CUCWEB: UN CORPUS DE LA LLENGUA CATALANA CONSTRUÏT A PARTIR DE LA WEB

Toni Badia / Gemma Boleda
Universitat Pompeu Fabra i Centre d'Innovació Barcelona Media

Aquesta nota pretén difondre dins la comunitat d'estudiosos de les llengües romàniques l'existència d'un nou recurs per a l'estudi del català: un corpus del català creat a partir de la Web. El corpus fou constituït per mitjans totalment automàtics a partir de la descàrrega i processament de les pàgines Web del territori espanyol, i es pot consultar *online* mitjançant una interfície molt flexible. Es tracta d'un projecte conjunt entre el GLiCom[1] (Grup de Lingüística Computacional) de la Universitat Pompeu Fabra i la Càtedra Telefónica de Comunicació Multimèdia.[2]

Els corpus en format electrònic són una eina cada vegada més usada per un ampli ventall d'estudiosos i professionals de la llengua: lexicògrafs (v. el projecte del *Corpus Textual Informatitzat de la Llengua Catalana*, Rafel, 1994),[3] lingüistes, traductors, mestres, etc. Quan disposen d'una interfície de consulta adequada, els corpus permeten accedir a molts exemples d'ús real de mots o construccions molt variats de manera ràpida i sistemàtica. Això permet complementar les dades obtingudes d'altres fonts documentals, lexicogràfiques, o d'altra mena.

El CUCWeb (Corpus d'Ús del Català a la Web) és un projecte dirigit a construir corpus a partir de la Web i fer-los accessibles mitjançant una interfície web a qualsevol usuari interessat en la llengua catalana. Aquesta interfície està disponible a: *http://www.catedratelefonica.upf.es/cucweb*. Per a detalls sobre el procés de construcció del corpus i el funcionament de la interfície, vegeu Boleda et al. (2006) i Badia / Boleda (en preparació); aquí oferirem només un resum d'aquests i d'altres aspectes relacionats amb el projecte.

En el procés de construcció del corpus pròpiament dit, en primer lloc es van descarregar les pàgines del domini *.es*, i després es va ampliar la col·lecció a d'altres dominis, com ara *.com*, *.org*, *.net*, i *.info*. Les dues compilacions es feren l'any 2004. Mitjançant un classificador de llengües automàtic es van identificar els documents en català (un 8% del total), que es van filtrar mitjançant un diccionari computacional del català per tal d'assegurar-ne la qualitat lingüística. Aquest procés va donar com a resultat dos corpus, un de 208 milions de mots (domini *.es*) i un de 166 milions de mots (altres dominis).

Els corpus es van processar mitjançant CatCG, un seguit d'eines computacionals per al català disponibles al GLiCom (Alsina et al., 2002), per tal d'afegir-hi informació lingüística,

1. *http://glicom.upf.es/*.
2. *http://www.catedratelefonica.upf.es/*. En aquest projecte hi han participat, a més dels autors: Stefan Bott, Carlos Castillo, Rodrigo Meza, i Barbara Poblete. Ha estat dirigit per Vicente López, director de la Càtedra Telefónica de Comunicació Multimèdia.
3. La interfície a aquest corpus està disponible a: *http://ctilc.iec.cat/*.

DOI: 10.2436/20.2500.01.20

Estudis Romànics [Institut d'Estudis Catalans]
Vol. 30 (2008), p. 291-293

sobretot el lema, categoria morfològica i funció sintàctica. Per exemple, per a la frase *La Maria és alta*, la CatCG permet codificar el mot *alta* com a adjectiu femení singular, amb lema *alt* i funció sintàctica d'atribut. Com que aquest procés és, de nou, completament automàtic, en l'anotació resultant hi haurà errors causats per deficiències del programa i ambigüitats inherents a la llengua. Això no obstant, els errors queden compensats per la quantitat de dades obtingudes. Certament, hi ha una correlació inversa entre la quantitat de text que és raonable d'obtenir i el nombre d'errors que hi poden aparèixer.

Tal i com hem dit més amunt, l'exploració de corpus grans és impossible sense una interfície adequada. La manera més fàcil i directa de crear una interfície és precisament a partir de la web. La interfície del CUCWeb permet fer cerques d'exemples cercant mots, lemes, categories morfològiques o funcions sintàctiques. Per exemple, es poden cercar ocurrències del lema verbal *interessar* i obtenir exemples d'ús d'aquest verb en totes les seves formes. Per a usuaris més avançats, es poden buscar combinacions d'unitats lingüístiques, fins a cinc unitats. Per exemple, es pot cercar el lema *interessar* seguit de preposició, per tal d'estudiar el règim preposicional d'aquest verb. Obtindrem així exemples en què s'usa *s'interessava per*, *interessat en*, *interessat a*, etc. Per a molts propòsits, però (per exemple, per a objectius lexicogràfics), són útils no només exemples, sinó també freqüències d'ús. La interfície de CUCWeb permet fer cerques estadístiques, en què es pot observar per exemple que *interessar* va seguit per la preposició *en* sobretot en la forma participial, mentre que en formes finites del verb la preposició més freqüent és *per*.

Els corpus del CUCWeb, per la seva mida i la seva constitució, complementen d'altres recursos disponibles per a la llengua catalana, el més important dels quals és el *Corpus Textual Informatitzat de la Llengua Catalana* (CTILC) de l'Institut d'Estudis Catalans, esmentat més amunt. El CTILC conté uns 50 milions de mots provinents de textos de diversos gèneres literaris i àrees acadèmiques, escrits en el darrer segle i mig (1832-1988). Ha estat compilat i etiquetat de forma semiautomàtica, amb un rigorós control manual sobre el seu contingut i el processament ulterior de les dades. En canvi, els textos del CUCWeb són extrets de la Web. Els corpus són més grans, però contenen més soroll (pàgines o fragments en d'altres idiomes, errors d'etiquetatge, etc.). Els textos que el formen han estat sotmesos a menys control editorial durant la seva redacció que la majoria de textos del CTILC. Això representa un desavantatge (presència d'errors ortotipogràfics, gramaticals, etc.), però també un avantatge, ja que compila el llenguatge usat per un ventall ampli de persones que no són autores d'articles o llibres literaris, acadèmics o científics. A més, el llenguatge de la Web és fortament sincrònic, tot i la presència esporàdica de textos més antics. Aquesta característica és de nou un desavantatge (perquè no permet estudis diacrònics, almenys no fins d'aquí uns anys) i un avantatge, per la possibilitat d'estudiar fenòmens recents (detecció de neologismes, canvis gramaticals en curs, etc.).

En resum, creiem que el CUCWeb pot representar un recurs valuós per a estudis i consultes lingüístiques sobre el català per a diverses menes d'usuaris, des d'investigadors (lingüistes, traductòlegs) fins a professionals (lexicògrafs, traductors, mestres) o fins i tot usuaris de la llengua que volen resoldre dubtes gramaticals o ortogràfics.

REFERÈNCIES BIBLIOGRÀFIQUES

ALSINA, Àlex / BADIA, Toni / BOLEDA, Gemma / BOTT, Stefan / GIL, Àngel / QUIXAL, Martí / VALENTÍN, Oriol (2002): «CATCG: a general purpose parsing tool applied». *Proceedings of*

Third International Conference on Language Resources and Evaluation (LREC). Vol. III. Las Palmas, p. 1130-1135.

BADIA, Toni / BOLEDA, Gemma (en preparació): *CUCWeb, Corpus d'Ús del Català a la Web*.

BOLEDA, Gemma / BOTT, Stefan / CASTILLO, Carlos / MEZA, Rodrigo / BADIA, Toni / LÓPEZ, Vicente (2006): «CUCWeb: a Catalan corpus built from the Web». *Proceedings of the Second Workshop on the Web as a Corpus at EACL'06*. Trento, Itàlia.

RAFEL, Joaquim (1994): «Un corpus general de referència de la llengua catalana». *Caplletra*. Vol. 17, p. 219-250.