CuCWeb: un corpus del català construït a partir de la web

G. Boleda[i], S. Bott[i], B. Poblete[ii], C. Castillo[ii], M.E. Fuenmayor[ii], T. Badia[i], V. López[ii]
[i]GLiCom, Departament de Traducció i Filologia
[ii]Departament de Tecnologia i Cátedra Telefónica de Producción Multimedia
Universitat Pompeu Fabra

Paraules clau: web, corpus, lingüística, sociolingüística, català

Keywords: web, corpus, linguistics, sociolinguistics, Catalan

Resum: Aquest article presenta el Corpus d'Ús del Català a la Web (CuCWeb), un corpus de 208 milions de paraules (125.000 documents) compilat automàticament a partir de la Web. Aquest corpus ha estat processat automàticament per tal de proporcionar informació lingüística addicional a la mera forma de les paraules, i s'ha habilitat una interfície de cerca molt flexible que permet cercar exemples per a determinades construccions o paraules i també extreure freqüències d'ús. Creiem que aquest recurs és molt útil sobretot per a) conèixer la llengua catalana (estudis lingüístics) i b) conèixer l'ús que es fa de la Web en català (estudis sociolingüístics).

Abstract: This paper presents the *Corpus d'Ús del Català a la Web* (CuCWeb), a 208 million word (125,000 documents) corpus automatically compiled from the Web. This corpus has been automatically processed so that additional linguistic information is available (apart from the word forms). A very flexible search interface has been implemented, which allows for different kinds of searches for constructions and words, as well as frequency information. We believe this resource is mostly useful to a) get to know the Catalan language (linguistic research) and b) get to know the use that is made of the Web in Catalan (sociolinguistic research).

Motivació del projecte

Els corpus lingüístics

En l'estudi del llenguatge, els corpus lingüístics són col·leccions de textos preparats per a l'anàlisi lingüística. Tot i que pròpiament tota col·lecció de textos adequada per a l'anàlisi lingüística es considera un corpus (independentment del suport en què estigui guardada), actualment

tots els corpus lingüístics realment útils estan en suport electrònic, ja que així poden ser consultats de manera molt més efectiva i eficient.

La compilació, organització i explotació de corpus lingüístics constitueix una de les activitats principals dels lingüistes i dels estudiosos de la llengua, amb l'objectiu de tenir-ne exemples reals d'ús. Els corpus permeten aproximar-se a la llengua d'una manera no normativa sinó descriptiva. Justament la possibilitat d'observar la llengua real i de descriure-la adequadament és l'avantatge principal que ofereixen els corpus lingüístics. Així, els corpus permeten de respondre preguntes com:

- com s'usen les paraules (o com no s'usen)?
- quines expressions s'usen (o quines no s'usen)?
- quines estructures es fan servir en un context determinat?

D'aquesta manera es pot descobrir que els catalans usem sovint el verb *berenar* com a transitiu, encara que el diccionari normatiu ens indiqui que és un verb intransitiu; o que hi ha variacions enormes en l'ús de la veu passiva amb verbs transitius, de manera que no tots els verbs ni tots els temps verbals tenen la mateixa freqüència d'ús de la veu passiva.

Els estudis lingüístics basats en corpus descobreixen els usos presents en el corpus sobre el qual s'han basat. D'aquí l'enorme importància de conèixer la manera com s'ha compilat un corpus. Així un corpus constituït per textos literaris dels anys 50 del segle passat permetrà descobrir i estudiar una llengua que és diferent en molts aspectes de la que podem descobrir a partir d'un corpus constituït per textos científics dels primers anys d'aquest segle. I no ens referim només a aspectes relacionats amb la temàtica o la ideologia subjacent en els textos, sinó també a l'ús lingüístic pròpiament dit: per exemple, a la freqüència amb què hi apareixen els pronoms febles *hi* i *en*, que han anat desapareixent de molts registres de la llengua; o a l'ús dels adjectius, que és clarament diferent en els contextos descriptius o en els argumentatius. Com que l'ús de la llengua no és

uniforme, sinó variat i divers, el tipus de textos recollits en un corpus condiciona les conclusions que es puguin treure de la seva observació. Així doncs, la representativitat dels corpus, juntament amb la quantitat de text i la qualitat del text que recullen, són elements que determinen el seu valor.

El valor dels corpus lingüístics

Entre els principals factors que incideixen en el valor final d'un corpus destaquen la seva mida, l'origen dels textos que el composen i el marcatge lingüístic que incorpora. La grandària del corpus afecta tant el tipus d'informacions que se'n poden extreure com la seva representativitat.

En general, per a cada qüestió hi ha una grandària adequada per a observar-la. Per exemple, no necessitem un corpus de textos catalans gaire gran per adonar-nos que els articles en català precedeixen el nom; com que es tracta d'un fenomen lingüístic molt freqüent, ja és evident en textos molt curts. En canvi, per poder distingir entre l'ús de *gens* i de *res*, ens caldrà un corpus més gran, ja que aquest parell de paraules poden no aparèixer en textos curts, i per poder observar el comportament general en el corpus necessitarem un nombre significatiu d'exemples, per evitar que inadvertidament convertim en normal un tractament excepcional. Així mateix, quan busquem correlacions estadístiques en l'ús d'una paraula juntament amb unes altres (és a dir, quan volem veure amb quines paraules apareix una paraula determinada) necessitem corpus molt grans, perquè altrament no tenim exemples suficients de l'ús de la paraula en qüestió per ser significatius estadísticament. Per exemple, si estem interessats en conèixer amb quines preposicions tendeix a aparèixer un verb determinat, com *interessar* (que pot anar amb *en*, *per*, *a*...), necessitarem molts exemples d'aquest verb perquè puguem afirmar que *interessar* va més amb *en* que amb *per*.

Per altra banda, un corpus relativament petit conté menys mots i, per tant, exemplificarà l'ús de menys mots de la llengua, cosa que en condiciona la representativitat. En general, com més mots hi ha en un corpus, més mots hi tenen una freqüència significativa i, per tant, se'n poden extreure més dades sobre el seu comportament lingüístic.

L'altre gran factor que determina la representativitat dels corpus és l'origen dels textos que el componen. Si aquests tenen unes característiques comunes específiques, el corpus representarà només el seu tipus de llenguatge; per exemple, d'un corpus format per converses telefòniques no se'n podran extreure conclusions sobre la llengua científica. En les dues darreres dècades del segle XX, els recopiladors de corpus tenien molta cura en incorporar-hi textos suficientment rics i variats amb proporcions predeterminades per tal de representar la llengua general en qüestió: aquest ha estat el principi fonamental que ha guiat la creació dels grans corpus recollits en aquells anys (entre els quals el Corpus Textual Informatitzat de la Llengua Catalana, de l'Institut d'Estudis Catalans; v. RAFEL, 1994). Això no obstant, cal tenir en compte que el concepte de representativitat és discutible, difícil d'objectivar, i que els corpus són representatiu d'un ventall sempre restringit de variants lingüístiques (registres, gèneres, temàtiques, dialectes), en funció del criteri de qui el construeixi.

En aquest sentit, també s'han anat constituint corpus especialitzats amb l'objectiu d'estudiar a fons aspectes que no són prou representats en aquests corpus generals; així, per al català, actualment podem trobar corpus de llenguatges científics, corpus de llengua oral, corpus periodístics, etc.

La tercera característica que determina el valor d'un corpus és la informació addicional que té, cosa que determina la informació que se'n pot extreure. En un corpus hi pot haver dos tipus d'informació addicional: la general, que codifica informació sobre el text en qüestió (data, autor, origen...), i la lingüística, que caracteritza amb més detall cada una de les paraules del corpus (per exemple, indicant si *roda* és un verb o un nom en un exemple particular, o si *cap* és un verb, un nom, un determinant o una preposició). Com més extensa i acurada sigui la informació addicional que complementa el corpus, més exacta i rica serà la informació que se'n podrà obtenir.

El marcatge dels corpus lingüístics

Marcar un corpus consisteix precisament en incorporar-hi aquesta informació addicional.

L'operació de marcar un corpus actualment se sol fer de forma automàtica, és a dir, utilitzant programes que efectuen aquest tipus de marcatge. El marcatge lingüístic d'un corpus consisteix en associar amb cada paraula informació sobre les seves propietats lingüístiques: de quina paraula es tracta, propietats morfològiques, sintàctiques... Vegem-ne un exemple, marcat amb la informació que conté el corpus que presentem en aquest article:

La fi de la guerra va suposar la fi de la lluita contra el règim

Forma	**Lema**	**Categoria**	**Codi de propietats morfològiques**	**Funció sintàctica**
La	el	Det	AFS	Determinant de nom
fi	fi	Nom	N5-6S	Complement directe o subjecte
de	de	Prep	P	Complement d'un nom a l'esquerra
la	el	Det	AFS	Determinant de nom
guerra	guerra	Nom	N5-FS	Terme de preposició
va	anar	Verb	VDR3Sa-	Verb auxiliar
suposar	suposar	Verb	VI----	Verb principal
la	el	Det	AFS	Determinant de nom
fi	fi	Nom	N5-6S	Complement directe o subjecte
de	de	Prep	P	Complement d'un nom a l'esquerra
la	el	Det	AFS	Determinant de nom
lluita	lluita	Nom	N5-FS	Terme de preposició
contra	contra	Prep	P	Complement d'un nom o adverbial
el	el	Det	AMS	Determinant de nom
règim	règim	Nom	N5-MS	Terme de preposició

Fig. 10. Exemple d'oració amb marcatge lingüístic.

En l'exemple apareix la informació en columnes:

- en la primera columna, hi ha la forma, és a dir, el mot tal com apareix en el corpus,
- en la segona, hi ha el seu lema (o forma de referència de la paraula),
- en la tercera, s'hi indica la categoria morfològica de la paraula,
- en la quarta, hi apareix un codi que reprodueix la categoria morfològica de la paraula, juntament amb les seves propietats morfològiques (gènere, nombre, persona, temps, mode...): AFS (article femení singular), N5-6S (nom comú masc/fem singular), P (preposició), N5-FS (nom comú femení singular), VDR3Sa- (verb indicatiu present 3a

persona singular), VI---- (verb infinitiu), AMS (article masculí singular), N5-MS (nom comú masculí singular),

- en la darrera, s'hi codifica les funcions sintàctiques que compleixen les paraules: determinant d'un nom a la dreta, objecte directe o subjecte, complement d'un nom a l'esquerra, etc.; en l'exemple, es veu com de vegades el programa de marcatge no és capaç de resoldre adequadament entre més d'una possibilitat (*fi* podria ser complement directe o subjecte en els dos casos, segons el programa).

Com veurem, una codificació com aquesta permet fer cerques detallades en el corpus; per exemple, un nom en funció de subjecte, precedit per l'article *el*.

El corpus de la web catalana

A part dels corpus més o menys petits que els grups de recerca d'institucions públiques o empreses privades han anat recollint per a les seves investigacions i activitats, el corpus català més representatiu de la llengua general és el CTILC, Corpus Textual Informatitzat de la Llengua Catalana (creat per l'Institut d'Estudis Catalans). Aquest corpus té uns 52 milions de paraules i ha estat recollit amb la finalitat de conèixer el català real (des de les darreries del segle XIX fins als anys 80 el segle XX) i poder fer un diccionari que reflecteixi l'ús de la llengua estàndard, tal com apareix en els textos literaris, científics, periodístics...

El corpus que avui presentem té unes altres característiques i respon a uns objectius molt diferents. Es tracta d'un corpus extret automàticament de la web i processat de forma automàtica també. Això ha permès constituir-lo en terminis de temps realment molt curts: l'extracció de la web s'ha fet al mes d'abril del 2004 i el corpus i el seu sistema de consulta és operatiu des de finals de setembre del 2004. És clar, per tant, que no serà un corpus tan ben acabat com el CTILC (ja que tots els processos s'han dut a terme automàticament i no s'ha pogut fer una revisió manual de tot el corpus i el seu marcatge), però serà molt més gran (en aquest moment té més de 200 milions de paraules), cosa que permet fer-hi estudis que no són possibles amb corpus menors. No és un corpus

representatiu de la llengua estàndard, com el CTILC, sinó un corpus representatiu de la llengua que utilitza la gent quan construeix pàgines Web en català, i dels documents que es troben a la Web. És per tant un corpus que mostra la llengua viva, en el seu ús immediat, en un context determinat; clarament, això és la seva limitació, però també el seu valor, des del punt de vista sociolingüístic.

En recopilar el Corpus d'Ús del Català a la Web, el CuCWeb, i posar-lo a disposició dels estudiosos i la societat catalana en general, la nostra intenció doncs ha estat la d'oferir un corpus amb unes característiques específiques, que el diferencien de tots els corpus catalans existents fins avui:

1. un corpus ampli de la llengua catalana, el més ampli possible actualment, obtingut amb recursos relativament poc costosos o ja disponibles
2. un corpus marcat amb informació lingüística bàsica (consistent en assignar de manera automàtica a cada paraula la informació sobre la seva forma, el seu lema, les seves propietats morfològiques i la seva funció sintàctica)
3. un corpus de llengua real, representatiu de la llengua habitual, ja que la web permet una agilitat i immediatesa en la producció de textos escrits que no són possibles en cap altre mitjà
4. un corpus fàcilment assequible a totes les persones interessades en aspectes concrets de la llengua:
   - lingüistes
   - lexicògrafs
   - professors i mestres
   - sociòlegs y sociolingüistes

Creació

La procés de creació del corpus s'ha dividit en tres etapes: el procés de recollida del domini *.es* (ap. 2.1), la classificació per idiomes (ap. 2.2), l'etiquetatge lingüístic i el processament amb un motor de cerca de corpus (ap. 2.3).

Procés de recollida del domini *.es*

El corpus s'ha obtingut a partir de les pàgines de la *World Wide Web* que són en servidors el nom dels quals (domini) acaba amb la terminació assignada a Espanya (*.es*). Es tracta d'una aproximació molt simple al que podria denominar-se "la web espanyola", un concepte que, en rigor, no existeix. Una aproximació alternativa a la web espanyola seria considerar les pàgines allotjades a una màquina en territori espanyol. Però això inclou també nombroses pàgines d'empreses o institucions angleses que compleixen aquesta condició. L'idioma tampoc no és un element identificatiu, ja que l'espanyol no es pot relacionar unívocament amb Espanya.

En escollir les pàgines del domini de primer nivell *.es*, seleccionem les que els seus autors associen voluntàriament amb una marca o empresa espanyola. Un estudi preliminar sobre el contingut de la web allotjada en màquines del territori espanyol indica que el domini *.es* representa aproximadament un 40% del total (v. POBLETE *et al.*, en preparació).

Es van recollir 7.752.967 pàgines del domini *.es* durant la primera quinzena d'abril del 2004, mitjançant els robots cedits per l'empresa especialitzada en motors de cerca Akwan.[1] Aquestes eines es van utilitzar per extreure el text dels arxius HTML i dels que tenen un format estàndard de text (TXT, DOC, RTF, etc.). La col·lecció inclou també 577.529 fitxers en format PDF (Adobe).

Els documents extrets ocupaven inicialment 247 Gigabytes, que van quedar reduïts a 36,2 Gigabytes en separar el text de la resta dels components de les pàgines (és a dir, imatges, vincles, etc.). Les pàgines (urls) recollides estan organitzades en 24.378 llocs web que al seu torn formen part de 14.094 dominis de segon nivell.

Classificació lingüística dels documents

Un cop recollits tots els documents del domini .es, es va procedir a la seva classificació en funció de la llengua, per tal de saber la composició lingüística del domini *.es* i identificar-ne la part

1 V. http://www.akwan.co.br i http://www.raditech.es.

catalana. Aquest procés es va fer en dues fases: mitjançant un sistema estadístic (ap. 2.2.1) i amb un filtre posterior sobre el corpus català, utilitzant informació lingüística.

Classificació mitjançant un sistema estadístic

Per poder identificar la llengua d'un document de manera automàtica, cal entrenar un sistema estadístic amb documents de cada una de les llengües que es vol reconèixer. Per tant, vam recollir una sèrie de corpus en deu idiomes, corresponents a les llengües més freqüents de la Web segons KILGARRIF i GREFENSTETTE 2003 (anglès, alemany, francès, castellà, italià, portuguès i neerlandès), més 3 llengües que suposàvem que serien freqüents al domini *.es*: l'èuscar, el gallec i el català.[2,3] Vam utilitzar el sistema de lliure distribució Bow (MCCALLUM 1996) per construir un classificador de llengües, basat en el mètode de Naive Bayes (v., p. ex., MITCHELL, 1997, o DUDA *et al.*, 2000).

Els corpus es van utilitzar per ajustar els paràmetres del classificador estadístic, i la classificació dels documents de la web es va realitzar de manera molt restrictiva per tal de minimitzar el nombre de falsos positius, és a dir, el nombre de documents classificats com a català que no ho fossin. La Taula 1 resumeix el resultat del procés de classificació inicial, amb la distribució del domini *.es* per idiomes en funció del nombre de documents, el nombre de paraules, i els Megabytes (MB) que ocupa cada idioma.[4]

2 En rigor, el gallec no és una llengua, sinó un dialecte del portuguès. Això no obstant, com que el classificador funciona amb qualsevol variant lingüística si es té prou corpus d'entrenament, vam diferenciar el gallec per tal de possibilitar l'estudi de la seva presència a la Web.

3 Els corpus utilitzats són els següents:
Anglès: *British National Corpus*. BURNARD (1995)
Alemany: textos del diari *Frankfurter Rundschau*, anys 1992-1993. Cedit per Universität Gesamthochschule (Paderborn, Alemanya).
Francès: textos dels diaris *Le Monde* i *Le Soir*, any 1995. Cedit per CENTAL (Lovaina, Bèlgica).
Castellà: corpus LexEsp. SEBASTIÁN et al., 2000.
Italià: fragment del corpus PAROLE. Cedit per ILC (Pisa, Itàlia).
Portuguès: textos del diari *O Publico*, any 1999. Cedit per CENTAL (Lovaina, Bèlgica).
Neerlandès: textos del diari *NRC Handelsblad*. Cedit per CENTAL (Lovaina, Bèlgica).
Euskera: fragments d'un corpus del grup IXA (Universidad del País Vasco). Cedit per aquest grup.
Gallec: fragment del corpus CLUVI. Cedit pel SLI (Vigo).
Català: fragment del corpus CTILC. Cedit per l'IEC.
Els autors expressem el nostre agraïment a totes aquestes institucions per la cessió dels corpus.

4 No hi ha les dades per a l'italià perquè el corpus es va rebre després d'haver construït el classificador. Els textos en

idioma	paraules (milions)	%paraules	paraules per document	documents (milers)	%documents	MB	%MB
Castellà	**3.179**	**48,8%**	**1.944**	**1.635**	**21,4%**	**19.441**	**46,6%**
Anglès	1.963	30,1%	1.421	1.381	18,1%	12.619	30,3%
Català	**366**	**5,6%**	**1.572**	**233**	**3,0%**	**2.387**	**5,7%**
Francès	158	2,4%	527	299	3,9%	1.052	2,5%
Gallec	**65**	**1,0%**	**2.854**	**23**	**0,3%**	**409**	**1,0%**
Alemany	22	0,3%	846	26	0,3%	165	0,4%
Èuscar	**22**	**0,3%**	**2.016**	**11**	**0,1%**	**172**	**0,4%**
Português	11	0,2%	1.921	6	0,1%	83	0,2%
Neerlandès	8	0,1%	1.623	5	0,1%	58	0,1%
Desconegut	726	*11,1%*	*180*	*4.027*	*52,7%*	*5.328*	*12,8%*
Total	6.519	100,0%		7.645	100,0%	41.714	100,0%

Taula 2. Distribució del domini .es per idiomes, ordenats de major a menor presència.

Com a resultat de la primera fase de classificació, es van obtenir 232.692 documents classificats com a català, el que correspon a un 3,04% dels documents. Si considerem com a mesura les paraules i no els documents, el volum de català és el 5,6% del total, i és el tercer idioma amb més presència al domini *.es*, després del castellà i l'anglès (en nombre de documents, el francès és lleugerament superior, però es tracta de documents amb menys text). També cal fer notar que, en funció del criteri escollit (nombre de documents, de paraules o de Megabytes), el català suposa entre un 10 i un 12% del volum total de textos escrits en un idioma oficial d'Espanya al domini *.es* (en negreta a la taula). El gallec i l'èuscar suposen un 1-2% i el castellà un 86-88% dels documents escrits en un idioma oficial.

Respecte als idiomes no oficials, només l'anglès i el francès tenen una presència significativa al domini *.es*, ja que l'alemany, el portuguès i el neerlandès no arriben ni al 1% de presència en cap dels criteris escollits. L'anglès, en canvi, representa gairebé un terç del domini *.es* mesurat en paraules i MB.

El criteri utilitzat, tan restrictiu, va fer que al 52,7% dels documents no se'ls assignés cap idioma

italià, doncs, es classifiquen de moment com a *desconegut*. En el futur integrarem aquest idioma.

(línia *Desconegut* a la taula). D'aquests, la majoria són documents amb molt poques paraules: el nombre mitjà de mots és de 180, mentre que la resta tenen una mitjana d'uns 1600 mots. Aquest baix nombre de paraules és, en la majoria dels casos, el que fa que no es puguin classificar fiablement. En aquest grup també hi ha els documents escrits en llengües per a les qual no teníem corpus d'entrenament, així com documents sense un idioma clar (documents formats bàsicament per imatges, documents multilingües).

Filtre lingüístic

La classificació automàtica permet fer-se una idea prou fiable de la distribució de la web en idiomes, però per construir un corpus cal filar una mica més prim. Això és el que es va dur a terme a la segona fase del procés de classificació, en què es va utilitzar un filtre addicional lligat a la informació lingüística pròpia del català. Per a aquesta finalitat, es va utilitzar una de les eines originalment creades per al marcatge lingüístic (v. ap. 2.3): el formari, diccionari de formes flexionades, que es va aplicar per marcar les paraules dels textos com a catalanes o no. Cap formari no pot ser complet, entre altres raons per la capacitat del llenguatge de crear noves paraules, però es va considerar que per tal de classificar un text com a adequat per formar part del corpus, un gran percentatge de les paraules havien de ser paraules existents i conegudes del català. Així, es van filtrar tots els documents que contenien més de 15% de paraules desconegudes.

A més del problema de textos no catalans, ens vam trobar altre cop amb el problema que hi ha documents que en realitat no es poden atribuir a una llengua en particular, sobretot llistats de noms propis, molta part dels quals no s'havien descartat a la primera fase (perquè tenien una frase de presentació en català, o alguna altra característica). Els noms propis no es poden llistar exhaustivament a un formari i per això es detecten mitjançant heurístiques. El problema central és que qualsevol nom propi es pot interpretar com a paraula existent del català, però també una paraula existent de totes les altres llengües. A més, fins i tot en cas de textos que es poden atribuir al català,

el problema és que no solen ser textos coherents si contenen molts noms propis perquè la majoria d'ells són o contenen mers llistats. No representen l'ús de la llengua en el sentit estricte. Per això tampoc no interessava incloure aquests documents en un corpus lingüístic, i es va descartar tots els documents que contenien més de 30% de paraules classificades com a noms propis.

Això va reduir la mida del Corpus fins a un volum de 165.386 documents i 262 milions de paraules. Finalment, com que a la web hi ha entre un 10 i un 30% de documents duplicats, vam aplicar un procés de detecció de documents duplicats, la qual cosa va reduir el corpus als 125.000 documents amb 208 milions de paraules que té actualment. Malgrat aquest estricte procés de classificació i depuració, al corpus hi ha documents multilingües, amb predomini del català però amb fragments en d'altres idiomes, cosa que cal tenir en compte a l'hora d'extreure'n informació.

Processament dels textos

Els textos extrets de la web i ja classificats com a català es van etiquetar amb un seguit de gramàtiques per al català desenvolupades al grup GLiCom de la Universitat Pompeu Fabra, el sistema CatCG (v. ALSINA *et al.*, 2002). Aquesta eina segmenta els textos en paràgrafs i oracions, detecta noms propis, contraccions, pronoms clítics i d'altres formes especials, i finalment assigna lema, categoria morfològica i funció sintàctica a cada mot. El resultat de l'aplicació de l'eina a l'oració *La fi de la guerra va suposar la fi de la lluita contra el règim* està exemplificat a la Fig. 1.

Finalment, els textos es van processar amb les eines del *Corpus WorkBench* (CWB), desenvolupades a l'Institut für Maschinelle Sprachverarbeitung de la Universitat de Stuttgart (CHRIST, 1994).[5] Aquestes eines, de lliure disposició per a finalitats de recerca, indexen els textos de manera que es puguin consultar de manera ràpida i eficient mitjançant CQP (*Corpus Query Processor*), el motor de cerca de corpus del CWB. El CQP permet una gran flexibilitat i expressivitat en les cerques, ja que s'hi pot utilitzar qualsevol expressió regular, i la consulta

5V. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

eficient de qualsevol de les informacions lingüístiques. Això no obstant, és una eina molt poc *user-friendly*, amb una sintaxi pròpia, per la qual cosa es va dissenyar una interfície web adequada a les necessitats de l'usuari potencial, tal i com veurem a l'apartat 4.

Característiques del fragment català del domini .es

Per caracteritzar la Web catalana, considerem els dominis de segon nivell que pengen del domini *.es* i en els quals hem detectat almenys una pàgina en català. Això fa un total de 1.548 dominis, dels quals 407 tenen una sola pàgina en català. A més, la majoria de les pàgines estan concentrades en uns quants dominis: 30 dominis contenen el 70% de les pàgines en català i amb només 13 dominis s'arriba a la meitat de les pàgines. A la Taula 2 es presenten dades dels 15 dominis de la web que tenen major nombre de pàgines en català.

domini	català	castellà	anglès	desconegut	percentatge	acumulat
upc	20.511	11.106	36.961	98.277	8,94%	8,94%
uab	18.583	7.872	27.024	61.525	8,10%	17,05%
caib	10.020	1.415	215	10.713	4,37%	21,41%
udg	9.676	2.058	9.137	24.873	4,22%	25,63%
uji	8.926	17.837	15.864	43.940	3,89%	29,53%
diba	8.696	838	337	8.724	3,79%	33,32%
ub	6.543	7.606	11.233	43.641	2,85%	36,17%
upf	5.999	2.258	7.902	22.062	2,62%	38,79%
gencat	5.913	442	731	4.320	2,58%	41,36%
fut	5.844	1.217	270	5.263	2,55%	43,91%
urv	5.597	2.490	9.360	26.623	2,44%	46,35%
xtec	5.477	1.580	852	14.670	2,39%	48,74%
udl	5.179	1.168	2.756	24.750	2,26%	51,00%
uib	4.817	2.028	5.196	16.626	2,10%	53,10%
bcn	4.782	1.384	1.620	7.041	2,09%	55,18%

Taula 2 Nombre de pàgines en català, castellà i anglès en els 15 dominis amb major nombre de pàgines en català. També s'inclou el nombre de pàgines que no es van poder classificar en un idioma i els percentatges de pàgines que té cada domini respecte del total de documents.

Tots aquests dominis corresponen a universitats i institucions públiques. De fet, les universitats tenen el 42% de totes les pàgines en català del domini *.es*. Respecte al caràcter bilingüe o multilingüe de la web catalana, cal fer notar que 266 dominis tenen més del 90% de les seves pàgines en català, 624 en tenen menys del 10%, i la resta (669) entre el 10% i el 90%.

Explotació

L'explotació del corpus es pot fer bàsicament de dues maneres: a través de la interfície web i a partir dels fitxers del corpus com a recurs en local, ja que el corpus és de lliure distribució per a finalitats de recerca. Per exemple, ja s'està fent servir per classificar els verbs automàticament en funció del tipus de complements que poden tenir (verbs transitius, intransitius, etc.); v. MAYOL *et al.* (en preparació). Això no obstant, es preveu que l'ús majoritari del corpus sigui a través de la interfície.

La interfície web, disponible a l'adreça http://www.catedratelefonica.upf.es/, és molt flexible respecte de la mena de cerques que es poden fer. Inclou dues interfícies de cerca (cerca d'exemples i cerca de freqüències), i per a la cerca d'exemples s'han previst dues modalitats de cerca: el mode **simple**, que permet fer cerques d'una o més paraules, lemes, categories morfològiques o funcions sintàctiques, i el mode **expert**, que permet buscar fins a cinc elements amb possibilitat de restringir qualsevol d'aquests elements, tal i com veurem a continuació amb més detall.

Cerca d'exemples, mode simple

Al mode simple, només cal introduir la paraula o paraules que volem cercar i prémer el botó "Mode simple":

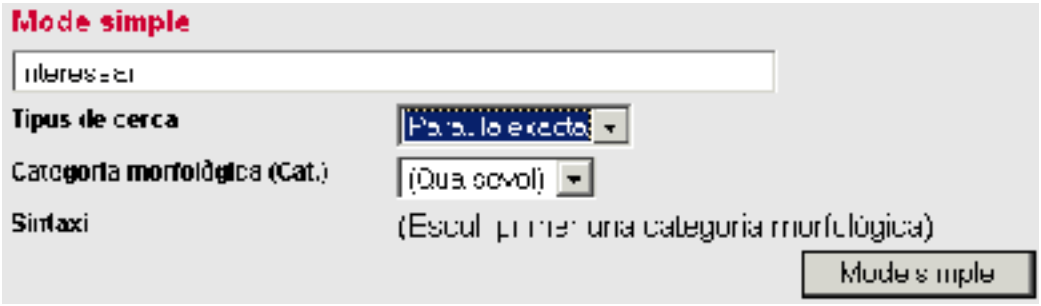

Fig. 10. Cerca simple: "interessar", paraula exacta.

Veiem a la següent figura el resultat (parcial) que obtindrem amb la cerca "interessar":

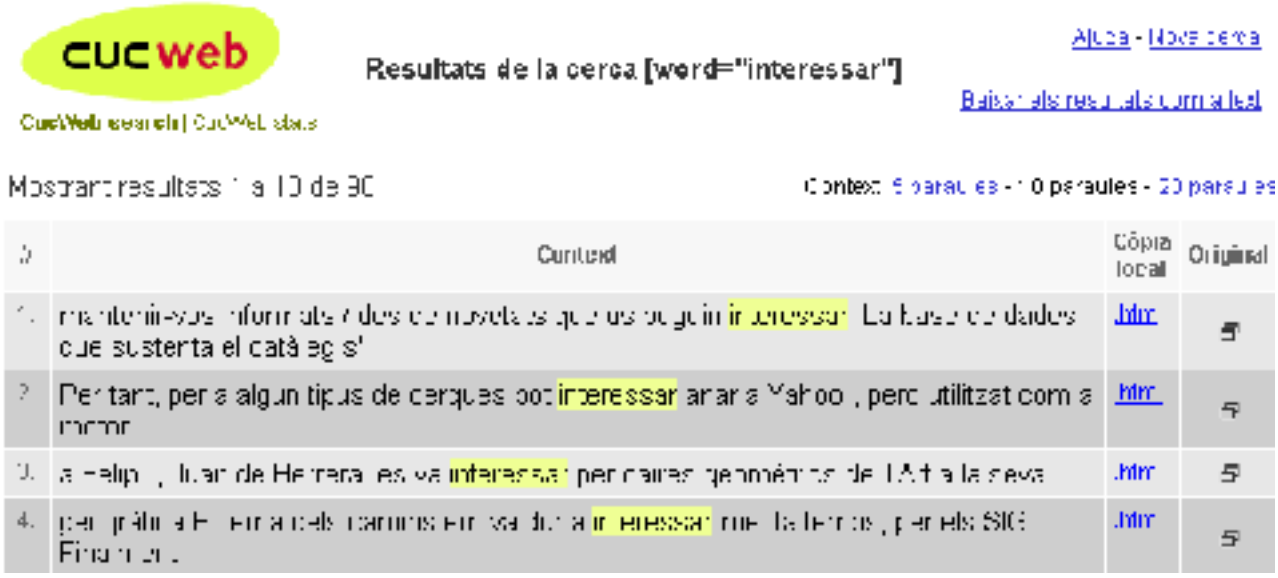

Fig. 10. Resultats per a la cerca simple "interessar", paraula exacta.

Observeu que s'ofereixen moltes facilitats per visualitzar i recuperar la informació: baixar els resultats en un fitxer de text, canviar el context, accedir a una còpia local del document on s'ha trobat l'exemple, i accedir a la url original.

Si en lloc de buscar per forma busquem per lema (selecció a partir de *Tipus de cerca*), obtindrem més resultats, ja que recuperarem qualsevol forma del verb *interessar*. Podem així mateix especificar més d'una paraula o lema, separats per espai. Si busquem "interessar per" i especifiquem el tipus de cerca "lema", obtindrem els resultats reflectits parcialment a la següent figura:

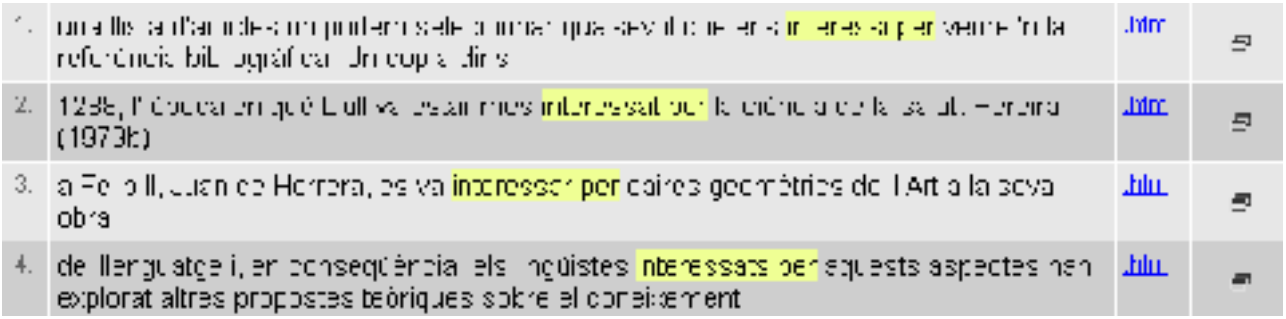

Fig. 10. Resultats per a la cerca simple “interessar per”, cerca per lema.

El tipus de cerca es pot restringir encara; per exemple, es pot buscar la paraula *roda* especificant alhora que sigui verb (opció “Categoria morfològica”), i s’obtenen exemples com els següents:

El cavall de regalo, quan es vell roda la sínia
Document lliurat per el Departament en roda de premsa
Immòbil sobre el punt d'una roda que roda sobre el pla de l'escenari faig la cicloide

Fig. 10. Alguns resultats per a la cerca simple “roda”, paraula exacta, especificant “Verb”.

Observi’s que hi ha errors, com a *en roda de premsa*, en què es tracta d’un nom. Això és degut que l’etiquetatge és automàtic, i l’etiquetador l’ha etiquetat com a verb.

Finalment, es pot restringir la funció sintàctica, per exemple buscant la paraula *roda* com a nom i en funció de complement directe, amb la qual cosa s’obtenen exemples com els següents:

desembragant la roda dentada que s'uneix al motor
eliminant l'obligació de incorporar el maleter i la roda de recanvi
plantejar a un grup de 5 xiquets que arreglaren una roda de platja punxada.

Fig. 10. Alguns resultats per a la cerca simple “roda”, paraula exacta, especificant “Nom” i “Complement Directe”.

Cerca d’exemples, mode expert

El mode expert és més flexible però una mica més complex i menys transparent que el mode simple. Està pensat per a usuaris que estiguin familiaritzats amb el mode simple o bé amb d’altres interfícies de cerca de corpus. Es poden fer cerques sobre fins a cinc unitats o paraules, i permet fer

cerques sobre diferents tipus d'informació en funció de la unitat. Per exemple, la següent cerca no es pot fer en el mode simple:

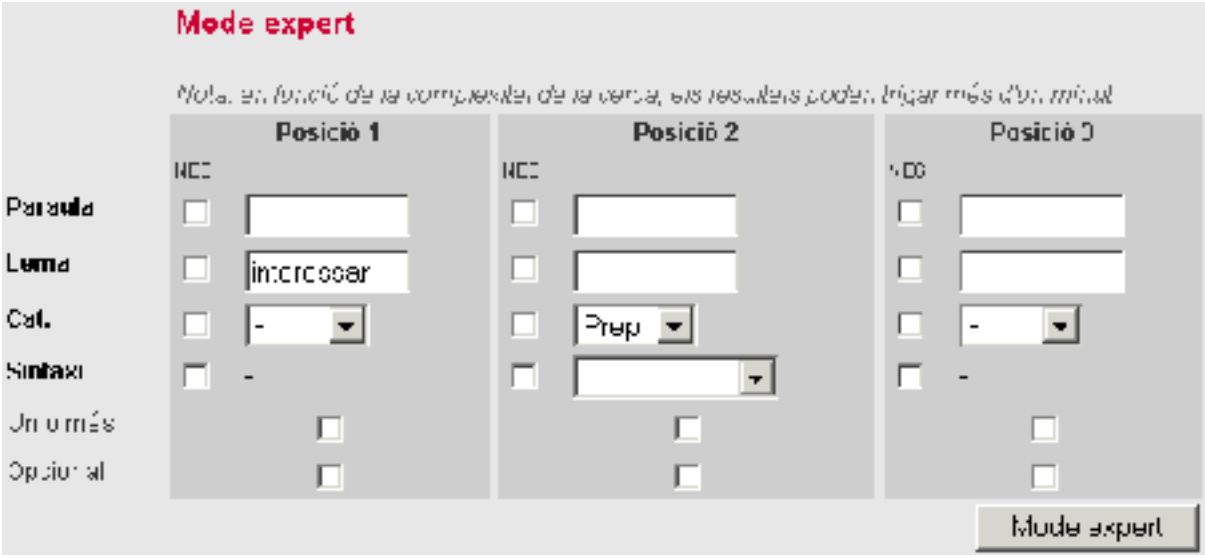

Fig. 10. Cerca experta: lema "interessar" seguit de preposició.

Així busquem el verb *interessar* seguit de preposició, sense especificar quina, a diferència d'abans, en què havíem d'especificar el lema de la preposició que ens interessava. Amb aquesta cerca obtenim resultats com els següents:

una llista d'articles on podem seleccionar qualsevol que ens interessi per veure 'n la referència bibliogràfica.
Web molt interessant sobre paremiologia catalana.
Els grups de recerca que estiguin interessats en usar aquest servei
Està especialment pensat per a persones interessades a conèixer l'obra d'autors llatinoamericans.
un punt de trobada, relació i intercanvi entre persones interessades en la novel·la històrica

Fig. 10. Alguns resultats per a la cerca experta de lema "interessar" seguit de preposició.

Aquest mode de cerca ofereix també la possibilitat de negar elements (p.ex., *interessar* no seguit de preposició), d'especificar més d'un element (per exemple, nom seguit d'un o més adjectius) i d'expressar opcionalitat (per exemple, nom seguit opcionalment d'un adverbi seguit d'adjectiu). Igual com la cerca simple, també permet buscar sobre funcions sintàctiques.

Cerca de freqüències

Per a segons quines finalitats, com per exemple l'elaboració de diccionaris, no cal només mirar exemples, sinó avaluar la freqüència relativa dels fenòmens que s'estudien. Per exemple, per a un lexicògraf és rellevant d'observar quina mena de complements són més freqüents amb un verb determinat, per tal d'ordenar i escollir les accepcions i els exemples. La interfície estadística del corpus està pensada per fer aquest tipus de cerques.

Seguint amb l'exemple d'*interessar*, si volem saber quines preposicions són més freqüents, podem fer la cerca següent.

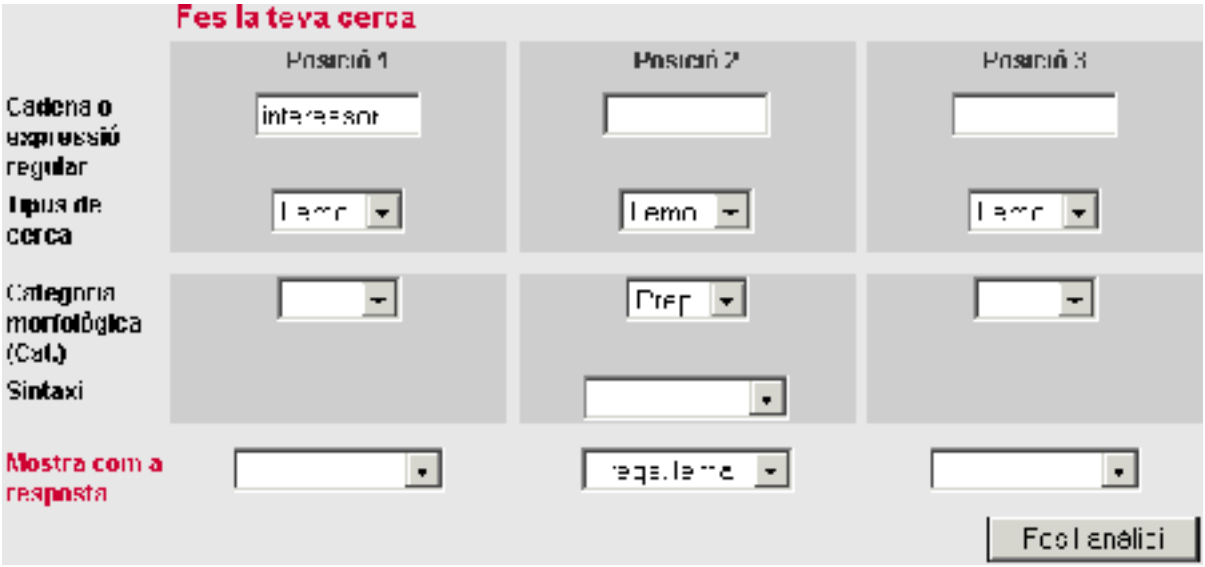

Fig. 10. Cerca de freqüències: lema "interessar" seguit de preposició. Retorna les freqüències dels lemes de les preposicions.

En aquesta cerca, demanem que ens llistin les freqüències dels lemes de les preposicions que segueixen el lema *interessar*. Els resultats que obtenim mirant els 10 primers milions del corpus són els següents.[6]

6 Aquesta cerca triga 18 segons. El fragment de corpus que sobre el qual es vol fer l'estadística es pot especificar, ja que les estadístiques sobre els 208 milions de corpus poden trigar 10 minuts o més.

Frequency			Position 1	Position 2	
Relative	Cumulative	Absolute	Lema	Lema	Categoria morfològica (Cat.)
41,20%	41,20%	151	interessar	en	[illegible]
31,88%	73,23%	117	interessar	a	[illegible]
17,21%	90,41%	63	interessar	per	[illegible]
5,73%	96,17%	21	interessar	de	[illegible]
1,36%	97,26%	4	interessar	amb	[illegible]
0,81%	98,08%	3	interessar	sobre	[illegible]
0,54%	98,63%	2	interessar	davant	[illegible]
0,27%	98,90%	1	interessar	segons	[illegible]
0,27%	99,18%	1	interessar	des	[illegible]
0,27%	99,45%	1	interessar	sense	[illegible]
0,27%	99,73%	1	interessar	dins	[illegible]
0,27%	100,00%	1	interessar	cap	[illegible]

Fig. 10. Resultats per a la cerca de freqüències del lema “interessar” seguit de preposició.

Veiem que la preposició més freqüent és *en*, però això pot resultar enganyós. Si mirem les **formes** del verb *interessar* que apareixen amb cada preposició (especificar addicionalment “Freqs. forma” a la Posició 1, opció “Mostra com a resposta”), veurem que el que apareix amb la preposició *en* són els participis (*interessada*, *interessats*), i en canvi la resta de formes (*interessa*, *interessar*, *interessen*) apareixen més freqüentment amb *per*. Aquesta mena d’informació és crucial per a un lexicògraf, i útil per a redactors i d’altres usuaris amb necessitats lingüístiques. També es pot explotar aquesta informació en la docència del català, tant per a ensenyament a nadius (sobretot a secundària i batxillerat) com per a aprenents del català com a llengua estrangera.

Conclusions i camins futurs

En aquest article hem presentat el CuCWeb, Corpus d’Ús del Català a la Web, un corpus construït i etiquetat de manera totalment automàtica a partir de la web. A més de construir el corpus, hem dissenyat i posat a disposició del públic una interfície que permet consultar-lo de manera molt flexible.

La web té molts avantatges davant d’altres recursos per construir corpus: és de lliure distribució, conté grans masses textuals i s’hi pot accedir de manera automàtica. També té desavantatges:

sobretot, el fet que no hi ha massa control sobre el seu contingut, ni d'edició ni de cap altra mena. Això vol dir que ens trobarem documents amb errors d'ortografia, gramaticals, amb barreges d'idiomes, estrangerismes, etc. Aquí és també, però, on resideix el repte i la riquesa d'aquest corpus, perquè permet estudiar la llengua en un entorn no academitzat.

Creiem que el corpus pot ser útil a professionals de la llengua (traductors, mestres, lexicògrafs, lingüistes) i a usuaris de la llengua en general. Ja és molt habitual utilitzar *Google* per a consultes lingüístiques, tant, que fins i tot s'ha habilitat una interfície per fer-les.[7] La interfície del CuCWeb, a http://www.catedratelefonica.upf.es/, permet també fer aquesta mena de consultes específicament per al català i amb possibilitats de cerca molt més riques.

Amb un corpus de 208 milions de paraules, però, queda molt de camí per recórrer. No es pot inspeccionar de manera manual, per la qual cosa cal desenvolupar mètodes automàtics per analitzar i classificar el seu contingut. La part lingüística està resolta amb l'etiquetador automàtic; què es pot dir, però, de la part extralingüística? És a dir, quina mena de documents hi ha? De quins gèneres i temàtiques? En el futur, caldrà dedicar-se a aprofundir en aquesta mena de qüestions, que estan a la frontera entre l'estudi lingüístic i el sociolingüístic. Una primera aproximació pot ser classificar els documents segons l'origen, i permetre per exemple fer cerques sobre pàgines d'universitats, o d'una determinada institució pública.

Un altre repte és el del multilingüisme. Fins ara, tot l'etiquetatge està fet com si els documents fossin escrits totalment en català. Caldrà identificar i separar les parts escrites en altres idiomes, i es podria pensar d'integrar-los a la interfície també, de manera que es poguessin fer cerques sobre aquests fragments de manera separada. Així mateix, es podria aprofitar el fet que hi ha moltes pàgines que tenen versió bilingüe (o trilingüe) per extreure automàticament corpus paral·lels, és a dir, corpus amb la mateixa informació en versió catalana, castellana, anglesa, etc. Aquesta mena de

7 A http://cli.la.asu.edu/togoogleornot.htm.

recurs és molt útil tant per a aproximacions científiques a la llengua (estudis interlingüístics) com per a aplicacions (traducció automàtica, diccionaris multilingües).

Finalment, seguint amb la diversitat lingüística, una altra línia de recerca que s'obre és l'ampliació del procés a d'altres dominis. Estem processant els dominis *.com*, *.net*, *.org* i d'altres per tal d'identificar-ne les parts en català, ja que la majoria de pàgines personals i d'empreses són en dominis que no són *.es*, per raons sociològiques. Igualment, es pot ampliar a d'altres idiomes o varietats, pensant per exemple en comparar el castellà de Xile i el castellà d'Espanya. Les possibilitats, com la web mateixa, són immenses.

Bibliografia

ALSINA, À., T. BADIA, G. BOLEDA, S. BOTT, À. GIL, M. QUIXAL, O. VALENTÍN, 2002, "CATCG: a general purpose parsing tool applied", a *Proceedings of Third International Conference on Language Resources and Evaluation (LREC)*, Las Palmas

BURNARD, L., 1995, *The BNC Reference Manual*, Oxford: Oxford University Computing Service

CHRIST, O., 1994, "A modular and flexible architecture for an integrated corpus query system", a *Proceedings of COMPLEX'94*, Budapest

DUDA, R. O., P. E. HART i D. G. STORK, 2000, *Pattern Classification*, New York: John Wiley & Sons

KILGARRIFF, A. i G. GREFENSTETTE, 2003, "Introduction to the Special Issue on the Web as Corpus", a *Computational Linguistics*, vol. 29, no. 3, pp. 333-348

MAYOL, L., G. BOLEDA i T. BADIA, en preparació, "Automatic Learning of Syntactic Verb Classes"

MCCALLUM, A., 1996, Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, disponible a http://www.cs.cmu.edu/~mccallum/bow/.

MITCHELL, T. M., 1997, *Machine Learning*, New York: Mc Graw Hill

POBLETE, B., M. E. FUENMAYOR, C. CASTILLO, R. BAEZA-YATES, V. LÓPEZ, "Características de la Web española", en preparació

RAFEL, J. 1994. Un corpus general de referència de la llengua catalana", a *Caplletra*, vol. 17, pp. 219-250

SEBASTIÁN, N., CUETOS, F., MARTÍ, M.A., CARREIRAS, M.F., 2000, *LEXESP: Léxico informatizado del español*, edició en CD-ROM, Barcelona: Edicions de la Universitat de Barcelona