

# El processament de corpus II: exemples pràctics d'exploració i ús

Judith Domingo Mañosa

PID\_00233515

---

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





# Índex

<b>Introducció</b> .....	5
<b>Objectius</b> .....	6
<b>1. Informació i formats</b> .....	7
1.1. Corpus escrits .....	7
1.2. Corpus orals .....	12
1.3. Corpus signats (de llengua de signes) .....	13
<b>2. Interfícies de consulta de corpus</b> .....	16
2.1. CTILC: Corpus Textual Informatitzat de la Llengua Catalana ....	16
2.2. AnCora .....	17
2.3. CucWeb .....	19
<b>3. Exploració d'un corpus</b> .....	22
3.1. Instal·lació de programari .....	22
3.2. Descripció de la interfície de cerca .....	22
3.3. Llistes de paraules ( <i>word list</i> ) .....	23
3.4. Concordances .....	25
3.5. Concordances gràfiques ( <i>concordance plot</i> ) .....	26
3.6. Agrupaments de segments (clústers i <i>n-grames</i> ) .....	27
3.7. Col·locacions ( <i>collocates</i> ) .....	28
<b>Resum</b> .....	30
<b>Activitats</b> .....	31
<b>Glossari</b> .....	32
<b>Bibliografia</b> .....	33





## Introducció

Ja hem vist al mòdul "El processament de corpus I: la lingüística empírica" que qualsevol estudi d'una llengua, si es vol basar en la realitat empírica, cal que disposi de gran número de dades que donin compte d'aquesta llengua. Precisament, un corpus és el recurs adequat per a fer aquests estudis ja que són col·leccions d'elements lingüístics que pretenen ser una mostra real de la llengua. Per tant, són recursos molt utilitzats en diverses disciplines: en estudis filològics s'empra per a estudis diacrònics; en estudis lingüístics per a la descripció sincrònica de la llengua; en estudis lexicogràfics per a la millora de diccionaris o en el processament del llenguatge natural per al desenvolupament de recursos secundaris (etiquetadors morfològics, sintàctics, etc.).

Actualment, disposem de molts corpus al nostre abast per poder fer els nostres estudis lingüístics tot i que de vegades no queda més remei que fer el nostre propi corpus. En aquest mòdul pretenem abordar tant els corpus existents en català com l'exploració d'un corpus propi. Concretament, a la primera part del mòdul, observarem com es codifiquen els corpus textuais, orals i de llengua de signes i quina informació lingüística i extralingüística podem trobar-hi. A la segona part veurem i utilitzarem algunes interfícies existents de corpus textuais en català. I a la tercera part, aprendrem a explorar el nostre propi corpus amb eines d'exploració de corpus, obtenint resultats textuais (exemples de frases) i estadístics.

Aquest mòdul, malgrat tenir coherència interna, s'ha concebut com a complementari del mòdul d'"El processament de corpus I: la lingüística empírica". Farem una aproximació als corpus més pràctica però cal haver assolit els coneixements teòrics del mòdul anterior. Com a mínim, cal haver assimilat els coneixements dels apartats 1.1 ("Conceptes fonamentals"), 2 ("Tipologia") i 3 ("Processament de corpus").

## Objectius

Els objectius que cal assolir mitjançant aquest mòdul didàctic són:

- 1.** Conèixer els formats que poden tenir els corpus i examinar la codificació de la informació que contenen.
- 2.** Familiaritzar-se amb les interfícies de corpus i analitzar les diverses presentacions dels resultats.
- 3.** Aprendre a utilitzar una eina d'exploració de corpus aplicant els coneixements adquirits en aquest mòdul i en el mòdul de *El processament de corpus I*.
- 4.** Extraure dades quantitatives i qualitatives d'un corpus propi.

## 1. Informació i formats

L'objectiu d'aquest apartat és veure exemples que il·lustrin els tipus d'anotació que poden tenir els corpus. Dividirem els tipus de corpus seguint els criteris del mòdul "El processament de corpus I: la lingüística empírica" en corpus escrits i corpus orals, però afegirem també els corpus en llengua de signes, ja que per la seva pròpia naturalesa segueixen uns criteris d'anotació ben diferents dels corpus orals o escrits.

### Vegeu també

Podeu consultar els criteris de divisió dels corpus a l'apartat 2 del mòdul "El processament de corpus I: la lingüística empírica".

### 1.1. Corpus escrits

Els corpus escrits, depenent del seu objectiu, poden tenir diversos tipus d'informació. Dividirem aquesta informació entre informació lingüística i metadades (no són excloents, un mateix corpus pot tenir metadades i informació lingüística).

#### 1) Amb informació lingüística

Com a exemple de corpus escrits ens fixarem en l'AnCora ja que actualment, per al català, és un dels corpus que conté més informació lingüística. Ha estat desenvolupat pel centre CLIC de la Universitat de Barcelona, el Grup de Processament del Llenguatge Natural de la Universitat Politècnica de Catalunya i el Lengoia Naturalaren Prozesamendurako Ixa Taldea de la Universitat del País Basc. El corpus complet comprèn tres llengües: català, castellà i basc, però nosaltres només ens fixarem en el català.

### Lectura complementària

M. A. Martí; M. Taulé; L. Márquez; M. Bertran (2007). "Ancora: A Multilingual and Multilevel Annotated Corpus".

L'AnCora en català (AnCora\_CA) conté 488.380 paraules en 16.788 frases i està anotat amb informació morfològica, sintàctica, de dependències, semàntica i de coreferències.

Totes aquestes anotacions s'han fet amb diverses eines d'anotació automàtica però a més també s'han revisat manualment, per això és tan important, ja que és un corpus d'una mida important amb unes anotacions de bona qualitat. Quant al gènere dels seus textos, aquests són majoritàriament periodístics<sup>1</sup>.

<sup>(1)</sup> Les fonts són *El Periódico*, l'agència EFE i l'Agència Catalana de Notícies (ACN).

Per entendre com es codifica la informació lingüística en un corpus, observem un fragment<sup>2</sup> del corpus AnCora:

<sup>(2)</sup> Per motius didàctics, es reproduïx una simplificació del corpus original.

Taula 1. Fragment del corpus AnCora

P	Forma	Lema	C	Trets	D	F	Etiq	Rol
1	Per	per	s	postype=preposition	7	cc	sps00	argM-cau

2	Aquest	aquest	d	postype=demonstrative   gen=m   num=s	3	spec	dd0ms0	-
3	motiu	motiu	n	postype=common   gen=m   num=s	1	sn	ncms000	-
4	,	,	f	punct=comma	1	_	f	_
5	el	el	d	postype=article   gen=m   num=s	6	spec	da0ms0	-
6	jurat	jurat	n	postype=common   gen=m   num=s	7	subj	ncms000	arg1-tem
7	està	estar	v	postype=main   num=s   person=3   mood=indicative   tense=present	ROOT	sentence	vmip3s0	-
8	format	format	a	postype=qualificative   gen=m   num=s   posfunction=participle	7	atr	aq0msp	-
9	per	per	s	postype=preposition	8	cag	sps00	arg0-agt
10	coneixedors	coneixedor	n	postype=common   gen=m   num=p	9	sn	ncmp000	-
11	de	de	s	postype=preposition	10	sp	sps00	-
12	cada	cada	d	postype=indefinite   gen=c   num=s	13	spec	di0cs0	-
13	estil	estil	n	postype=common   gen=m   num=s	11	sn	ncms000	-
14	.	.	f	punct=period	7	_	f	_

a) A la primera columna trobem la posició que ocupa cada element dins de l'oració.

b) A la segona columna trobem la forma, la paraula tal com apareix en el corpus.

c) A la tercera columna trobem el lema (forma no flexionada de la paraula).

d) A la quarta columna trobem la categoria gramatical major, aquí teniu les correspondències dels codis:

n > nom

d > determinant

s > preposició

f > puntuació

v > verb

a > adjectiu

e) A la cinquena columna hi ha els trets de cada paraula, determinats per la categoria que tenen. Per exemple, les paraules amb categoria d (determinant) tenen subtrets diferenciats ja que cada determinant és de tipus diferent (demostratiu, indefinit i article):

Taula 2. Determinants en el fragment del corpus AnCorà

aquest	d	postype=demonstrative   gen=m   num=s
--------	---	---------------------------------------

el	d	postype=article gen=m num=s
cada	d	postype=indefinite gen=c num=s

f) A la sisena columna trobem les dependències, és a dir, les relacions sintàctiques entre els mots de la frase. Per veure l'anàlisi sintàctica cal que relacionem la posició que té l'element a la frase (columna 1) amb la relació que guarda amb els altres elements (columna 6), és a dir, la relació de dependència entre els elements de l'oració. De moment, observem només una part de l'oració per veure com es relacionen els elements.

Taula 3. Sintagma nominal *Per aquest motiu* del corpus AnCora

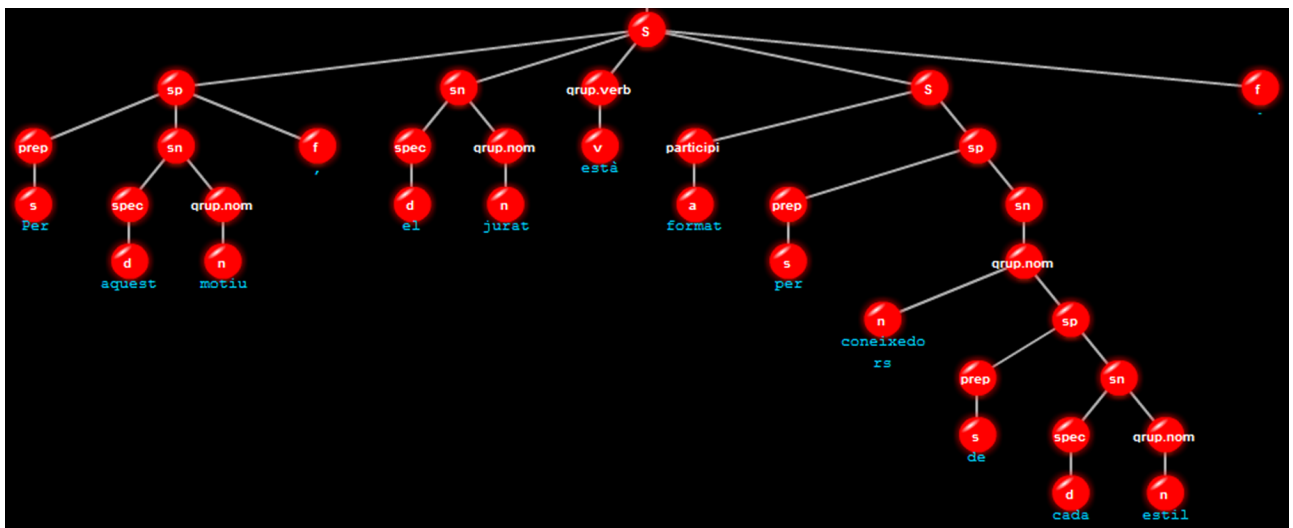
1	Per	per	s	Postype=preposition	7	cc	sps00	argM-cau
2	aquest	aquest	d	Postype=demonstrative gen=m num=s	3	spec	dd0ms0	-
3	motiu	motiu	n	Postype=common gen=m num=s	1	sn	ncms000	-

Si mirem la sisena columna, veiem que *aquest* (posició 2 a la frase) té una relació de dependència amb *motiu* (3), és a dir, el determinant depèn del nom i *motiu* (posició 3 a la frase) té una relació de dependència amb *Per* (1). Podem dir que el nostre sintagma preposicional té l'estructura següent:

[aquest (3) > motiu (1)] > per

Resseguint aquestes relacions de dependència podem obtenir l'arbre de la frase:

Figura 1. Anàlisi sintàctica del corpus AnCora



L'últim element de la frase (el node principal) és el verb conjugat que rep l'atribut ROOT (observeu que és l'única que no té un número).

g) A la setena columna trobem la informació sobre les funcions sintàctiques; només s'etiqueta la funció sintàctica en el nucli del sintagma. Per exemple, en el sintagma que hem vist abans (*Per aquest motiu*), només el nucli del sintagma *per* té etiquetada la funció sintàctica de *cc* (complement circumstancial).

En aquells elements que no són nuclis, només trobem la relació sintagmàtica, per exemple, *aquest* és un *especificador* (*spec*) de *motiu* i a la vegada *motiu* forma un *sintagma nominal* (*sn*) complement de *per*.

h) A la vuitena columna trobem una codificació de la categoria morfològica i els trets corresponents, és a dir, una codificació de les columnes 4 i 5.

i) A la novena columna, trobem els rols semàntics de l'oració.

El verb *formar* assigna els papers temàtics següents:

**Tema (arg1-tem)** a *El jurat*

**Causa (argM-cau)** a *Per aquest motiu*

**Agent (arg0-agt)** a *Per coneixedors de cada estil*

## 2) Amb metadades

Hem vist ja que els corpus poden contenir informació lingüística però sovint també poden contenir metadades; és a dir, informació extralingüística que parla del mateix text: any, autor, títol, etc.

El corpus de mostra que farem servir és del projecte BancTrad. El projecte BancTrad consistí en la creació d'una interfície per a corpus paral·lels i monolingües, amb finalitats tan diverses com la didàctica de la traducció o la recerca lingüística. El corpus paral·lel té 3.000.000 de paraules i comprèn les llengües següents: català, castellà, anglès, francès i alemany (les traduccions són del català o del castellà cap a les altres llengües).

Vegeu a continuació com es codifiquen les dades:

```
<text lar="de" lpa="ca" for="web" ftr="web" prof="cape" dif="m" reg="es"
esp="b" tem="g" tiptxt="ss" datorig="1950" dattrad="2000" autororig="ss"
traductor="ss" titorig="ss" tittrad="ss">
<s id="cade7193-1>
```

Llatinoamèrica	Llatinoamèrica	Nom
busca	buscar	Verb
la	el	Det
seva	seu	Adj
esquerra	esquerra	Nom

.	.	.
---	---	---

</s>

</text>

A BancTrad les dades lingüístiques també es codifiquen en format tabular (forma, lema i categoria morfològica). En canvi, les metadades tenen format XML i són a nivell de text, no pas de paraula.

Observem les metadades que ofereix BancTrad:

Taula 4

Atributs	Valors
lar (llengua d'arribada)	de (deutch)
lpa (llengua de partida)	ca (català)
for (font original)	web
fttr (font de la traducció)	web
prof (professor)	cape (carles pérez)
reg (registre)	es (estàndard)
dif (grau de dificultat)	m (mitjana)
esp (especialització)	b (baix)
tem (temàtica)	g (general)
tiptxt (tipus de text)	ss (sense especificar)
datorig (data de l'original)	1950
dattrad (data de la traducció)	2000
autororig (autor original)	ss (sense especificar)
traductor (traductor)	ss (sense especificar)
titorig (títol de l'obra original)	ss (sense especificar)
tittrad (títol de la traducció)	ss (sense especificar)

Les metadades estan formades per atributs i valors associats, els atributs sempre són els mateixos per a tots els textos però els valors varien; per exemple, podem tenir diverses llengües d'arribada (alemany, anglès, etc.), també podem tenir fonts diferents, etc.

## 1.2. Corpus orals

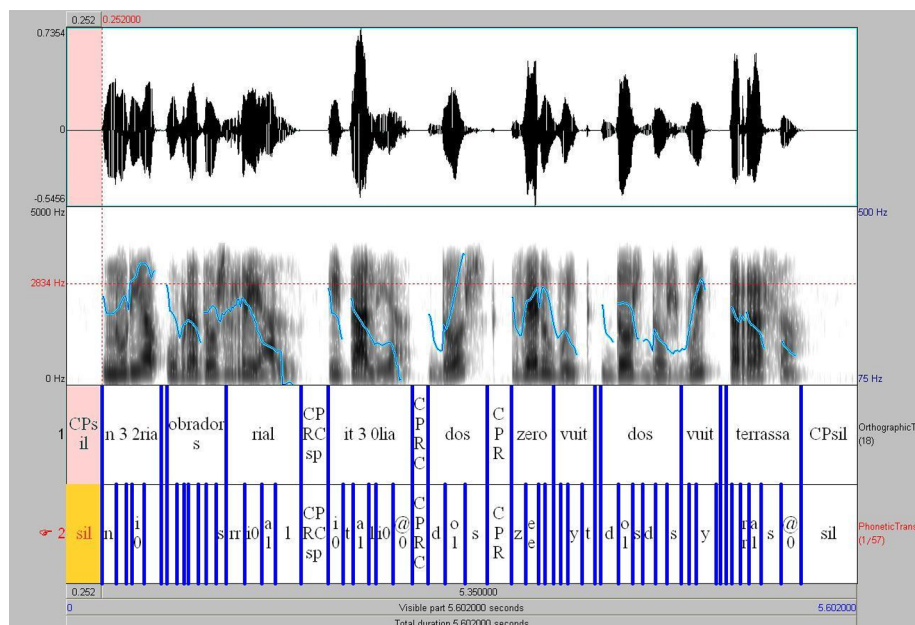
Tal com hem descrit anteriorment, podem distingir els corpus orals entre els que pertanyen a la lingüística de corpus i els que pertanyen a les tecnologies de la parla. El corpus que veurem aquí s'inclou dins de l'àmbit de les tecnologies de la parla. Es tracta d'un corpus que van desenvolupar Barcelona Media - Centre d'Innovació i Cereproc. L'objectiu era crear un conversor text-parla en català i en castellà i, a més, crear una veu sintètica. Es van gravar 4 hores d'àudio en català i 4 en castellà amb un locutor bilingüe que tingués bona competència en les dues llengües, que no tingués un accent determinat, que tingués bona prosòdia i capacitat d'interpretació.

### Vegeu també

La tipologia dels corpus s'ha tractat al subapartat 4.2.1 mòdul didàctic "El processament de Corpus I: la lingüística empírica".

Vegem una mostra de les anotacions del corpus amb el programa Praat:

Figura 2. Anotador de corpus orals Praat



### Programa Praat

Per a més informació sobre Praat, consulteu la seva web: <http://www.praat.org>.

A la imatge hi ha la seqüència:

*NÚRIA OBRADORS RIAL; Itàlia, dos, zero vuit dos dos vuit, Terrassa.*

La pantalla de Praat està dividida en quatre parts: a la primera part hi ha l'ona sonora; a la segona, el sonograma; a la tercera, la transcripció ortogràfica del segment, i, finalment, la transcripció dels fonemes amb un inventari propi (similar al SAMPA). Programes com Praat permeten veure tota la informació dels segments d'un corpus orals, això fa que siguin molt útils per fer revisions de les anotacions.

### SAMPA

És un alfabet fonètic llegible per ordinador, desenvolupat per al projecte ESPRIT. Està basat en l'alfabet fonètic internacional (IPA), però només fa servir caràcters ASCII de 7 bits. Per a més informació, vegeu la seva web. <http://www.phon.ucl.ac.uk/home/sampa/index.html>.



Un altre format de visualització del corpus és en text pla. Observeu a la taula següent les columnes en què tenim informació similar a la que hem vist amb el programa Praat (pertany a la paraula *vuit* del segment mostrat anteriorment):

Taula 5. Fragment de corpus anotat Praat

Inici-final del fonema	Transcripció del fonema	Duració	Mot
3.452000 3.502000	B	3.472000	vuit
3.502000 3.552000	u1	3.522000	vuit
3.552000 3.662000	J	3.602000	vuit
3.662000 3.752000	D	677000	vuit

- A la primera columna trobem el temps d'inici i de finalització del fonema dins de l'enregistrament (en mil·lisegons).
- A la segona columna, tenim la transcripció del fonema amb l'inventari propi.
- A la tercera columna, hi ha la duració del fonema (també en mil·lisegons).
- A la quarta columna, trobem la paraula (escrita ortogràficament) de la qual forma part el fonema.

El resultat d'aquest projecte fou el conversor, esmentat anteriorment, i la veu de la Mar, una veu femenina que no està especialitzada en cap domini i que es pot fer servir, per exemple, en els missatges que anuncien parades en el transport públic, en els serveis d'atenció al client que ens guien en processos de reclamacions, etc.

### 1.3. Corpus signats (de llengua de signes)

Fins ara hem vist corpus orals i corpus escrits, però també cal tenir en compte els corpus lingüístics signats, en llengua de signes. Actualment, hi ha una gran disponibilitat de corpus en llengües orals, però el desenvolupament de corpus per a les llengües de signes porta un endarreriment força notable, causat per dificultats com:

- La falta de sistemes estàndard de representació de llengües de signes.
- La simultaneïtat de trets (mans, expressions facials, ulls, etc.).
- La manca d'aprofundiment en l'anàlisi de l'estructura lingüística de les llengües de signes.

Com a exemple, volem mostrar el corpus ECHO. ECHO és un corpus que es va desenvolupar dins del marc del projecte europeu que rep el mateix nom. L'objectiu d'aquest projecte era crear un corpus en diferents llengües de signes europees.

El corpus ECHO consisteix en anotacions de les llengües de signes holandesa (NGT), britànica (BSL) i sueca (SSL). Per a cadascuna d'aquestes llengües es van gravar cinc faules, un petit lèxic i entrevistes amb signants. A més també es van incloure textos poètics per a la llengua de signes britànica i sueca.

La informació que es va anotar fou la següent:

- traducció de l'anglès, del suec o de l'holandès;
- glossa (tenint en compte la mà dreta i l'esquerra per separat);
- repetició (mà dreta i esquerra per separat);
- direcció i localització espacial (mà dreta i mà esquerra per separat);
- moviment i posició del cap;
- posició de les celles;
- obertura d'ulls;
- direcció de la mirada;
- forma de la boca;
- moviment de les galtes;
- rol (anotacions sobre l'estil directe).

L'eina que es va fer servir per a l'anotació fou ELAN. Vegeu la imatge següent.

Figura 3. Eina d'anotació de corpus ELAN

The screenshot shows the ELAN software interface. At the top, there is a menu bar with 'File', 'Edit', 'Search', 'View', 'Options', and 'Help'. Below the menu is a video window labeled 'CAM 3' showing a woman signing. To the right of the video is a control panel with tabs for 'Grid', 'Text', 'Subtitles', and 'Controls'. The 'Text' tab is active, showing a grid of annotations. The grid has columns for time (00:00:14.000 to 00:00:21.000) and rows for different annotation types: Translation Dutch, Translation English, Gloss RH, Mouth, Gloss RH English, Gloss LH English, and Gloss RH. The annotations include Dutch and English translations, glosses for the right and left hands, and mouth movements.

	00:00:14.000	00:00:15.000	00:00:16.000	00:00:17.000	00:00:18.000	00:00:19.000	00:00:20.000	00:00:21.000
Translation Dutch	emand te zien.		Hij rende snel de winkel in, pakte het bot en rende er zo snel als hij kon mee weg. Hij rende ver weg tot aan de br					
Translation English	nobody there.		He ran into the shop, took the bone and took off as fast as he could. He ran far away up to the bridge.					
Gloss RH English	NOTHING		(p-) running dog	CATCH	(p-) running d	(p-) dog disappears	BRIDGE	(p-) run
Gloss LH English	NOTHING		(p-) running dog		(p-) running d		BRIDGE	
Gloss RH	NIETS		(p-) rennen hond	GRJUPEN	(p-) rennen ho	(p-) hondje verdwijnen in d	BRUG	(p-) ren

L'ús de corpus en llengua de signes són interessants no només per estudiar intrínsecament les llengües signades sinó per desenvolupar traductors (llengua signada a llengua oral o a la inversa), avatars, etc. que són essencials per a fer més accessible la informació a la comunitat de signants.

## 2. Interfícies de consulta de corpus

En aquest apartat ens centrarem en les interfícies de corpus en català. Veurem detalladament les interfícies del CTILC, l'AnCora i de BancTrad.

### 2.1. CTILC: Corpus Textual Informatitzat de la Llengua Catalana

El CTILC (Corpus Textual Informatitzat de la Llengua Catalana: <http://ctilc.iec.cat/>) és un corpus desenvolupat per l'Institut d'Estudis Catalans per crear el *Diccionari descriptiu de la llengua catalana*. Conté més de 52 milions de paraules i conté textos literaris i no literaris del 1832 al 1988.

#### Vegeu també

Trobareu més informació del CTILC al subapartat 2.2.2 del mòdul "El processament de Corpus I: la lingüística empírica".

Observem-ne la interfície:

Figura 4. Interfície de consulta del corpus CTILC

Podem buscar per forma o per lema. En el nostre exemple, buscarem el lema *casar*. A continuació, marquem quin lema volem i premem *Seleccionar*. Ens apareixerà la pantalla següent:

Figura 5. Cerca del lema *casar* al CTILC

The screenshot shows the 'Selecció de lemes i formes' window. It has two tabs: 'Selecció de lemes' (active) and 'Llistat de concordances'. In the 'Selecció de lemes' tab, there is a text input field for 'Lema' containing 'casar' and a dropdown menu for 'Categoria gramatical'. Below these are 'Netejar' and 'Cercar' buttons. The 'Llistat de concordances' tab is also visible, showing a table with columns 'Lema' and 'Categoria gramatical'. The first row shows 'casar' and 'VVP - Verb trans./intr. pronom.'. There is also a checkbox for 'Incloure lemes secundaris?' and an 'Eliminar-los tots' button.

I a continuació, si cliquem *Executar*, ens demanarà el nombre de resultats que volem i obtindrem el resultat de la cerca.

Figura 6. Resultat de la cerca del lema *casar* al corpus CTILC

The screenshot shows the 'Consultes al corpus' window. It has two tabs: 'Selecció de lemes i formes' and 'Llistat de concordances' (active). The 'Llistat de concordances' tab shows a 'Reconstrucció de context' window with a list of search results. Each result consists of a snippet of text with the lemma 'casar' highlighted in blue. The results are as follows:

Snippet	Lemma	Context
té la forma exacta d'una ampolla de Benedictine. Pomaré V, el	casaren,	per raons polítiques, amb la princesa Joana Marau Taaroa.
més vell que no la seva muller. Havia estat a Cuba, on s'havia	casat.	No deia mai cap paraula: pot dir-se mai. Però se sabia que tenia
teniu totes ben explicades les raons serioses que va tenir per	casar-	la amb l'hereu de can Turell de la Serra. Certament s'havien
diada de sant Valentí. Hom creu que és aquest el dia en què es	casen	els ocells. A Itàlia aquests casaments s'anomenen /valneti/;
carrera... No sentir més que em parla del sacrifici d'haver-se	casat	amb mi. Sacrifici!... no el sabrà mai aquest... Emprendrem una
gras i de més durada; Crescència, la petita de la colla, es	casà	amb un tal Joan Gulam de Verdú, i no puc dir si el va fer feliç
preguntà d'improvis, sense ni donar-se compte: --Y bé i quan te	casas,	Arnau?-- L'Arnau tingué com un surt; donà llambregada a la dona,
la segona part de la predicció de la Pitia: Edip es	casarà	amb la seva mare! Una volta completat aquest segon crim condemnat
a casar i els sacerdots que han reprès la seva llibertat de	casar-	se sense passar per l'humiliant procediment de la "reducció a
precària situació, si no se li hagués presentat l'avinentesa de	casar-	se amb la vidua del seu ex-patró. Convertit ja en mercader

És interessant destacar que aquest corpus també ens permet filtrar les nostres cerques per metadades: autor, obra, gènere literari i cronologia. Per poder-les filtrar, no executeu la cerca, continueu clicant *següent* i us permetrà seleccionar més paràmetres de la cerca.

## 2.2. AnCora

Ja hem vist anteriorment les característiques i l'estructura interna del corpus AnCora: <http://clic.ub.edu/ancora/>. Vegem-ne l'interfície:

Figura 7. Interfície de consulta del corpus AnCorà

**Cerques**

Selecciona els corpora:  
 selecciona / deselecciona tots els corpus  
 Veure resultats en grups de 5

**Corpus**

CESS\_EU  
 AnCorà\_CA  
 AnCorà\_ES

Cerca totes les frases que continguin la paraula:  Cercal

Cerca totes les frases que continguin el lema:  Cercal

Quins --Sintagma -- contenen --Sintagma -- Cercal

Quins --Sintagma -- fan funció de --Funció -- Cercal

Quins sintagmes fan funció de --Funció -- Cercal

Quines frases contenen --Funció -- i (opcional) --Funció -- Cercal

Buscar estructura sintàctico-semàntica del verb  Cercal

Quins papers temàtics té la funció sintàctica --Funció -- Cercal

Quines funcions té el paper temàtic --Paper temàtic -- Cercal

En aquesta interfície podem buscar per tots els atributs que ofereix el corpus, però no podem combinar-los entre ells; per exemple, no podem buscar el lema *casar* dins d'un SN. Només podem buscar el lema fent una cerca per lema i els sintagmes nominals en una altra cerca.

Destacarem d'aquesta interfície, la informació de sortida que ens dóna. Per exemple, si busquem el verb *casar* com a lema, obtenim aquesta sortida:

Figura 8. Resultat de la cerca del verb *casar* del corpus AnCorà

<b>Cerques</b>	Cerca totes les frases que continguin el lema: 'casar'
<b>Corpora</b>	AnCorà_CA
<b>Matches</b>	S'ha trobat 12 vegades (llistant-ne 12)
<b>Info</b>	Mostrant resultats 0 .. 4.
<input type="button" value="&gt;"/> <input type="button" value="&gt;&gt;"/> <input type="button" value="10..11"/>	
<b>Accions</b>	<b>Arbre</b>
<input type="button" value="Veure arbre complet (gràfic)"/> <input type="button" value="Veure arbre complet (jeràrquic)"/> <input type="button" value="Veure frase completa (text)"/>	" Em vaig casar el 1992 per legalitzar la meva situació perquè si no no hi havia manera de treballar aquí ", confessa.
AnCorà_CA 136_19991201.tbf <a href="#">top</a> <input type="button" value="Veure arbre complet (gràfic)"/> <input type="button" value="Veure arbre complet (jeràrquic)"/> <input type="button" value="Veure frase completa (text)"/>	L'únic nen que sembla estar en camí és el de Lydia Bosch, i això que des que mig va abandonar la seva feina i es va casar amb un arquitecte que la porta tots els dies a missa, l'en un altre temps eixerida actriu travessa una època mística que feia pensar que vivia un matrimoni cast a imatge del que recentment ha beatificat el Papa.
AnCorà_CA 148_20011102.tbf <a href="#">top</a> <input type="button" value="Veure arbre complet (gràfic)"/>	

Tenim com a sortida les frases del corpus que contenen el verb *casar* però a més a més ens permet veure l'arbre de l'anàlisi sintàctica gràficament o jeràrquicament.

Una altra informació interessant que ens ofereix aquest corpus és la informació de subcategorització verbal. Ens permet buscar la informació sintacticosemàntica d'un verb. Només cal que introduïm el verb que vulguem a la casella corresponent i premem *Cerca*. Si busquem, de nou, el verb *casar*, obtenim

l'estructura sintacticosemàntica d'aquest verb en totes les frases del corpus. Des d'aquesta pantalla de resultats, també podem accedir als arbres sintàctics i a la frase completa del text.

Figura 9. Marcs de subcategorització verbal del corpus AnCora

Accions	Arbre
<input type="button" value="Veure arbre complet (gràfic)"/> <input type="button" value="Veure arbre complet (jeràrquic)"/> <input type="button" value="Veure frase completa (text)"/>	<pre> <b>su</b>j  <b>casar</b>  <b>cc</b>  <b>cc</b>  <b>cc</b> <b>sn</b>  vaig casar  <b>sn</b>  <b>sp</b>  <b>S</b> <b>arg0</b>  a2  <b>argM</b>  <b>argM</b>  <b>argM</b> <b>agt</b>  tmp  fin  cau           </pre>
AnCora_CA 136_19991201.tbf <a href="#">top</a>	
<input type="button" value="Veure arbre complet (gràfic)"/> <input type="button" value="Veure arbre complet (jeràrquic)"/> <input type="button" value="Veure frase completa (text)"/>	<pre> <b>su</b>j  <b>casar</b>  <b>creg</b> <b>sn</b>  va casar  <b>sp</b> <b>arg0</b>  a2  <b>arg1</b> <b>agt</b>  ø           </pre>
AnCora_CA 148_20011102.tbf <a href="#">top</a>	
<input type="button" value="Veure arbre complet (gràfic)"/> <input type="button" value="Veure arbre complet (jeràrquic)"/> <input type="button" value="Veure frase completa (text)"/>	<pre> <b>creg</b> <b>su</b>j  <b>casar</b>  <b>cc</b>  <b>cc</b> <b>sp</b>  <b>sn</b>  va casar  <b>sn</b>  <b>sp</b> <b>arg1</b> <b>arg0</b>  a2  <b>argM</b>  <b>argM</b> <b>ø</b>  <b>agt</b>  tmp  adv           </pre>
AnCora_CA 15_19990301.tbf <a href="#">top</a>	

### 2.3. CucWeb

El CucWeb (<http://ramsesii.upf.es/cucweb/>) és el Corpus d'Ús de Català a la Web desenvolupat per la Universitat Pompeu Fabra (Càtedra Telefònica de Producció Multimèdia i Grup de Lingüística Computacional). Amb 225 milions de paraules, és el corpus del català més gran que existeix actualment i per la seva mida podem dir que és representatiu de la llengua catalana. Per tant, és molt útil per a estudis lingüístics i sociolingüístics sobre el català i el seu ús a Internet.

#### CucWeb

El CucWeb s'ha integrat dins d'IAC (interfície d'accés a corpus). IAC és una eina desenvolupada per Barcelona Media - Centre d'Innovació i la Universitat Pompeu Fabra que permet crear de manera fàcil i dinàmica interfícies de consulta de corpus. Durant la redacció del mòdul, s'estava fent aquesta integració, és possible que encara no s'hagi acabat i veieu una interfície diferent. Igualment, podreu fer les cerques que esmentem al mòdul.

Les interfícies que hem vist fins ara, la del CTiLC i la de l'AnCora només permeten fer cerques KWOC<sup>3</sup>, cerques sense tenir en compte el context en què apareix la paraula. En canvi, la del CucWeb permet fer cerques KWIC<sup>4</sup> en la cerca avançada i l'estadística.

Centrem-nos de primer en la cerca avançada. Si observem els atributs pels quals es pot filtrar la cerca, observem que podem buscar per forma, lema, etiqueta morfològica (amb els subtrets corresponents: gènere, nombre, mode, etc.) i funció sintàctica.

Vegem-ne un exemple: imaginem que estem fent un estudi sobre les preposicions amb què pot aparèixer el verb *pensar*: hauríem de buscar *pensar* com a lema + preposició com a categoria.

#### Vegeu també

Per a refrescar el concepte de *representativitat*, consulteu el subapartat 1.4 del mòdul "El processament del Corpus I: la lingüística empírica".

<sup>(3)</sup>De l'anglès *key words out of context*.

<sup>(4)</sup>Cerques en context.

Figura 10. Consulta *pensar* + preposició en la cerca avançada del CucWeb

The screenshot shows the CucWeb search interface. At the top, there are tabs for 'Cerca simple', 'Cerca avançada', 'Estadístiques', 'Configuració', and 'Definició corpus'. The 'Cerca avançada' tab is selected. Below the tabs, there is a dropdown menu for 'Selecciona un corpus:' set to 'cucweb'. There are also buttons for 'Metadades' and 'Opcions presentació'. The main search area is divided into two conditions, 'Condicció 1' and 'Condicció 2'. In 'Condicció 1', the word 'pensar' is entered in the 'Lema' field, and the 'Categoria' is set to 'prep'. In 'Condicció 2', the 'Categoria' is also set to 'prep'. There are checkboxes for 'Neg Amb' and 'Opcionalitat' set to 'Escull una opció'. At the bottom, there are settings for 'Context' (30 paraules), 'Resultats per pàgina' (20), and 'Nombre màxim de resultats' (100). A 'Cerca' button is at the bottom right.

Cada *Condicció* representa una paraula; per a la nostra cerca, necessitem el verb *pensar* com a lema i la categoria *preposició*. El resultat són exemples del corpus que coincideixen amb la nostra cerca:

Figura 11. Resultats de la cerca *pensar* + preposició

The screenshot shows the search results page. At the top, there is a tab labeled 'Resultats'. Below it, the text 'Resultats de la cerca:' is followed by the search query '[Lema = "pensar"] [pos = "P"]'. On the right, there is a link for 'Context: 15 paraules - 30 paraules - 50 paraules'. The results are displayed in a table with two columns: '#', which is the result number, and 'Context:', which is the text snippet. The results are numbered 21 through 26. The text snippets contain the word 'pensar en' highlighted in orange. The snippets describe various contexts related to philosophy, economics, and information technology.

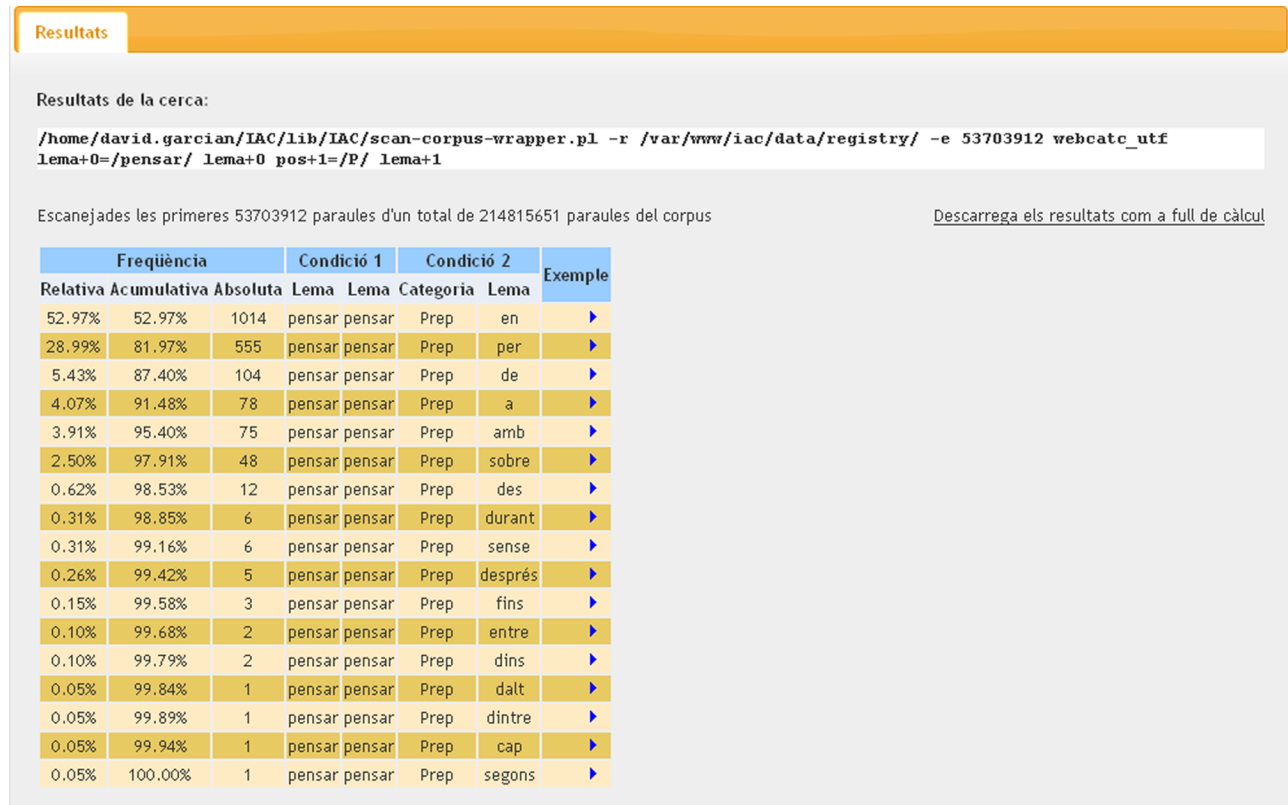
#	Context:
21	, barri nou, estadi, mercat., l'autor exposa temes, presenta filòsofs i obre debats i competicions. Lúdica i interactivament, se'ns porta a <b>pensar en</b> qüestions d'ètica i filosofia. LL. Vallmajó Data: 10 7 2000 Filosofia i Història de la Ciència i de la Tècnica
22	recordar que el futur econòmic ve marcat per la combinació intel·ligent i orientada al mercat d'idees i serveis. Dues són les raons principals que permeten <b>pensar en</b> una futura interrelació estreta entre les indústries turística i informàtica. D'una banda, el fet que el turisme sigui la primera indústria mundial el converteix en principal client
23	els factors tradicionals preu, diferenciació i nínxol un paper complementari dels nous factors qualitat, innovació, personalització, informacionalització, etc, fan necessari <b>pensar en</b> les TI com a generadores de noves oportunitats de negoci. Es tracta, doncs, d'un canvi d'una visió merament tàctica a una d'estratègica
24	del contingut al continent És molt possible que els centres d'informació empresarial siguin els primers centres d'informació que reorientin la seva forma de <b>pensar en</b> els seus recursos d'informació, migrant des del tradicional enfoc en el "llibre" (el producte d'informació per excel·lència) cap a un
25	Catalunya, de com han anat distribuint informació via disc o web. Són els primers passos, però n'han de venir més. Caldrà que els centres <b>penstin de</b> manera molt original com es pot donar millor servei. Per exemple, algun dels grans centres d'informació per empreses a Catalunya, com
26	el calendari de desplegament dels tubs. Perquè va essent clar que els països que en treuran profit de les autopistes de la informació són aquelles que <b>pensar en</b> termes d'infraestructura més que en els d'infraestructura. Aquells que inverteixen en facilitar l'aparició de fàbriques de transformació massiva dels suports paper a

D'altra banda, també ens pot interessar obtenir resultats estadístics per saber amb quina freqüència apareixen els diversos tipus de preposicions. CucWeb també ens ofereix el tipus de cerca estadística. Cal que fem la cerca següent:

Lema *pensar* [freqüència lema] + *preposició* [freqüència lema]

El resultat és una taula de freqüències amb el verb *pensar* i la freqüència amb què apareixen les diverses preposicions.



Figura 12. Resultats de *pensar* + preposició en la cerca estadística

### 3. Exploració d'un corpus

Fins ara, hem estat veient els diversos formats de corpus que hi ha i les interfícies amb què es poden consultar. Sens dubte, tots aquests corpus i interfícies són de gran utilitat per als investigadors però sovint tenen els seus propis corpus i no tenen una interfície per fer consultes fàcilment. Afortunadament, hi ha eines que ens poden ajudar a obtenir dades interessants del nostre corpus, sense necessitat de dissenyar una interfície. En aquest apartat farem ús d'una d'aquestes eines: l'AntConc, un programa d'exploració de corpus. Aprendre-m a:

- Comptar el nombre d'elements d'un corpus.
- Extreure el context en què apareix una paraula en el corpus.
- Buscar una paraula en el corpus i calcular-ne la freqüència.
- Fer grups de paraules (clústers) de diversos elements (*n*-grames) més freqüents del corpus.
- Fer cerques de grups de paraules més freqüents que apareixen en el corpus.

#### 3.1. Instal·lació de programari

L'AntConc és un programa d'exploració de corpus per a Windows (98/Me/2000/NT/XP/7), Macintosh OS X i Linux. Per descarregar-lo aneu directament a la pàgina de l'eina (<http://www.antlab.sci.waseda.ac.jp/software.html>) o bé copieu-lo del CD de l'assignatura.

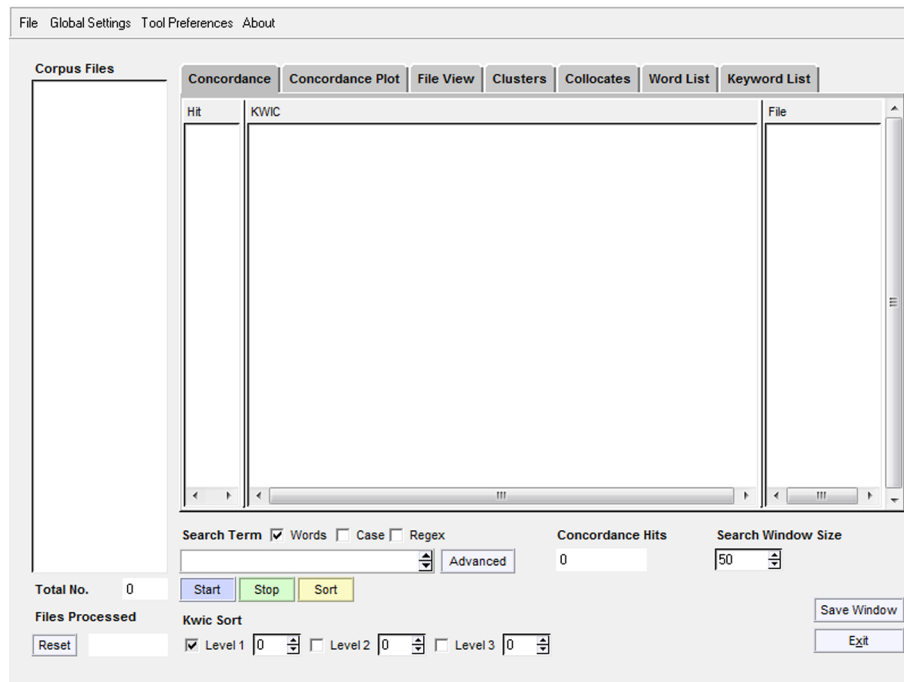
Per instal·lar-lo, feu doble clic sobre l'arxiu que heu descarregat i s'obrirà l'AntConc.

#### 3.2. Descripció de la interfície de cerca

La interfície de cerca de l'AntConc es basa en tres parts:

- A l'esquerra apareixen els noms dels arxius del nostre corpus (Corpus Files).
- A la part central superior, el resultat de la cerca i les pestanyes per poder fer diversos tipus de cerca.
- A la part central inferior, les condicions de cerca (p. ex. paraula que volem buscar, ordenació dels resultats, etc.).

Figura 13. Mostra de l'AntConc



Obriu el corpus que us ha proporcionat el professor, anant al menú *File* (Arxiu) > *Open Directory* (Obrir directori) > Seleccioneu el directori que heu descarregat i premeu *Open* (Obrir).

Veureu a la part esquerra que s'ha carregat correctament el vostre corpus si apareixen els noms dels arxius i a *Total No.* (nombre total d'arxius) apareix **Total No. 183**

A partir d'ara, ja podem començar a treballar.

### 3.3. Llistes de paraules (*word list*)

Inicialment, començarem esbrinant dades generals del nostre corpus. Aneu a la pestanya *Word List* (Llista de paraules) i feu clic al botó *Start* (sobretot no escriviu res al quadre de text). El resultat ens dóna molta informació sobre el corpus:

Figura 14. Llista de paraules més freqüents del corpus

Concordance				Concordance Plot				File View				Clusters				Collocates				Word List				Keyword List			
Hits		Total No. of Word Types: 12456												Total No. of Word Tokens: 69759													
Rank	Freq	Word	Lemma Word Form(s)																								
1	2730	de																									
2	2178	que																									
3	2124	la																									
4	1671	a																									
5	1659	i																									
6	1551	el																									
7	1004	va																									
8	999	l																									
9	992	en																									
10	935	un																									
11	860	d																									
12	794	per																									
13	785	una																									
14	670	amb																									
15	659	no																									
16	656	les																									
17	639	del																									

**Nota**

**Molt important:** comproveu que teniu seleccionada la codificació adequada (UTF-8). Per fer-ho aneu al menú *Global Settings*, opció *Language Encodings*, *Chosen Language Encoding* [*Unicode (utf-8)*]. Si hi teniu una altra codificació, cliqueu *Edit* i seleccioneu *Standard Encodings* > *Unicode (utf-8)* > *Apply*.

El resultat ens indica els *types* i els *tokens* del corpus. Quan es parla de *tokens*, es fa referència al nombre de paraules que té el corpus. En el nostre corpus tenim 69.759 tokens (*total no. of word tokens*) i, d'aquests, 12.456 són *types*, és a dir, ocurrències úniques d'un *token*.

A més, també s'ha generat una llista de paraules del corpus amb la freqüència absoluta associada. Per exemple, observem la paraula més freqüent en el corpus (*de*) amb 2.730 ocurrències. Però de vegades, la freqüència absoluta no ens dóna massa informació, ja que una paraula és més o menys freqüent depenent d'on la trobem. És a dir, no podem dir que una paraula és molt o poc freqüent en un corpus amb relació a les altres, si no tenim en compte aquesta paraula entre totes les del corpus. Caldria que en calculéssim la freqüència relativa; ho fem de la manera següent:

$$\frac{\text{nre. d'ocurrències de la paraula (freq. absoluta)}}{\text{nre. total d'elements del corpus (nre. de total de tokens)}}$$

Per exemple, per calcular la freqüència relativa del nostre corpus fariem:

$$2.730 \text{ (de)} / 69.759 = 0,039$$

La freqüència relativa del nostre corpus és de 0,039.

Vegem un altre exemple. Imaginem que tenim dos corpus: el Corpus 1 de 69.759 paraules, que és el corpus que hem estat explorant fins ara, i el Corpus 2 de 5.000.

Ens interessa saber en quin corpus és més freqüent la paraula *de*.

Primer de tot, fem un recompte de *de*:

*Corpus 1: 2.730 ocurrences de de*

*Corpus 2: 2.730 ocurrences de de*

Casualment, en ambdós corpus *de* apareix el mateix nombre de vegades, és a dir, tenen la mateixa freqüència absoluta. Vegem, doncs, que la freqüència absoluta no ens serveix com a mesura per saber en quin corpus la paraula *de* és més freqüent. Necessitem comparar-ne les freqüències relatives.

Taula 6

	<b>Corpus 1 (nre. tokens: 69.759)</b>	<b>Corpus 2 (nre. tokens: 5.000)</b>
<b>Freqüència absoluta</b>	2.730	2.730
<b>Freqüència relativa</b>	0,039 (2.730/69.759)	0,54 (2.730/5.000)

Observem que a la taula anterior, tot i que *de* té la mateixa freqüència absoluta en ambdós corpus, si en mirem la freqüència relativa, aquesta és més alta en el Corpus 2. Podem dir, doncs, que en el Corpus 2 *de* és més freqüent que en el Corpus 1.

### 3.4. Concordances

També ens pot interessar buscar concordances, és a dir, un mot en el seu context (KWIC). Per exemple, podem buscar les concordances en el nostre corpus per a la paraula *no*<sup>5</sup>.

<sup>(5)</sup>Ens pot interessar com a lingüistes tenir un recull de frases que continguin aquesta paraula per a estudiar l'ús de l'adverbi *no* en català.

Per fer-ho, aneu a la pestanya *Concordance* i introduïu al quadre de text de la cerca la paraula *no* i premeu *Start*. A continuació, ens apareix:

- A l'esquerra, els arxius que formen el nostre corpus.
- Al centre, els resultats de la cerca.
- A la dreta, el nom de l'arxiu en què apareix concretament la cerca.

Figura 15. Concordances amb la paraula *no* (sense tenir en compte el context de la dreta)

The screenshot shows the AntConc software interface. The main window displays concordance results for the search term "no". The interface includes a menu bar (File, Global Settings, Tool Preferences, About), a Corpus Files list on the left, and a main window with tabs for Concordance, Concordance Plot, File View, Clusters, Collocates, Word List, and Keyword List. The Concordance tab is active, displaying a table with columns Hit, KWIC, and File. The search term "no" is entered in the Search Term field, and the search is set to "Words" case. The total number of hits is 1686, and the search window size is 50. Below the table, there are buttons for Start, Stop, Sort, and a Kwic Sort section with checkboxes for Level 1, Level 2, and Level 3.

Ens pot interessar ordenar aquests resultats segons el context que tinguin. Ho farem indicant-ho a *Sort*. Marcarem la casella *Level 1* i, tot seguit, seleccionarem *IR* amb les fletxes –d'aquesta manera, li estem indicant que volem ordenar els resultats per la paraula que té a la dreta– i, a continuació, cliquem *Sort*. La cerca és la mateixa però en aquesta ocasió està ordenada alfabèticament per la paraula de la dreta.

Per exemple, podem observar que del *hit* 475 al 492 *no* va acompanyat d'*obstant*.

### 3.5. Concordances gràfiques (*concordance plot*)

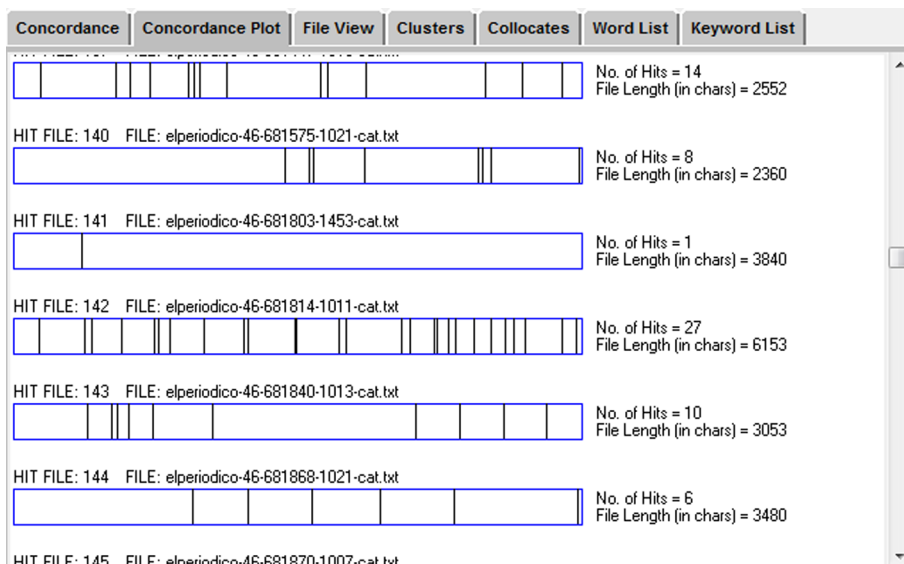
Seguint amb la cerca de la paraula *no*, l'AntConc també ens proporciona una manera gràfica de veure la cerca per arxius. Sense esborrar la cerca anterior, cliqueu a la pestanya *Concordance plot*.

És interessant veure la mateixa cerca gràficament perquè ens dona una idea clara de com es distribueix la nostra cerca en els diversos arxius.

A més, aquesta interfície també ens mostra el nombre de vegades (*no. of hits*) que apareix en cada arxiu la paraula que hem buscat i a més també ens indica la llargada de l'arxiu en nombre de caràcters (*file length [in chars]*).

Si volem veure les ocurrències concretes, cal que cliquem a sobre la barra i s'obrirà la finestra *File* (Arxiu) en què veurem l'arxiu complet amb les ocurrències marcades.

Figura 16. Concordances gràfiques de la cerca *no*

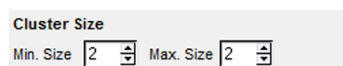


### 3.6. Agrupaments de segments (clústers i *n*-grames)

També ens pot interessar buscar agrupaments de paraules en el corpus. Aquests grups s'anomenen *clústers* i les seqüències de mots que els formen s'anomenen *n*-grames.

Per exemple, podem buscar quin és el conjunt d'*n*-grames més freqüents del corpus i que aquests estiguin agrupats pels *n*-grames de 2 elements (bigrames). Dit d'una altra manera, volem buscar grups (clústers) de 2 paraules (bigrames, *n*-grama de 2 elements). Per fer-ho cal que anem a la pestanya *Cluster* (segons com tingueu configurada l'eina, es pot dir *N-gram*) i que configurem la mida de l'*n*-grama (no us oblideu de tenir la casella *N-gram* marcada):

Figura 17. Mida dels *n*-grames



Si volguéssim buscar trigrames, canviaríem *Min. size* i *Max. size* a 3 i, si volguéssim buscar bigrames i trigrames a la vegada, canviaríem *Min. size* a 2 i *Max. size* a 3.

Segons els nostres resultats, el bigrama més freqüent és *de la* amb una freqüència absoluta de 414.

Figura 18. Resultat de la cerca de bigrames

Rank	Freq	N-gram
1	414	de la
2	275	a la
3	201	la seva
4	170	de l
5	150	a l
6	140	en el
7	128	de les
8	122	que no
9	118	que el
10	114	hi ha
11	111	es va
12	110	que va
13	108	d un
14	106	el seu
15	104	per a
16	98	en la
17	96	i la

### 3.7. Col·locacions (*collocates*)

Aquesta eina es fa servei per generar una llista de col·locacions que apareixen amb la paraula que ens interessa. Entenem per *col·locacions* aquell conjunt de paraules (de dos o més elements) que acostumen a aparèixer juntes. Per exemple, les preposicions que apareixen amb els verbs o la paraula *figa/figues* en expressions col·loquials com *pesar figues o fer figa*.

Imaginem que volem fer un estudi sobre quines preposicions apareixen amb més freqüència amb la paraula *interès*.

Per veure com es comporta aquesta paraula en el corpus, cal que anem a la pestanya *Collocates* i introduïm *interès* al quadre de text. A més, cal que definim també el context en què volem treballar. En el nostre cas, volem buscar *interès* + un element a la dreta. Caldrà que configurem la cerca de la manera següent:

Figura 19. Des de 0 a 1 element a la dreta (*right*)

Window Span  Same  
 From... 0 To... 1R

I a continuació cliquem *Start*.



Figura 20. Cerca de col·locacions amb la paraula *interès*

The screenshot shows the 'Collocates' window of a corpus analysis tool. The search term is 'interès'. The results table is as follows:

Rank	Freq	Freq(L)	Freq(R)	Collocate
1	7	0	0	interès
2	2	0	2	pels
3	1	0	1	propi
4	1	0	1	per
5	1	0	1	pel
6	1	0	1	Es
7	1	0	1	de

Additional interface details: Search Term: *interès*; Total No. of Collocate Types: 7; Total No. of Collocate Tokens: 14; Search Term: *interès*; Window Span: From 0 To 1R; Min. Collocate Frequency: 1.

A la columna *Freq* s'indica la freqüència absoluta de la col·locació. És a dir, *interès* apareix 7 vegades en el corpus, de les quals 2 vegades va seguida de *pels*.

*Freq (L)* ens indica quantes vegades apareix a l'esquerra i *Freq (R)* quantes a la dreta; en el nostre cas, *Freq(L)* sempre és 0 perquè només ens hem fixat en el context de la dreta (R) i, finalment, a *Collocate* ens diu la segona paraula de la col·locació. Observem que la preposició més freqüent és *per/pels* seguit de *de*. Fent una abstracció una mica "casolana"<sup>6</sup> dels resultats, podem dir que normalment *interès* va seguit de la preposició *per*.

<sup>(6)</sup>Aquests resultats quantitius no són representatius ja que el nostre corpus és molt petit.

## Resum

En aquest mòdul ens hem endinsat, tenint en compte que l'estudiant ja ha assolit els coneixements del mòdul "El processament de corpus I: la lingüística empírica", en el vessant més pràctic de la lingüística de corpus. No hem pretès donar una visió exhaustiva de tots els recursos i formats que hi ha en català, sinó fer unes pinzellades perquè l'estudiant, de manera autònoma, pugui continuar ampliant els seus coneixements en aquest camp.

Hem començat des dels fonaments més bàsics d'un corpus: la informació que conté (per exemple, categoria morfològica, any del document, etc.) i el format amb què es pot anotar aquesta informació, fins a arribar a l'exploració del nostre propi corpus.

A la primera part del mòdul, hem parlat dels formats que poden tenir els corpus (sense pretendre tractar de tots els formats disponibles) i de les anotacions lingüístiques i extralingüístiques que poden contenir.

A la segona part, hem après a fer cerques en interfícies de corpus ja existents en català, tots ells amb alguna peculiaritat pel que fa a la informació que contenen: p. ex. metadades, sintaxi, categories morfològiques, etc. Pel que fa a les cerques, hem vist cerques sense context (KWOC) i cerques en context (KWIC), i ja centrant-nos en els resultats, hem après a obtenir exemples de cerques, subcategoritzacions verbals, arbres sintàctics i dades quantitatives.

I finalment hem après a explotar el nostre corpus amb el programa AntConc, fent cerques KWIC i KWOC, i extraient resultats qualitius i quantitius.

## Activitats

### Sobre l'apartat "Informació i formats"

1. Per conèixer una mica més els corpus signats, consulteu els projectes següents:

- Corpus BSL: <http://www.bsllproject.org/>
- Corpus NGT: <http://www.ru.nl/corpusngt/>
- Corpus Auslan: <http://www.auslan.org.au/about/corpus/>
- Signs of Ireland Corpus: [http://www.tcd.ie/sllscs/cds/research/featuredresearch\\_signcorpus.php](http://www.tcd.ie/sllscs/cds/research/featuredresearch_signcorpus.php)
- American Sign Language Linguistic Research Project: <http://www.bu.edu/asllrp/>

2. I si us interessen els corpus orals, consulteu el corpus RETOC (<http://retoc.iula.upf.edu/html/>), desenvolupat per l'Institut Universitari de Lingüística Aplicada.

També us pot interessar l'article ([http://latel.upf.edu/terminotica/membres/DE\\_YZA/PUBLI/retoc.pdf](http://latel.upf.edu/terminotica/membres/DE_YZA/PUBLI/retoc.pdf)) sobre el corpus: Ll. de Yzaguirre; A. J. Farriols; J. Martí (2004). En aquest article també trobareu detalls interessants sobre l'eina Praat i el procés d' anotació de corpus orals.

### Sobre l'apartat "Interfície de consulta de Corpus"

3. Quant a les interfícies de consulta de corpus, n'hem vist algunes. Torneu a les que heu vist en el mòdul 1:

- Busqueu informació sintacticosemàntica sobre el verb *interessar* (corpus AnCora).
- Busqueu exemples del verb *jugar* en les obres de Josep Carner (corpus CTILC).

4. En aquest mòdul no hem vist totes les interfícies de corpus en català. Aquí en teniu algunes més per visitar:

- bwanaNet: <http://bwananet.iula.upf.edu/>
- BancTrad: <http://mutis2.upf.es/cgi-bin/bt/search-form.pl>

### Sobre l'apartat "Exploració d'un Corpus"

5. Agafeu deu documents de Word que tingueu a l'ordinador i poseu-los en un directori. Obriu aquest directori amb l'AntConc i intenteu esbrinar:

- Quina és la freqüència absoluta de la paraula *la*? I la relativa?
- Quin és el trigràma més freqüent del corpus?
- Quin arxiu o arxius contenen més aparicions de *la*?

6. Segons els vostres propis interessos lingüístics, en què creieu que us podria ajudar tenir un corpus. Quins tipus de cerques hauríeu de fer per poder-ne obtenir resultats interessants?

## Glossari

**col·locació** *f* Relació lèxica entre dos o més mots que apareixen junts amb certa freqüència.

**key word in context** *m* Cerca d'una seqüència de paraules que apareixen en un context determinat.  
sigla **KWIC**

**key word out of context** *m* Cerca en un corpus sense tenir en compte el context.  
sigla **KWOC**

**KWIC** *m* Vegeu **key word in context**.

**KWOC** *m* Vegeu **key word out of context**.

**lema** *m* Representació formal del conjunt de formes flexionades d'un mot (per exemple, el lema de la forma *cantava* és *cantar*).

**n-grama** *m* Qualsevol seqüència de mots que apareix en un corpus. Si la seqüència és formada per un mot s'anomena *unigrama*; si és formada per dos mots, *bigrama*; si és formada per tres mots, *trigrama*.

**ocurrència** *f* Nombre de vegades que apareix un element en un text.

**token** *m* Aparició concreta d'una paraula en un text.

**type** *m* Unitat abstracta que recull totes les aparicions d'una paraula en un text.

## Bibliografia

En aquesta bibliografia, trobareu una introducció breu a la lingüística de corpus en els articles de Rojo (2002) i Saurí (2004). Les altres referències són una ampliació dels coneixements sobre els corpus existents en català que hem vist en el mòdul. Si necessiteu més bibliografia sobre lingüística de corpus, consulteu la proposada per J. Rafel i J. Soler en el mòdul "El processament de corpus I: la lingüística empírica".

**Badia, T.; Boleda, G.** (2008). *CUCWEB: un corpus de la llengua catalana construït a partir de la web* (vol. 30) [en línia]. Editorial IEC.

**Garmendia, M.; Badia, A.; Colominas, C.; Brumme, J.; Boleda, G.; Quixal, M.** (2002). "Banc Trad: un banco de corpus anotados con interfaz web". *Procesamiento del lenguaje natural* (núm. 29, pàg. 293-294) [en línia].

**Martí, M. A.; Taulé, M.; Márquez, Ll.; Bertran, M.** (2007). "Ancora: A Multilingual and Multilevel Annotated Corpus" [en línia].

**Rojo, G.** (2002, octubre). "Sobre la lingüística basada en el anàlisi de corpus". A: *Jornadas sobre corpus lingüísticos* [en línia]. Ponència plenària. Sant Sebastià: Uzei.

**Saurí Colomer, R.** (2004). "Un corpus para el asturiano: Las tecnologías lingüísticas en la consolidación de las lenguas minorizadas". *Revista de Filología Asturiana* (vol. 3/4, anys 2003/2004, pàg. 135-174) [en línia].

**Soler, J.** (2002). "El Corpus Textual Informatitzat de la Llengua Catalana" [en línia]. Hizkuntza-corpusak. Oraina eta geroa (2002-10-24/25).

**Yzaguirre, Ll. de; Farriols, A. J; Martí, J.** (2004). "El corpus RETOC: Un corpus oral per a la recerca i la docència". A: S. Martí; M. Cabré; F. Feliu; N. Iglesias; D. Prats (eds.). *Actes del 13è Col·loqui de l'AILLC* (Girona 2003) [en línia]. PAM

