilita el fet de poder arribar a esquemes de classificació amb un alt grau d'homogeneïtat. Suposem que constituïm un corpus de textos escrits, i que establim proporcions entre cada un dels possibles mitjans de difusió textual; distingirem fàcilment entre els següents grups:

- Diaris.
- Revistes.
- Llibres.
- Correspondència.
- Textos electrònics (WWW, e-mail, etc.).
- Fullets informatius.

L'establiment d'aquests grups es basa en criteris objectius i, per tant, possibilita el fet que es puguin dissenyar corpus amb un grau màxim de rigor en els aspectes referents a la selecció textual. Els corpus constituïts en el projecte PAROLE, per exemple, han estat dissenyats i harmonitzats a partir d'aquest paràmetre.

Vegeu també

El projecte PAROLE es tractarà al subapartat 2.2.2 d'aquest mòdul didàctic.

Tornem a trobar aquí la tendència a seleccionar els tipus de difusió que ofereixen més possibilitats d'aprofitar materials existents.

Actualment, les possibilitats de reunir textos recents en suport electrònic (diaris i publicacions en format electrònic, textos en suport magnètic, textos de WWW, etc.) determina de manera preferent la constitució de corpus a partir de criteris mediàtics, ja que són fàcilment i objectivament aplicables.

1.3.3. Paràmetres de gènere

El gènere és un paràmetre que s'ha aplicat amb certa assiduïtat als corpus que contenen un nombre important de textos literaris. Aquesta classificació sol seguir la divisió tradicional entre els quatre gèneres literaris bàsics:

- Assaig.
- Narrativa.
- Poesia.
- Teatre.

En l'actualitat aquest paràmetre no és un dels que s'utilitzen de forma preferent en el disseny de proporcions de corpus.

1.3.4. Paràmetres temàtics

Els paràmetres que classifiquen un text a partir del tema de què tracta (de manera similar a com les biblioteques utilitzen en les seves catalogacions bibliogràfiques la Classificació Decimal Universal, per exemple) corresponen prototípicament als criteris interns. Aquest tipus de classificació ha estat molt criticada, pel fet que està basada forçosament en descripcions i classificacions de les àrees de coneixement que han de reduir forçosament una realitat complexa a una estructura jeràrquica monodimensional. Els problemes subjacents a l'establiment de les característiques temàtiques del text es traslladen, així, al subjectivisme que representa una divisió arbitrària de les àrees de coneixement humà.

A la pràctica, però, és un tipus de classificació textual molt utilitzada, especialment en els corpus més extensos. Pensem, per exemple, en les aplicacions que pot tenir en terminologia, en anàlisis del vocabulari, en treballs específics, etc. La indicació del "domini" a què pertany cada un dels textos d'un corpus pot ser d'una gran ajuda a determinades finalitats d'utilització del corpus, ja que possibilita el fet de treballar amb subconjunts del corpus definits a partir del paràmetre temàtic.

1.3.5. Altres criteris

Finalment, hi ha altres paràmetres específics basats en criteris externs que poden ser d'alguna utilitat: per exemple, el sexe de l'autor; la procedència geogràfica; l'estatus social dels textos, etc. Aquests paràmetres de classificació poden ser de gran utilitat per a determinades prospeccions sobre un subconjunt de textos d'un corpus.

1.4. Representativitat dels corpus

Una qüestió molt debatuda en el marc de la lingüística de corpus és fins a quin punt podem dir d'un corpus que és més o menys *representatiu* de la realitat lingüística que vol descriure. Podem dir que hi ha dissenys de corpus que són més representatius que altres? Quins són els criteris que determinen aquesta representativitat?

Podríem establir el factor determinant de la representativitat com la relació que existeix entre el disseny d'un corpus i les finalitats que s'han previst com a objectius fonamentals de la seva explotació. Així, el fet que hi hagi corpus més o menys extensos pel que fa al nombre de mots, més o menys restringits en el paràmetre cronològic, d'unes o altres proporcions de tipologies textuals, obeeix en general a les diferències sobre el tipus de resultats que hom n'espera d'obtenir.

La representativitat és un concepte lligat de manera indissoluble a allò que es vol representar; es tracta d'una hipòtesi de repartició de tots els paràmetres que només en la praxi es pot determinar com a més o menys afortunada.

Hi ha aplicacions d'un corpus que configuren en major o menor mesura un tipus de disseny determinat. Per exemple, els corpus que s'elaboren amb la intenció de servir com a font d'informació principal en l'elaboració d'un diccionari reuneixen una sèrie de característiques comunes:

- Nombre de mots elevat.
- Predomini dels textos escrits per sobre dels orals.
- Diversitat de grups temàtics.
- Inclusió de textos literaris.

La reunió de totes aquestes característiques i algunes altres poden fer un corpus més o menys adequat (o representatiu) per a la descripció lexicogràfica d'una llengua. Però la constitució d'un corpus elaborat a partir d'aquests criteris, que corresponen essencialment als d'un corpus de referència, requereix un esforç considerable (des dels punts de vista humà i financer) que no és a l'abast de tots els àmbits de recerca. Cal fer notar, però, que la reunió d'un gran volum de dades no és, en moltes ocasions, una condició *sine qua non* per a la utilització d'un corpus que puguem considerar representatiu de la realitat lingüística; hi ha un gran nombre de corpus de volum reduït que són utilitzats per a determinats tractaments de Processament del Llenguatge Natural (PLN), o per a la generació d'informació lèxica per a integrar en aplicacions de l'Enginyeria Lingüística.

2. Tipologia

2.1. Corpus orals

2.1.1. Conceptes fonamentals

Cal aclarir, d'entrada, que la definició, els objectius i el contingut dels corpus orals varia segons la tradició i la perspectiva en la que ens situem. En aquest apartat considerarem dues grans categories: per una banda, els corpus orals tal com s'entenen en la lingüística de corpus i, per una altra, els que s'utilitzen en les investigacions sobre fonètica i tecnologies de la parla.

Des del punt de vista de la lingüística de corpus, un corpus oral constitueix, habitualment, la transcripció ortogràfica –també anomenada transliteració– d'un enregistrament de la llengua parlada. Aquesta transcripció es pot enriquir amb diversos aspectes que reflecteixen el procés de producció de la parla i que varien en funció dels objectius del corpus. En darrera instància, el corpus constitueix una representació simbòlica de l'ús oral de la llengua i pretén, normalment, ésser representatiu d'un estil, d'un registre o d'una comunitat de parla.

En canvi, en el marc de la fonètica i de les tecnologies de la parla, l'aspecte més essencial d'un corpus oral és l'enregistrament mateix, ja que l'objectiu és obtenir informació fonètica o desenvolupar aplicacions relacionades amb la síntesi, el reconeixement o el diàleg. La representació simbòlica es sol fer mitjançant un alfabet fonètic, tot i que es crea també una representació ortogràfica per raons pràctiques.

2.1.2. Utilitat dels corpus orals

En aquest subapartat presentem molt breument algunes de les principals aplicacions dels corpus orals, centrant-nos en la divisió que hem establert anteriorment: corpus orals per a la fonètica i les tecnologies de la parla i corpus orientats a l'estudi de la llengua oral, inscrits en la tradició de la lingüística de corpus.

1) Corpus orals per a la fonètica i les tecnologies de la parla

Els investigadors en fonètica utilitzen els corpus orals com a eina indispensable per a la descripció segmental i suprasegmental de les llengües, tan en el nivell articulatori com en l'acústic.

En les diverses branques de la fonètica aplicada, la descripció contrastiva, l'anàlisi de la producció i de l'adquisició de la parla en psicolingüística, els estudis sobre interferència fonètica en l'aprenentatge de segones llengües, sobre patologies de la parla, i les recerques sobre sociolingüística i dialectologia centrades en la fonètica requereixen disposar de corpus orals enregistrats, fonèticament transcrits i codificats a diferents nivells.

Webs recomanats

A continuació teniu un llistat d'adreces web on podeu ampliar la informació sobre els diferents estudis que s'han portat a terme en els camps que hem esmentat:

a) Estudi de la prosòdia:

MULTEXT Prosodic Database.

MARSEC, Machine Readable Spoken English Corpus

b) Descripción fonètica contrastiva:

University of Victoria Phonetic Database.

IRIS, Immigrant Voices in Sweden

c) Estudi dels errors de producció de la parla:

Base de Données de Lapsus.

Estudi del llenguatge infantil. CHILDES, Child Language Data Exchange System.

d) Corpus en anglès recollit amb parlants no natius:

TED, Translanguage English Database

Pel que fa a les tecnologies de la parla, un corpus oral és necessari, per exemple, en el desenvolupament d'aplicacions de conversió de text a parla per a l'extracció de les unitats fonètiques que constitueixen el diccionari d'unitats de síntesi, per estudiar com s'enllacen o es concatenen aquestes unitats en la parla natural i per obtenir dades sobre aspectes prosòdics que permetin la creació de models de durada dels sons o d'entonació de les frases.

En el reconeixement, un corpus oral és una eina fonamental per crear models acústics de les unitats de reconeixement, per obtenir dades que permetin caracteritzar els parlant i l'entorn i per constituir models de llenguatge i crear els lèxics que utilitzen els reconeixedors. Si volem aplicar el reconeixement a serveis telefònics, cal, naturalment, recollir el corpus a partir de trucades, per tal de reproduir de la manera més fidel possible les condicions reals de funcionament del sistema.

Web recomanat

Corpus per al desenvolupament d'aplicacions del reconeixement de la parla per telèfon: The SALA Project - SpeechDat accross Latin America.

Per a la identificació i la verificació de la identitat del locutor, una estretament relacionada amb el reconeixement, és necessari disposar de corpus amb mostres de la manera de parlar de moltes persones diferents.

Lectura recomanada

Identificació i verificació de locutors: **J. Ortega García *et al.*** (1998). "AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification", in *Proceedings of*

ICAPSSP-98. IEEE International Conference on Acoustics Speech and Signal Processing. May 1998. pp. 773-776.

El disseny d'un sistema de diàleg requereix igualment corpus orals que recullin interaccions persona-persona o persona-màquina, per tal de comprendre millor la tasca que ha de realitzar el sistema i les estratègies emprades pels usuaris en demanar informació o en realitzar una determinada transacció.

Webs recomanats

Desenvolupament de sistemes de diàleg:

InfoTren: Person Dialogue Corpus

TRAINS Spoken Dialogue Corpus

2) Corpus per a l'estudi de la llengua oral

L'ús de corpus ortogràficament transcrits i codificats, tal com indicàvem que eren els característics de la tradició de la lingüística de corpus, és propi de l'anàlisi del discurs i de la conversa, especialment en tendències com l'etnografia del parlar.

La sociolingüística, especialment pel que fa a l'estudi del registre és una altra de les branques de la lingüística que s'ha caracteritzat darrerament per un ús freqüent dels corpus orals. També en dialectologia s'empren aquesta mena de corpus, que permeten caracteritzar fenòmens que apareixen en diversos nivells de l'anàlisi lingüística.

Lectura recomanada

Anàlisi sociolingüística i dialectal: **E. Boix** (1996). "Els materials de llengua oral dels corpus de català contemporani de la UB (CUB)". A: Payrató, Ll.; Boix, E.; Lloret, M.-R.; Lorente, M. (Eds.). *Corpus, Corpora*. Actes del 1er i 2on Col·loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 93-114.

L'anàlisi gramatical es pot fonamentar igualment en corpus orals, complementant les dades tradicionalment obtingudes de la llengua escrita.

Web recomanat

Descripció gramatical: CHRISTINE.

Finalment, un recull organitzat de mostres de llengua oral que combini la transcripció i l'enregistrament original pot ésser també un excel.lent material autèntic per a l'aprenentatge de llengües[6].

(6)Oral Language Archive.

2.1.3. Característiques específiques dels corpus orals

1) Corpus orals per a la fonètica i les tecnologies de la parla

Un corpus oral adaptat a les necessitats de la fonètica i de les tecnologies de la parla conté, com dèiem en l'apartat anterior, el senyal sonor enregistrat; en determinats casos, el senyal pot acompanyar-se també de dades articulatòries[7] que facilitin l'estudi del procés de producció de la parla.

(7)ACCOR, Articulatory-acoustic correlations in coarticulatory processes - a cross-linguistic investigation.

Pel que fa a les característiques lingüístiques, es consideren sovint els estils de parla, entesos com una sèrie de dimensions que varien en relació amb l'espontaneïtat, la formalitat i el grau de preparació o planificació del discurs oral.

Web recomanat

Estils de parla: PACOMUST, Corpus de Parole Continue Multistyle.

El contingut lingüístic dels corpus orals emprats en fonètica i en tecnologies[8] de la parla abasta des dels sons aïllats fins al discurs espontani, incloent elements específics com ara els logatoms –mots sense sentit però fonològicament ben formats, com ara "sula" en català– o les frases marc, frases d'estructura controlada en la qual s'insereixen els elements que s'analitzaran, conegudes també com a frases portadores, un exemple típic de les quals sol ser "va dir ___ i va sortir".

[8] EUROM1 Spanish.

Els corpus emprats en el desenvolupament d'aplicacions, especialment en el camp del reconeixement, poden incloure igualment les anomenades frases fonèticament equilibrades –en les quals la freqüència d'aparició dels sons en el corpus és equivalent a la de la llengua en general– o fonèticament riques –amb una bona representació de totes les combinacions dels sons de la llengua–, dígits, nombres connectats, seqüències alfanumèriques, lletres i paraules dites lletra per lletra, dates i hores, antropònims, topònims i mots relacionats amb l'aplicació.

Web recomanat

Desenvolupament d'aplicacions en tecnologies de la parla: Spanish SpeechDat.

L'adquisició del corpus es realitza, en general, en entorns acústicament controlats –una cambra anecoica o una sala insonoritzada per evitar la influència dels sorolls de l'ambient– tot i que es pot dur a terme en entorns naturals, amb les dificultats per a l'anàlisi acústica que això comporta, o aprofitant les emissions dels mitjans de comunicació. El desenvolupament d'algunes aplicacions requereix l'enregistrament per telèfon o la introducció de soroll de fons, de manera que es reprodueixin les condicions real d'ús de l'aplicació per a la qual es crea el corpus. Existeixen, a més, tècniques específiques per a l'adquisició de determinats tipus de corpus com la "tasca del mapa[9]" –en la qual un parlant ha d'explicar a un altre com s'arriba a un lloc a partir d'un mapa compartit– o el "protocol del Mag d'Oz", aquest darrer utilitzat en la recollida d'interaccions simulades entre un usuari i un sistema de diàleg.

[9] Map Task Corpus.

La transcripció és un element essencial en qualsevol corpus oral, i pot realitzar-se en diferents nivells: ortogràfic, fonèmic o fonològic, al.lofònic, fonètic o prosòdic. En la transcripció fonològica es representen únicament els fonemes, mentre que en l'al.lofònica es reflecteixen tots els al.lòfons que apareixen de manera sistemàtica com a resultat de processos regulars (per exemple, les assimilacions). En canvi, la transcripció fonètica pretén de representar de la manera més exacta possible allò que realment ha estat dit. En la transcripció prosòdica s'inclou específicament informació sobre moviments tonals de la corba melòdica.

Lectura recomanada

Transcripció: **J. Llisterri** (1997). "Transcripción, etiquetado y codificación de corpus orales". Fundación Duques de Soria, Seminario de Industrias de la Lengua, 15 de julio de 1997.

Una altra operació important és l'anomenada "alineació", que consisteix en sincronitzar la transcripció amb l'enregistrament, de manera que es puguin consultar com si es tractés d'una partitura amb la música i la lletra.

2) Corpus per a l'estudi de la llengua oral

Com s'indicava al principi, en la tradició de la lingüística de corpus, la transcripció ortogràfica enriquida i anotada constitueix la base del treball que es realitza sobre la llengua oral. Per això, fins fa poc, l'èmfasi en aquesta mena de corpus es posava en la representació escrita, més que no pas en disposar de l'enregistrament original. Darrerament, però, existeixen cada cop més projectes que es proposen proporcionar les dues fonts d'informació.

Els corpus per a l'estudi de la llengua oral reflecteixen, típicament, la variació sociolingüística entesa en un sentit ampli. En el seu disseny s'inclouen materials representatius de la màxima diversitat de situacions comunicatives, de manera que s'obtingui una mostra el més propera possible als diversos usos del llenguatge parlat.

Webs recomanats

Corpus per a l'estudi de la llengua oral:

The Spoken Component of the British National Corpus.

CSAE, Santa Barbara Corpus of Spoken American English.

Al costat dels corpus generals, n'existeixen d'especialitzats en els quals s'intenta de recollir dades sobre un àmbit específics com, per exemple, el discurs polític, el discurs acadèmic, la llengua dels mitjans de comunicació, les converses informals, la llengua dels joves, i tots aquells àmbits que ofereixen un especial interès per a la descripció lingüística.

Webs recomanats

Discurs acadèmic i discurs polític:

CPSA, Corpus of Professional Spoken American-English.

a) Discurs polític:

ADPA, Análisis del discurso público actual.

b) Llengua als mitjans de comunicació:

DIES, Difusión Internacional del Español.

La transcripció ortogràfica del corpus sol ésser el primer pas en la preparació de les dades per tal de fer-les accessibles als usuaris. Tot i que pot semblar una operació trivial, no ho és tant en el cas de corpus que recullen llengua oral espontània. En el marc de l'anàlisi del discurs i de la conversa s'han creat diversos sistemes de notació que enriqueixen la representació ortogràfica amb els elements necessaris per a l'anàlisi.

Lectura recomanada

Transcripció: **Ll. Payrató** (1996). "Transcripció del discurs col.loquial". A: Payrató, Ll.; Boix, E.; Lloret, M.-R.; Lorente, M. (Eds.). *Corpus, Corpora*. Actes del 1er i 2on Col·loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 181-216.

També des de la pròpia lingüística de corpus han sorgit sistemes per incloure en la representació ortogràfica diversos aspectes de la llengua oral.

Aquests elements es solen codificar mitjançant marques que segueixen estàndards com els de la Text Encoding Initiative i que permeten recuperar la informació desitjada i intercanviar fàcilment els corpus. Fenòmens com el torn de paraula, la identitat del parlant, la superposició entre locutors, les interrupcions en el discurs i altres elements propis de la llengua oral com els que fem servir per expressar dubte, negació, assentiment, les pauses, els errors en la producció o els fragments que no queden clars al transcriptor són alguns dels elements que es solen codificar en la transcripció d'un corpus oral.

Web recomanat

Transcripció i codificació d'un corpus oral: CSAE, Corpus of Spoken American English, Transcription conventions.

Hem de destacar, finalment, que els materials als quals ens estem referint s'integren, en alguns dels projectes més importants, com una part d'un corpus general que recull tant llengua escrita com parlada. Aquest és el cas del *British National Corpus* o del CREA (*Corpus de Referencia del Español Actual*), per esmentar només dos exemples.

2.2. Corpus escrits

2.2.1. Els principals corpus escrits

La presència i el grau d'importància d'una llengua en l'anomenada *societat de la informació* es pot determinar a partir de la importància, la magnitud i el grau d'actualització dels *recursos lingüístics* que s'han constituït per a ser utilitzats en diversos tipus d'aplicacions. Entre aquests recursos, el desenvolupament de grans corpus de referència ha esdevingut un dels primers objectius a acomplir per part de les llengües d'un pes cultural i demogràfic més destacat.

Webs recomanats

Desenvolupament de sistemes de diàleg:

InfoTren: Person Dialogue Corpus.

TRAINS Spoken Dialogue Corpus.

Un dels primers grans corpus europeus desenvolupats per a finalitats lexicogràfiques fou el FRANTEXT, constituït per l'*Institut National de la Langue Française* (INaLF), i que ha estat utilitzat per a la redacció del *Trésor de la Langue Française*. Actualment, el FRANTEXT és un corpus d'uns 150 milions de mots obert a consulta pública (per als subscriptors), i s'hi pot accedir telemàticament.

Les possibilitats actuals de constituir corpus a partir del reprocessament de materials textuals que es troben ja en suport electrònic ha permès la constitució de corpus de grans dimensions a partir de sistemes de processament i etiquetatge altament automatitzats, com ara el *British National Corpus* (col·lecció de 4.000 fragments de textos d'una extensió total de 100 milions de mots).

Un altre exemple d'aquest tipus el constitueix el corpus anomenat *Bank of English*, constituït a partir del projecte COBUILD de la Universitat de Birmingham, i amb una extensió total de més de 300 milions de mots.

Corpus	Nombre de mots	Llengua
FRANTEXT	150 M	Francès
British National Corpus	100 M	Anglès

Corpus	Nombre de mots	Llengua
Bank of English (COBUILD)	300 M	Anglès

La dimensió comercial i les necessitats d'actualització d'aquests corpus han fet desenvolupar el concepte de *monitor corpus*. Un *monitor corpus* és un corpus que es manté actualitzat permanentment, i que per tant va creixent indefinidament, i que es configura com un servei que pot ser utilitzat amb les finalitats habituals (lexicografia, processament del llenguatge natural, etc.).

En l'actualitat la quantitat de corpus existents i de projectes de constitució de corpus creix cada dia, fins al punt que es fa difícil de donar-ne una relació. Hi ha adreces WWW específiques que són actualitzades periòdicament on es pot trobar informació sobre diferents corpus.

2.2.2. La situació del català

L'Institut d'Estudis Catalans va desenvolupar, entre l'any 1985 i el 1996, el **Corpus Textual Informatitzat de la Llengua Catalana** (CTILC) (Rafel, 1994), com a primera fase del programa de recerca **Diccionari del Català Contemporani** (DCC). El DCC té com a objectiu l'elaboració d'un diccionari descriptiu de la llengua catalana moderna.

El CTILC ha estat desenvolupat com un corpus especialment orientat a l'explotació lexicogràfica, però el seu caràcter multifuncional i la seva extensió fan que sigui susceptible de ser utilitzat per a qualsevol treball que requereixi el concurs de dades textuals. El CTILC està format per un total de 3.399 textos catalans publicats entre 1830 i 1985, i té una extensió total de 50.000.000 de mots, distribuïts en un 40% de llengua literària (poesia, narrativa, teatre, assaig) i un 60% de llengua no literària (classificada en 10 grups temàtics diferents). EL CTILC és un corpus consultable telemàticament i està disponible per a usos de recerca.

Altres corpus públicament accessibles i que han estat desenvolupats en llengua catalana són el corpus català del projecte europeu PAROLE, que consta d'un total de 21.000.000 de mots que majoritàriament corresponen a premsa i a altres publicacions periòdiques. El projecte PAROLE ha desenvolupat una sèrie de corpus que cobreixen un total de 13 llengües europees (alemany, anglès, català, danès, finès, francès, grec, irlandès, italià, neerlandès, noruec, portuguès, i suec), que s'han constituït amb la finalitat de dotar aquestes llengües d'un cert nombre de recursos lingüístics harmonitzats (dissenyats a partir de criteris comuns) i màximament reutilitzables. Cada un dels corpus té una extensió aproximada de 20 milions de mots, seleccionats a partir de criteris me-

diàtics i cronològics (textos publicats a partir de 1980 fins a 1998, moment de constitució del corpus). Les proporcions de representació de cada un dels valors mediàtics és la següent:

Mitjà	% mínim	% màxim
Premsa diària	16	22
Publicacions periòdiques	58	72
Llibres	4	10
Miscel·lània	8	12

L'Institut Universitari de Lingüística Aplicada (IULA) de la Universitat Pompeu Fabra també té (en fase de creació) un projecte de corpus multilingüe format essencialment per textos d'especialitat i orientat bàsicament a la recerca en terminologia i neologia.

En altres centres s'elaboren, també, corpus de dimensions més reduïdes, en general lligats a projectes concrets, i que no corresponen a les característiques generals dels corpus de referència.

3. Processament de corpus

3.1. Etiquetatge morfològic i sintàctic. Lematització

Ja hem vist abans que un corpus no és una mera col·lecció de textos sense cap tipus d'organització ni de marcatge intern, sinó que el concepte de corpus va lligat a l'estructura i a l'etiquetació que en possibiliten l'explotació. La utilització de corpus per a la recerca lingüística o per a d'altres finalitats exigeix, en moltes ocasions, que els mots pertanyents als textos que s'hi ha introduït estiguin *etiquetats* de manera que hom pugui conèixer, per a cada mot, les informacions de naturalesa gramatical que s'hi relacionen (quina part de l'oració és, amb quines característiques de flexió es presenta, etc.). Cada mot d'un corpus organitzat sol incorporar una *etiqueta* (d'aquí el terme *etiquetatge*) que n'indica les seves característiques.

Les característiques que més correntment se solen indicar fan referència a les propietats morfosintàctiques del mot tal com apareix en el text. La forma que pot tenir un fragment de text etiquetat seguint un sistema de marcatge SGML és més o menys la següent:

```
<w lemma="feixuc" msd="AQPFS0.">Feixuga </w>
<w lemma="tasca" msd="NCFS00.">tasca </w>
<w lemma="el" msd="TDFS0">la </w>
<w lemma="nostre" msd="DP1FS0P">nostra</w>
<w msd="FI">, </w>
<w lemma="anar" msd="VAIP1S.">vaig </w>
<w lemma="pensar" msd="VMN....">pensar </w>
<w lemma="mirar" msd="VMG....">mirant </w>
<w lemma="el" msd="TDCS0">l'</w>
<w lemma="assalariat" msd="NCMS00.">assalariat </w>
<w lemma="del" msd="SPCMS">del </w>
<w lemma="xalet" msd="NCMS00.">xalet</w>
<w msd="FE">. </w>
```

Aquest és un fragment brevíssim d'un text etiquetat del corpus català de PAROLE, concretament de la frase "*Feixuga tasca la nostra, vaig pensar mirant l'assalariat del xalet.*" Es tracta d'un text etiquetat morfosintàcticament i lematitzat de manera que les informacions morfosintàctiques es fan explícites. En SGML els *tags* o etiquetes es marquen entre els signes < i >. Els tags <w> i </w> identifiquen el principi i el final de cadascun dels diferents mots i dels signes prosòdics del text. El tag <w> incorpora en el seu contingut una sèrie d'*atributs*; el primer d'aquests atributs, anomenat *lemma* conté com a valor el lema o forma canònica del mot; a continuació, l'atribut *msd* introdueix la in-

formació gramatical segons un sistema formalitzat que segueix els estàndards d'EAGLES (AQPFS0, per exemple, significa "adjectiu qualificatiu de grau positiu i femení singular").

A partir de tècniques basades en el processament del llenguatge natural s'han desenvolupat sistemes per a l'etiquetatge automatitzat dels textos d'un corpus. Aquests sistemes es basen en dues estratègies bàsiques, i per tant es poden classificar en dos tipus fonamentals:

1) Etiquetadors basats en regles: consten de diversos components, el més important dels quals és un conjunt de regles que determinen l'etiquetatge de les diferents formes del corpus. Un altre dels components fonamentals d'aquests sistemes és un lexicó (o diccionari informatitzat) que conté informació sobre els diferents lemes i formes morfològicament diferenciades. Aquest diccionari interacciona amb el component de regles de manera que permet verificar-ne l'aplicabilitat.

Un exemple d'etiquetador basat en regles: suposem que donem a un sistema automatitzat una frase com "es va menjar una amanida"; el diccionari pot informar el sistema de les diferents possibilitats d'etiquetar aquesta seqüència, paraula per paraula: la forma *va* pot pertànyer al verb *anar*, o bé a l'auxiliar de formació del perfet perifràstic; *menjar* pot ser la forma infinitiva del verb o el substantiu *menjar*, i *amanida* pot ser tant el substantiu com la forma femenina singular del participi passat del verb *amanir*. Amb tota aquesta informació, el sistema podria assignar a cada una de les formes les diferents possibilitats d'etiquetació que presenten; el resultat fóra un etiquetatge que en alguns casos resultaria ambigu. Si tenim en compte l'existència del mòdul de regles, però, bastarà determinar que en contacte amb una forma que pot ser infinitiva, la forma *va* serà considerada com a auxiliar. Pel que fa a la forma *menjar*, el fet d'anar acompanyada de la forma *va* permetria resoldre-la com a infinitiu. Alhora, la formulació d'una regla que determini que un possible nom precedit d'un article indeterminat s'ha de solucionar com un nom permetrà que *amanida* sigui caracteritzat com a nom i no com a participi de passat del verb *amanir*.

2) Etiquetadors estadístics: a diferència dels anteriors, els etiquetadors estadístics no requereixen l'enunciació sistemàtica de les regles gramaticals que podem trobar en una llengua determinada, sinó que basen les seves decisions en un sistema constituït automatitzadament.

Els etiquetadors d'aquest tipus han de ser *entrenats* a partir d'un cert volum de textos prèviament etiquetats; l'anàlisi automatitzada d'aquests textos permet de fer les inferències necessàries per a la determinació de les solucions més plausibles per als diferents contextos en què poden aparèixer els mots.

Els etiquetadors són eines utilitzades de manera especial en la constitució de grans corpus. Un dels paràmetres que dóna una idea de quina és la seva efectivitat és el tant per cent d'etiquetes assignades correctament; es considera que un etiquetador ha assolit un nivell de perfeccionament acceptable quan se situa entre el 95 i el 98 % de solucions correctament assignades.

3.2. Informació estadística

El fet que els corpus integrin en una estructura organitzada un gran nombre de dades sobre el llenguatge possibilita que s'utilitzin per a fer-nos una idea de l'estructura del vocabulari en termes quantitatius. Un concepte bàsic i absolutament primari per a l'extracció de dades quantitatives d'un corpus és el concepte de **freqüència** d'un element dins d'un corpus, que fa referència al nombre de vegades que aquest element hi apareix.

El concepte de freqüència no és exclusiu d'un tipus d'entitat determinada, sinó que es pot referir a informacions molt diferents. Així, podem parlar de la freqüència d'una forma, d'un lema, o fins i tot d'una categoria gramatical (nom, adjectiu, etc.) o d'una determinada combinació de lletres.

Una altra constatació evident és que la freqüència, tal com l'hem presentada, és una dada *absoluta*, el valor de la qual depèn fortament de l'extensió del corpus a què es refereix. Suposem que un determinat element lèxic *a* apareix 50 vegades en un corpus de 50.000.000 de mots, mentre que un altre element lèxic *b* apareix també 50 vegades en un petit corpus de 5.000 mots. Encara que *a* i *b* tinguin la mateixa freqüència en termes absoluts, la seva importància relativa en cadascun dels dos corpus és força diferent: *a* apareix una vegada cada milió de mots, mentre que *b* apareix una vegada cada cent mots.

El concepte de *freqüència relativa* dóna raó d'aquesta relació existent entre la freqüència d'aparició i la magnitud del corpus en què apareix. La freqüència relativa s'ha d'expressar com un valor relacional, de manera que podem expressar-la en tant per cent, per mil, etc. Així, direm que la freqüència relativa, expressada en tant per cent, de la forma *a* en l'exemple anterior és de 0,0001 (és a dir, que apareix 0.0001 vegades cada cent mots), mentre que la de la forma *b* és de d'1 (és a dir, que apareix un cop cada 100 mots). Aquesta manera d'expressar el valor relatiu d'un element lèxic ens dóna, de fet, una idea molt més indicativa a l'hora de valorar la importància d'un mot en el conjunt del vocabulari d'una llengua.

Un altre element essencial a tenir en compte al costat del nombre de vegades que apareix un mot és allò que podríem anomenar la seva *distribució*, això és, quin és el seu grau de presència en els diferents tipus de textos que integren un corpus. Suposem que, en un corpus amb els textos classificats temàticament, dos mots presenten freqüències similars, però un d'ells concentra gairebé totes les seves aparicions en un tipus temàtic (per exemple en matemàtiques, o bé dret, o bé psicologia, etc.), mentre que l'altre es presenta repartit més o menys

equitativament entre la totalitat dels grups. Immediatament diríem que el segon mot té un caràcter més *general* en el vocabulari que el primer; aquest, en canvi, podria tractar-se amb força probabilitat d'un mot específic d'una matèria determinada. Consideracions d'aquest tipus esdevenen fonamentals per a la determinació del vocabulari bàsic d'una llengua; aquests vocabularis són de gran interès en molts camps de la lingüística aplicada, com l'ensenyament de llengües, la traducció, etc.

Diccionaris de freqüències

Són repertoris que ens donen informacions freqüencials sobre els elements lèxics. En català existeix un diccionari de freqüències elaborat a partir del Corpus Textual Informatitzat de la Llengua Catalana, de l'IEC.

El català disposa d'un diccionari de freqüències elaborat a partir de les dades proporcionades pel CTILC. Aquest diccionari, publicat alhora en format de llibre i en format electrònic (CD-ROM), recull dades estadístiques dels lemes del CTILC, les quals fan referència a la freqüència i a la distribució de cada un dels lemes dins del corpus. La possibilitat de consultar el diccionari com una base de dades accessible informàticament fa que puguem extraure'n de manera automàtica informacions de gran interès per a determinades investigacions. Com a exemple, donem el resultat d'una consulta sobre la versió electrònica del *Diccionari de freqüències*, que correspon a una sentència d'interrogació que demana els adjectius invariables terminats en *–forme* i que tinguin una freqüència absoluta en el corpus superior a 10; demanem també que els resultats s'ordenin per ordre decreixent de freqüències:

Lema	CG	F.abs.	F. rel.
conforme	ai	745	0.001454
uniforme	ai	740	0.001444
informe	ai	205	0.000400
deforme	ai	151	0.000295
multiforme	ai	80	0.000156
aeriforme	ai	54	0.000105
campaniforme	ai	45	0.000088
filiforme	ai	40	0.000078
disconforme	ai	34	0.000066
cuneïforme	ai	29	0.000057
puntiforme	ai	23	0.000045
fusiforme	ai	17	0.000033
bacciforme	ai	11	0.000021
cruciforme	ai	11	0.000021
piriforme	ai	11	0.000021

Lectures complementàries

Diccionari de freqüències en català:

Joaquim Rafel i Fontanals (dir.): *Diccionari de freqüències, 1. Llengua no literària*, Barcelona, Institut d'Estudis Catalans, 1996.

Joaquim Rafel i Fontanals (dir.): *Diccionari de freqüències, 2. Llengua literària*, Barcelona, Institut d'Estudis Catalans, 1998.

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

El fet de poder comptar amb un diccionari de freqüències que proporcioni informacions adequades i representatives sobre la llengua possibilita la realització de treballs en molts àmbits de recerca, des de l'aprenentatge de llengües fins a la psicologia cognitiva, ja que permet extreure conclusions sobre el caràcter més o menys bàsic de cada element lèxic en relació al conjunt del vocabulari.

3.3. Els resultats d'un corpus

Els resultats d'un corpus

Depenen de la interfície d'accés. La interfície és el conjunt de programes que ens permeten d'extreure informació d'un corpus informatitzat. Un programa de consultes, per exemple, actua com una interfície en el sentit que converteix les nostres demandes en instruccions interpretables per la base de dades que gestiona el corpus.

Se sol dir que els resultats d'un corpus es presenten en forma d'índexs, i que hi ha dos tipus d'índexs, que es denominen en general a partir de la seves sigles en anglès:

- Índexs KWOC (*Keyword out of context*) que contenen informació que no incorpora el context en què apareix un element lèxic (paraula clau).
- Índexs KWIC (*Keyword in context*) que presenten els elements lèxics (paraules clau) dins del context en què apareixen.

El terme *índex* fa referència a les llistes (de mots, de categories, etc.) que es poden extreure d'un corpus, i que històricament s'han elaborat fins i tot amb anterioritat a l'aparició dels mitjans informàtics (són especialment destacables els *índexs de concordances* de la Bíblia elaborats manualment). En l'actualitat les possibilitats creixents de l'explotació automatitzada fan que els resultats que es poden obtenir dels corpus siguin molt variats.

La divisió genèrica en índexs KWIC i KWOC, tot i que resulta essencialment vàlida, identifica dos casos extrems de resultats; actualment, el desenvolupament d'interfícies d'accés a corpus i les aportacions de la lingüística de corpus han possibilitat l'obtenció d'informacions cada vegada més complexes i que no corresponen estrictament a cap dels dos tipus.

Per exemple, si apliquem a un corpus un procés consistent a determinar quines possibilitats té cadascun dels lemes de coaparèixer amb un altre lema qualsevol en una determinada posició anterior o posterior, obtindrem uns resultats que poden tenir formes i presentacions diverses en funció, fins i tot, dels paràmetres que s'apliquin al procés, però que en tot cas no corresponen a cap cas prototípic dels dos models exposats, tot i que per a la seva obtenció hagi calgut avaluar i quantificar els contextos de cadascun dels mots del corpus. Un altre exemple d'un cas difícil de classificar en aquests termes és, per exemple, una llista de freqüències de cada una de les categories gramaticals o morfolò-

giques amb què ha estat etiquetat el corpus: els resultats d'aquest procés no requereixen específicament l'anàlisi de context, però tampoc no es refereixen a cap paraula clau en concret.

En funció d'aquestes consideracions, les possibilitats actual fan més adequat parlar de *tipus de processos*, d'una banda, i de *tipus de resultats*, de l'altra, com a criteri de classificació de les diferents informacions que ens pot donar un corpus.

a) *Tipus de processos*: hi ha processos no contextuals en els quals tots els paràmetres interns de selecció afecten un únic element considerat en termes absoluts, sense tenir en compte els elements que hi coapareixen (per exemple: una llista de freqüències de lemes, una llista de freqüències de categories gramaticals, la distribució de categories morfològiques dins de cada lema, una llista de concordances referida a un sol lema, etc.). D'altra banda, hi ha processos en els quals la selecció es realitza sobre un patró que no té en compte un sol element, sinó també els elements que l'envolten (de manera immediata o no); aquest és el cas, per exemple, d'una llista de concordances d'un lema seguit d'una categoria gramatical determinada, o la de les expectatives d'aparició conjunta que tenen dos lemes, etc.

b) *Tipus de resultats*: de manera similar als tipus de processos, també hi ha resultats de dos tipus, en funció del fet que mostrin o no el context en què apareix la informació que s'ha demanat. Així, un llistat de concordances correspondrà a un resultat de tipus contextual, mentre que una llista qualsevol de coaparicions, freqüències, etc. correspondrà a un resultat de tipus no contextual, amb independència que per a la seva obtenció hagi calgut la prospecció del context (per exemple, la llista de lemes substantius amb què apareix un lema adjectiu determinat).

L'òptima utilització d'un corpus està en relació directa amb les capacitats de la interfície amb què s'hi accedeix per a l'execució de processos de selecció i per a la presentació dels diferents tipus de resultats possibles.

3.4. Prospeccions en un corpus. Un cas pràctic

Dels diferents tipus d'informació lingüística que ens pot proporcionar un corpus, tal vegada el més representatiu sigui aquell que dóna idea dels contextos en què apareixen els elements lèxics. A més de donar informació sobre la semàntica dels mots, les sortides de resultats amb context permeten l'observació de fenòmens com ara les *col·locacions* en què intervé un mot, les característiques dels arguments que formen part del les seves propietats d'inserció, els règims preposicionals dels seus arguments, etc.

En l'exemple que segueix, extret del CTILC, alguns contextos de l'adjectiu *gruixut* han estat seleccionats en forma de concordances. S'han aplicat uns criteris de selecció consistents a recuperar només els casos en què l'adjectiu *gruixut*,

en qualsevol de les seves formes, va precedit d'un nom substantiu qualsevol, i com a criteri d'ordenació de la informació s'ha especificat que els contextos s'ordenessin a partir del mot que apareix a l'esquerra de la paraula clau.

```
    i cremar- los. Si el xancre es troba sobre una branca gruixuda que convé conservar o bé en el tronc, amb una eina ben
   travessos a nivell (fig. 130), sostinguts per branques gruixudes clavades a terra; són, doncs, pels materials emprats,
     número 51). Convé estellar els troncs i les branques gruixudes fent-ne tions, per a afavorir el bon contacte de la
  químics. El primer consisteix en descorsar les branques gruixudes i la soca a fi que les orugues no s'hi puguin amagar i
    per engalzar- hi hàbilment un teixit dens de branques gruixudes i ara seques. Té el desavantatge que, sobre, el tendal
 i no gasto vergonya i sentiment; si la gàbia té brèdoles gruixudes tant se val que demani acorriment. D'ençà, mestressa,
       les matèries permaneixen llarg temps en el budell gruixut, junt amb les bactèries, i les secrecions albuminoses
 tòxics de reabsorció de productes estancats en el budell gruixut, sobre tot en el cec i còlon ascendent. Aquesta
 per un bon drenatge, per un buidament regular del budell gruixut. Tota la terapèutica quirúrgica gira al voltant d'aquest
     directa de les matèries encara líquides en el budell gruixut anastomosat no deixa de ser gros inconvenient i és
          degut, com diu Nannon, a que el tros de budell gruixut que queda per sota l'anastomosi fins l'anus, sofreix una
   d' absorció. Altres, com Ducbert, creuen que el budell gruixut que hi ha per sobre de l'anastomosi, es deixa dilatar al
    celles espesses, hirsutes, el cap poblat d' un cabell gruixut i dur que es resistia a esdevenir blanc. Però en aquella
  d' anar a Candia a comprar el material necessari: cable gruixut d'acer, corrioles, coixinets, claus, ganxos... Aniré i
    tall i plegat, adequada a la prova de cadenes, cables gruixuts, biguetes, eixos de carruatges i vagons, etc. Premses
   dies de vent, si són penjats del gros feix dels cables gruixuts com braços, el vent hi fa una remor tota estranya, un
     juntes i fortíssimes, amb un gran barret. Una cadena gruixuda, amb un medalló a dins del qual hi havia el retrat de
 remuga; el captaire que simula una marfuga, i el camàlic gruixut, de cara jovial, que ronca endormiscat en un portal.
      Duia els mitjons a la deriva i ensenyava les cames gruixudes, d'una carn blanca i fofa. Aquell militar original
 Candeles llises, primes i petites per al poble; candeles gruixudes i ornades per a la gent de pes; ciris majestuosos i
  més resisteixen el bolcament són les que tenen la canya gruixuda, l'estructura interna de la qual presenta a més, un bon
  barrejar- se als altres i han d' ésser emprats amb capa gruixuda. Les mescles amb blanc opacifiquen el to, mentre que el
   dos paraments del macís a construir, extenían una capa gruixuda (10 a 15 centímetres al menys) de morter o trencadissa
 l' ull amb un instrument òptic, hem d' imaginar una capa gruixuda de teixit transparent, amb un nervi sensible a la llum,
   del Nord fa 1700 anys que va ésser coberta d' una capa gruixuda de cendra volcànica calenta. Hi van penetrar
   nuesa. Al menjador hi ha la taula parada sota una capa gruixuda de pols i més endins, en un dormitori molt vast, amb
   que trenca la discòrdia i, ben embolcallat amb la capa gruixuda de l'escalf de la terra, murmura la lloança, vigila,
aquella més vella, que és la pabordessa, li serva la capa gruixuda i rebel, que a sobre li posen com una disfreça de verd
 fins ara. Aquesta conca potàssica, constituïda per capes gruixudes de sal marina, i també d'altres sals (sosa, potassa,
podia moure les parpelles: s' hi havia adherit dues capes gruixudes de crema platejada. Sota les celles, hi duia una línia
   roques arenisques i pinyolenques, dipositades en capes gruixudes i horitzontals, estableix en les formes del terreny
  mitjà d' altres colors com l' ocre o el blanc (en capes gruixudes o poc barrejat amb blanc, sembla lleugerament negre),
```

L'anàlisi d'aquesta mena de resultats pot proporcionar informacions de gran interès per a la descripció lingüística de les propietats dels mots i de les categories lèxiques. Una anàlisi acurada dels substantius amb què apareix (o *col·loca*) l'adjectiu *gruixut* ens pot ser de gran ajuda de cara a determinar:

- En quin tipus de combinacions lèxiques apareix: per ex. *budell gruixut*.
- Quines són les seves col·locacions: *capa gruixuda*.
- Quins són els diferents significats que adquireix en funció de les característiques del nom amb què apareix: *gruixut* pot significar coses diferents segons els diferents substantius a què s'apliqui.

El fet de poder comptar amb aquest tipus d'informacions, i amb la interfície d'accés al corpus que ens permeti adequar-la a cada cas i a cada necessitat, és d'una importància cabdal en el treball lexicogràfic, ja que permet la descripció sistemàtica del lèxic sobre bases empíriques. Sense la concurrència d'aquestes dades, el lexicògraf hauria de recórrer, per a la realització de la seva tasca, a la introspecció o a la utilització amb caràcter exclusiu d'altres diccionaris.

4. Utilitat dels corpus

4.1. Els corpus en els estudis filològics

Quan hem presentat els diferents tipus de corpus ja hem vist que poden tenir un caràcter sincrònic o diacrònic segons la cobertura cronològica dels materials que contenen. Tot i que la majoria dels treballs sobre corpus i sobre lingüística de corpus se centren en aspectes referits a la descripció sincrònica, la utilització de corpus en estudis de lingüística històrica, dialectologia, estilística, etc. és cada vegada més freqüent. Les característiques generals de disseny de cada corpus, així com el tipus d'anotació de què ha estat objecte, són els factors que determinen aquestes possibilitats d'utilització.

Sovint s'han fet crítiques a la utilització de corpus en els treballs de descripció diacrònica o diatòpica (geolingüística); en concret, s'ha objectat el perill que l'ús de corpus pugui suplantar el coneixement aprofundit de la història de la llengua, o el fet que s'usi per a extreure conclusions representatives sobre un període determinat, sense tenir en compte les seves limitacions respecte als tipus de textos que conté. Malgrat que aquestes crítiques poden estar, segons els casos, més o menys fonamentades, no invaliden la conveniència de la utilització de corpus textuals en lingüística diacrònica. Això no constitueix en sí, però, una objecció metodològica al treball amb corpus. Més aviat cal prendre aquests tipus d'observacions com a advertiment del perill que pot representar el fet de treure conclusions d'un conjunt de dades que, encara que sigui molt voluminós, pot ser que no contingui elements d'anàlisi decisius.

La dimensió diacrònica, social, etc. d'un corpus pot propiciar la seva utilització com a font de diccionaris etimològics o històrics o per a la producció de materials dialectològics.

D'altres utilitzacions més concretes dels textos d'un corpus consisteixen a donar suport a tasques que són objecte habitual de preocupació filològica: establiment d'autories a partir d'anàlisis textuals o estilístiques, edicions crítiques, datacions d'aparició de mots, etc.

4.2. Els corpus en els estudis lingüístics

L'anomenada lingüística de corpus ha experimentat, de manera paral·lela al desenvolupament de les tècniques de processament de corpus, un creixement en els darrers anys que l'ha situada entre les grans disciplines de la lingüística aplicada. L'objectiu de la lingüística de corpus és la prospecció i el processament de corpus per a la descripció, a partir de dades objectives, de les estructures i de les categories (sintàctiques, lèxiques, morfològiques, etc.) de la llen-

gua. Un corpus serveix, així, com a element de contrast de les hipòtesis del lingüista i, alhora, com un element que pot conduir determinades recerques lingüístiques, per la immediatesa dels tipus d'evidències que proporciona.

A més de donar resultats dins de l'àmbit de la lingüística de corpus, cal assenyalar també quina pot ser la contribució de l'explotació d'un corpus en diferents àrees de la lingüística. Per exemple, les capacitats que pot tenir un corpus (en funció de com hagi estat concebut o dissenyat) de representar la variació lingüística (en relació als diferents estrats socials dels parlants, o als diferents registres comunicatius) el converteixen en una eina indispensable per a qualsevol tipus de recerca referent a gramàtica, semàntica, pragmàtica, sociolingüística, dialectologia, etc.

4.3. Els corpus i la lexicografia

Durant els darrers decennis la lexicografia ha experimentat una important renovació metodològica que ha influït decisivament en la manera d'elaborar els diccionaris; l'aplicació d'aquests nous mètodes ha donat ja fruits en algunes llengües amb l'elaboració de nous i de millors diccionaris. Les dades empíriques proporcionades pels corpus informatitzats constitueixen un dels elements fonamentals d'aquesta renovació, i han incidit de manera decisiva en la millora de les capacitats descriptives dels diccionaris. Els corpus ens informen sobre l'ús real dels mots en el discurs i permeten de caracteritzar-los adequadament des del punt de vista lingüístic.

Un dels usos d'un corpus que es pot considerar bàsic en l'activitat del lexicògraf és l'anàlisi de les concordances d'un mot o conjunt de mots, a fi de determinar-ne els seus diferents aspectes (comportament sintàctic, sentits bàsics, etc.); en aquest sentit, per a la realització de diferents projectes lexicogràfics moderns s'han desenvolupat *estacions de treball lexicogràfic*, que són entorns integrats de treball informàtic en els quals el lexicòggraf té accés a corpus i d'altres recursos lingüístics essencials per a la realització del seu treball.

Però l'anàlisi semàntica de les concordances d'un mot (a fi de determinar-ne els sentits diferenciats, o d'establir la freqüència dels diferents usos i sentits), o l'establiment dels seus patrons de comportament sintàctic no són les úniques aportacions que la utilització de corpus pot fer a la lexicografia. Els corpus poden tenir també una importància determinant en la selecció del conjunt d'entrades que formen part d'un diccionari (anomenat **nomenclatura**) ja que, tal com hem vist en l'apartat dedicat a la informació estadística, permeten de treure conclusions sobre la *importància* d'un element lèxic determinat en el conjunt del lèxic de la llengua. Més enllà d'aquestes utilitzacions, els corpus també permeten de discernir entre dos conjunts: el del **lèxic real** (o **vocabulari** d'una llengua) i el del **lèxic virtual**. El primer integra els elements que realment són o han estat utilitzats en els actes de parla d'una llengua, mentre que el segon està format per tots aquells elements lèxics *possibles* d'obtenir a partir dels processos de derivació propis de la llengua. Els diccionaris recullen,

sovint, elements del segon conjunt que no tenen de fet realització pràctica en el primer; l'ús d'un corpus suficientment representatiu pot corregir aquesta mena de desviacions.

4.4. Els corpus i el processament del llenguatge natural

Els corpus han tingut i tenen un paper decisiu en el desenvolupament d'aspectes relacionats amb el Processament del Llenguatge Natural (PLN). Sota la denominació global de PLN s'agrupen les branques de la lingüística computacional, que constitueix una de les línies bàsiques de desenvolupament de la denominada **enginyeria lingüística**. Considerat com a *recurs lingüístic*, un corpus és un recurs de tipus primari, que proporciona informacions de base que tenen una importància decisiva en el desenvolupament de recursos secundaris. Els recursos secundaris (lèxics electrònics, diccionaris, etc.) són productes més elaborats que poden ser utilitzats modularment en aplicacions desenvolupades per l'enginyeria lingüística que serveixen per a ser integrades en productes i serveis informatitzats en els quals intervenen mòduls que tenen com a funció principal l'anàlisi o la generació de llenguatge natural.

El grau de desenvolupament de les eines informàtiques per al processament de corpus fa que sigui cada cop més factible la constitució de corpus més i més grans, els quals permeten, al seu torn, el desenvolupament d'aplicacions computacionals més potents i elaborades, algunes de les quals estaran orientades a permetre l'etiquetació de volums d'informació cada vegada més importants amb menys esforç.

Els corpus s'integren, així, en la dinàmica de reutilització de recursos de les anomenades indústries de la llengua, intervenint com a components en el desenvolupament d'aplicacions relacionades amb les tecnologies del llenguatge. Així doncs, cada cop es tenen més en compte en la constitució de corpus els aspectes relacionats amb l'estandardització de les eines de processament i dels formalismes d'etiquetatge.

Resum

Un corpus és una col·lecció d'elements lingüístics seleccionats i ordenats d'acord amb criteris lingüístics explícits amb la finalitat de ser usat com a mostra de la llengua. El desenvolupament de corpus ha adquirit, en els darrers decennis, una importància decisiva. Les actuals possibilitats de tractament informatitzat permeten que un gran volum de dades lingüístiques siguin analitzades i processades en un temps cada vegada més breu, i això ha propiciat el desenvolupament de tot tipus de corpus, entre els quals destaquen els anomenats *corpus de referència*, que pretenen aportar dades representatives per a una anàlisi de qualsevol aspecte de la llengua.

La representativitat d'un corpus està determinada pel disseny i per l'equilibri intern. Un corpus s'organitza a partir d'una sèrie de paràmetres (cronològics, mediàtics, etc.) que giren al voltant de dos tipus de criteris fonamentals: els criteris externs (que depenen de factors extratextuals) i els criteris interns (determinables a partir de l'anàlisi del text o del seu suport). En funció dels valors que dissenyem per a cada paràmetre, podem establir classificacions tipològiques dels corpus.

La naturalesa dels resultats que pot proporcionar un corpus varia en funció del procés a què el sotmetem. Hi ha un estret lligam entre el processament de corpus, d'una banda, i els mitjans informàtics i els tractaments computacionals utilitzats en la seva constitució, de l'altra. L'etiquetatge d'un corpus, sovint realitzat a partir de processos semiautomàtics o automàtics, és un dels aspectes fonamentals que en determina les possibilitats d'explotació i la fiabilitat de les dades. Els corpus poden proporcionar, també, informacions de caràcter quantitatiu (de vegades publicades en forma de diccionaris de freqüències), les quals permeten extreure conclusions de gran interès sobre l'estructura del vocabulari.

L'anàlisi de la contribució dels corpus a diferents disciplines relacionades amb el llenguatge dóna una idea de la seva importància en els diferents àmbits de recerca lingüística. El fet que aquesta aportació es produeix en terrenys molt diversos, fa pensar que els corpus textuals s'han consolidat plenament com un dels principals recursos de què pot disposar avui la lingüística per a l'obtenció de dades empíriques.

Activitats

1. Mira de buscar a Internet informació sobre els diferents corpus que s'han elaborat o estan en fase d'elaboració en l'àmbit de les llengües europees. Mira de fer una llista de llengües i de recursos.

2. Busca informació sobre els diferents corpus elaborats en català, i estableix-ne les característiques principals.

3. Intenta consultar un corpus qualsevol accessible públicament a Internet, i familiaritza't amb la seva interfície de consulta. Mira d'extreure un llistat de concordances d'un mot qualsevol.

Exercicis d'autoavaluació

1. Què volem dir quan diem que un corpus és *representatiu* d'una llengua determinada? Quines són les característiques bàsiques que ha de reunir un corpus per poder ser considerat representatiu?

2. Quins són els criteris interns i el criteris externs que són a la base dels diferents paràmetres de disseny d'un corpus? Quines són les tendències actuals en el disseny de corpus en relació a aquests criteris?

3. Què diferencia un corpus estructurat i organitzat d'una simple reunió de textos sobre suport magnètic?

4. Les diferents aplicacions d'un corpus en diferents aspectes de la lingüística teòrica i aplicada.

5. Processos i resultats en l'explotació d'un corpus; diferents dades que es poden obtenir de la prospecció d'un corpus, utilitat d'aquestes dades.

6. Indiqueu les principals aplicacions d'un corpus oral en l'estudi de la fonètica.

7. Indiqueu les principals aplicacions d'un corpus de llengua oral tal com s'entenen aquests recursos en l'àmbit de la lingüística de corpus.

8. Assenyaleu les principals diferències entre els corpus desenvolupats en l'àmbit de la fonètica i les tecnologies de la parla i els que s'han creat per a l'estudi de la llengua oral.

Solucionari

Exercicis d'autoavaluació

1. La representativitat dels corpus és un concepte que, des del punt de vista metodològic, és a la base de la lingüística empírica. Un corpus que hagi de ser considerat com un *corpus de referència* ha de reunir unes característiques determinades que es poden considerar més o menys objectives. No es pot parlar, però, de representativitat al marge de la consideració de la finalitat amb què ha estat constituït un corpus.

2. L'exposició dels diferents criteris i paràmetres de disseny de corpus es fa a l'apartat 2. En l'actualitat hi ha una clara preferència pels criteris externs en la constitució de corpus; aquesta tendència està motivada pel fet que els criteris externs són considerats més objectius.

3. El disseny d'un corpus obeeix a una determinada hipòtesi de distribució de tipologies textuals; a més, en un corpus solen ésser expressades de manera explícita determinades informacions textuals i lingüístiques mitjançant l'etiquetatge.

4. Les possibilitats d'explotació diacrònica d'un corpus determinen que pugui ésser utilitzat en lingüística històrica. En lingüística sincrònica, els corpus són una eina habitual. En àrees específiques com la lexicografia, el corpus esdevé un element de descripció indispensable. Finalment, els corpus són utilitzats àmpliament en processament del llenguatge natural.

5. Processos i resultats amb context i sense context; els tipus KWIC i KWOC d'índexs de resultats. Descripció del llistat de concordances, possibilitats de presentació etc. Les diferents possibilitats que ofereix l'anàlisi dels resultats d'un corpus.

6. Estudi fonètic: descripció articulatòria; descripció acústica; descripció dels elements segmentals; descripció dels elements suprasegmentals. Fonètica aplicada: fonètica contrastiva, adquisició de segones llengües, adquisició i desenvolupament de primeres llengües, producció de la parla, patologies de la parla, sociolingüística, dialectologia.

7. Anàlisi del discurs; anàlisi de la conversa; sociolingüística; dialectologia; estudi dels registres; estudi dels estils; estudi d'àmbits específics (discurs polític, acadèmic, dels joves, etc.); anàlisi gramatical de la llengua oral; ensenyament de llengües.

8. Corpus per a la fonètica i les tecnologies de la parla: contenen l'enregistrament del senyal sonor; transcripció ortogràfica, fonètica, fonològica i prosòdica; alguns es recullen en les condicions específiques que exigeix una determinada aplicació (per telèfon, amb soroll de fons, imitant un diàleg amb un sistema informàtic); els corpus per a l'estudi fonètic es recullen en situacions controlades en un laboratori. Corpus per a l'estudi de la llengua oral: contenen la transcripció ortogràfica complementada amb marques que reflecteixen l'ús oral de la llengua (torns de paraula, superposició de parlants, interrupcions del discurs, etc..); intenten representar l'ús espontani de la llengua i es recullen en contextos el més naturals possibles.

Glossari

corpus *m* Col·lecció d'elements lingüístics seleccionats i ordenats d'acord amb criteris lingüístics explícits amb la finalitat de ser usat com a mostra de la llengua.

corpus automatitzat (o informatitzat) *m* Corpus que s'ha codificat de manera homogènia per a diferents tasques de recuperació de la informació.

corpus de referència *m* Corpus dissenyat per a proporcionar informació sobre els diversos aspectes d'una llengua, de manera que en representa totes varietats de registre, de tipus de discurs, de vocabulari, etc.

corpus oral *m* Corpus que es constitueix a partir d'un enregistrament de la llengua parlada i que conté l'enregistrament original o una transcripció.

etiquetatge *m* Informació afegida al text que explicita, en forma de categories lingüístiques o textuals, característiques del text o dels mots que en formen part, que d'altra manera romandrien implícites tal com es produeixen en els enunciats lingüístics corrents.

freqüència absoluta *f* Nombre absolut de vegades que apareix un element dins del corpus.

freqüència relativa *f* Nombre relatiu de vegades que apareix un element en relació a l'extensió total del corpus, expressat generalment en tant per 100, tant per 1.000, etc.

lematització *f* Assignació, en forma d'etiqueta, de lema (o forma canònica) a un mot tal com el trobem en el discurs textual.

Bibliografia

Bibliografia bàsica

Biber, D.; Conrad, S.; Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Lamel, L.; Cole, R. A. (1997). "Spoken Language Corpora". A: Cole, R. A.; Mariani, J.; Uszkoreit, H.; Zaenen, A.; Zue, V. (Eds). *Survey of the State of the Art in Human Language Technology*. Cambridge: Cambridge University Press. pp. 450-454.

Rafel i Fontanals, Joaquim (1994). "Un corpus general de referència de la llengua catalana". *Caplletra* (17., tardor, 219-250.). València.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Bibliografia complementària

Gibbon, D.; Moore, R.; Winski, R. (eds.) (1998). *Spoken Language Sustems and Corpus Design*. Berlin: Mouton De Gruyter. (Handbook of Standards and Resources for Spoken Language Systems, Volume I).

Leech, G.; Wilson, A. (1996). EAGLES *Recommendations for the Morphosyntactic Annotation of Corpora*. s. l.: EAGLES Document EAG-TCWG-MAC/R.

Llisterri, J. (1996). "Els corpus lingüístics orals". A: Payrató, Ll.; Boix, E.; Lloret, M. R.; Lorente, M. (eds.). *Corpus, Corpora*. Actes del 1er i 2on Col.loquis Lingüístics de la Universitat de Barcelona (CLUB-1, CLUB-2). Barcelona: Promociones y Publicaciones Universitarias SA. pp. 27-70.

Llisterri, J. (1999). "Los corpus orales. Introducció als corpus lingüístics 1998/99". Departament de Filologia Espanyola, Universitat Autònoma de Barcelona.

Moreno Fernández, F. (1997). "La formación de corpus de lengua hablada". A. Moreno Fernández, F. (ed.). *Trabajos de sociolingüística hispánica*. Alcalá de Henares: Universidad de Alcalá, Servicio de Publicaciones (Ensayos y Documentos, 27) pp. 93-114.

Torruella, J.; Llisterri, J. (1999). "Diseño de corpus textuales y orales". A: Blecua, J. M.; Clavería, G.; Sánchez, C.; Torruella, J. (eds.). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio. pp. 45-77.

Ooi, Vincent B. Y. (1998). *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Rafel i Fontanals, Joaquim (1996). "Introducció". *Diccionari de freqüències 1. Llengua no literària (I-CLIII)*. Barcelona.

Sinclair, J. (1996). *EAGLES Preliminary Recommendations on Corpus Typology*. S. l.: EAGLES Document EAG-TCWG-CTYP/P.

Sinclair, J. McH.; Ball, J. (1996). *EAGLES Preliminary Recommendations on Text Typology*. s. l.: EAGLES Document EAG-TCWG-TTYP/P.

Sperberg-McQueen, C. M.; Burnard, L. (eds.) (1994). *Guidelines for Electronic Text Encoding and Interchange*. Chicago-Oxford: Text Encoding Initiative.

Vidal Beneyto, J. (ed.) (1991). *Las industrias de la lengua*. Madrid: Fundación Germán Sánchez Ruipérez.