Eines de lingüística computacional per a la traducció: corpus paral·lels anotats

Toni Badia, Gemma Boleda, Carme Colominas, Mireia Garmendia, Agnès González, Martí Quixal
Universitat Pompeu Fabra
Rambla 30-32 ,
E-08002 Barcelona
{toni.badia,carme.colominas,marti.quixal}@trad.upf.es, gemma.boleda@iula.upf.es

1 Resum

Les eines desenvolupades al camp de la lingüística computacional tenen un gran potencial per a la traducció, tant per a la didàctica com per a la recerca com per a la pràctica professional. En aquest article presentem un projecte que pretén explorar aquest camí: BancTrad (http://glotis.upf.es/bt/index.html). Aquest projecte pretén oferir accés via web a corpus paral·lels, amb finalitats tan diverses com la didàctica de la traducció o la recerca lingüística.

La particularitat de BancTrad és que, com que integra eines de processament del llenguatge natural, permet cercar no només sobre paraules (cadenes de caràcters), sinó també sobre dues altres menes de característiques:

a) lingüístiques: sobre lema, categoria morfològica i –en el cas del català- funció sintàctica
b) extralingüístiques: sobre trets tals com gènere textual, registre, tema, etc. (quinze paràmetres en total)

Aquest article s'estructura de la manera següent: a l'apartat 2 s'hi explica el procés de compilació, marcatge i alineació dels textos de BancTrad. L'apartat 3 està dedicat a l'etiquetatge lingüístic i la construcció dels corpus. L'apartat 4 exposa aspectes tècnics tals com el motor de cerca i la interfície web. L'apartat 5 explica les possibilitats de cerca que s'ofereixen a BancTrad i les aplicacions previstes de l'eina. Finalment, l'article acaba amb algunes conclusions i agraïments

2 Compilació de textos, marcatge i alineació

Les llengües dels corpus de BancTrad són les llengües de treball a la Facultat de Traducció i Interpretació (FTI) a la Universitat Pompeu Fabra: català, castellà, anglès, francès i alemany (tot i

que estan previstes totes les combinacions, actualment només hi ha textos paral·lels del català i al castellà a l'anglès, francès, alemany i viceversa).

Es pretén que els textos de BancTrad siguin representatius en tant que textos traduïts, és a dir, no tenen un caràcter normatiu sinó descriptiu. Per això es va decidir recollir documents de fonts molt diverses, que representessin un ventall ample de tipus de text, temes i registres. Les fonts principals de textos són el cos docent de la facultat, la feina que es fa a les classes de traducció, algunes editorials i Internet. La feina que fan els docents com a *free-lance* ens proporciona traduccions d'alta qualitat; a més, el fet d'incloure traduccions fetes (de manera supervisada) a classe pot ser molt útil per propòsits docents i acadèmics (v. apartat 4). Quant a Internet, els textos que en provénen passen abans per un procés de selecció per assegurar-ne la qualitat.

Un cop els textos estan seleccionats, es processen per tal d'etiquetar-los (en el format SGML) amb informació extralingüística, mitjançant un formulari de MS Word (v. Fig. 1).

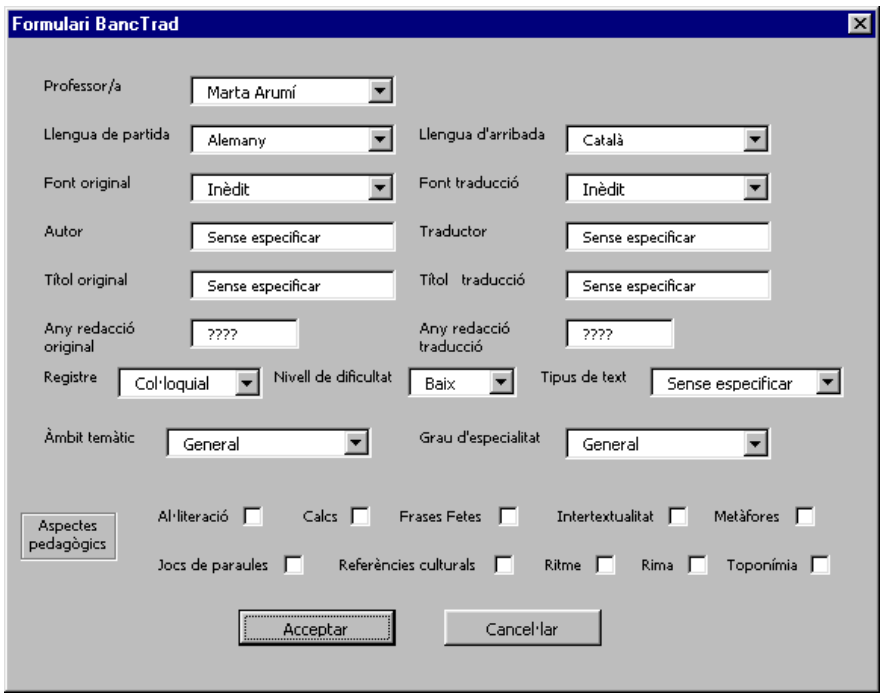

Figura 1: Formulari de MS Word que s'usa per al marcatge de trets extralingüístics

Els paràmetres que es marquen i la informació que contenen es van consensuar amb el cos docent i investigador de la FTI. Són els següents:

- Nom de qui introdueix el text

- Llengües de partida i d'arribada
- Referències de l'original i la traducció
- Registre (col·loquial, estàndard, formal, acadèmic)
- Tipus de text (normatiu, descriptiu, de divulgació, literari)
- Tema (economia, ciència, política, etc.)
- Grau d'especialització (baix, mitjà, alt)
- Aspectes pedagògics: presència rellevant de fenòmens tals com metàfora, calcs, frases fetes, etc.

Mitjançant el formulari i el llenguatge VisualBasic s'agilita la feina de formateig (tots els formats es converteixen a ascii) i marcatge, alhora que s'eviten errors de tecleig, ja que només cal seleccionar la informació, i el marcatge es realitza automàticament. A més, el formulari marca també l'estructura de paràgrafs del text, que si no es perdria en el procés d'alineació.

Quant a aquesta última tasca, els texts s'alineen a nivell oracional mitjançant l'eina *DéjàVu Database Maintenance*, d'Atril (http://www.atril.com). Aquest programa realitza una prealineació automàtica, que es pot editar de manera *user-friendly*, evitant d'aquesta manera els errors de l'alineació completament automàtica (sobretot en la identificació d'oracions), que haurien incrementat molt els errors en l'anàlisi lingüística (v. apartat següent).

Cal fer notar que les tasques descrites en aquest apartat no exigeixen coneixements d'informàtica especialitzats i, tot i que només són parcialment automàtiques, no resulten massa costoses: el temps que es necessita per marcar i alinear un text de 400 paraules és d'entre 5 i 10 minuts.

Un cop passat aquest procés, els textos es transfereixen al servidor Linux per tal de continuar-ne el processament, que a partir d'ara serà completament automàtic.

3 Processament lingüístic i construcció dels corpus

Els passos que cal fer encara abans de poder fer cerques són l'anàlisi lingüística i el formateig dels corpus. A continuació exposem els detalls de cada pas.

3.1 Anàlisi lingüística

No totes les llengües segueixen el mateix procés d'etiquetatge. Els textos en català s'analitzen mitjançant CATCG (Badia *et al.* 2000), un *shallow parser* basat en el formalisme de la Constraint Grammar i desenvolupat a la UPF. Es preveu que els textos en espanyol segueixin un procés paral·lel amb una versió espanyola d'aquest parser, d'aquí a un any aproximadament. En canvi, els textos en anglès, alemany i francès s'etiqueten mitjançant TreeTager, un etiquetador morfològic desenvolupat a l'IMS (v. Schmid 1995, 1997).

Cal fer notar que, tot i que els processos són diferents, totes les llengües reben el mateix tipus d'informació lingüística: lema, categoria morfosintàctica i trets morfològics tals com gènere i nombre (els textos catalans també es marquen quant a la funció sintàctica). Això permet processar i fer cerques de manera uniforme sobre totes les llengües. Aquest disseny modular del procés fa que es puguin canviar les eines usades en qualsevol dels moments del procés sense haver de canviar ni la interfície ni la resta de mòduls.

3.2 Construcció dels corpus

Un cop anotats, els fitxers de text es formategen i es processen amb les eines del *Corpus WorkBench* (CWB), desenvolupades a l' IMS (Christ 1994; Christ *et al*. 1999). Així, els corpus es fan consultables mitjançant CQP (*Corpus Query Processor*) una de les eines del CWB. Aquesta eina permet una gran flexibilitat i expressivitat en les cerques, ja que es pot utilitzar qualsevol expressió regular, i la consulta eficient de qualsevol de les anotacions exposades més amunt. A més, pot processar corpus alineats, cosa que el fa especialment adient per a BancTrad. Això no obstant, és una eina molt poc *user-friendly*, per la qual cosa es va dissenyar una interfície web adequada per als usuaris potencials de BancTrad.

4 El motor de cerca i la interfície web

Tècnicament, la novetat de BancTrad és la integració de diverses eines per a accedir a corpus paral·lels a través d'Internet. Això vol dir que el sistema ha de poder (1) interpretar la cerca feta per l'usuari, (2) efectuar la cerca i (3) presentar-ne els resultats. Per a acomplir això, calien dues eines: una interfície gràfica (GUI, *Graphical User Interface*) amb un formulari i una EPI (*External Program Interface*) per dur a terme la comunicació entre el servidor i el navegador. Aquesta arquitectura està esquematitzada a la Fig. 2.

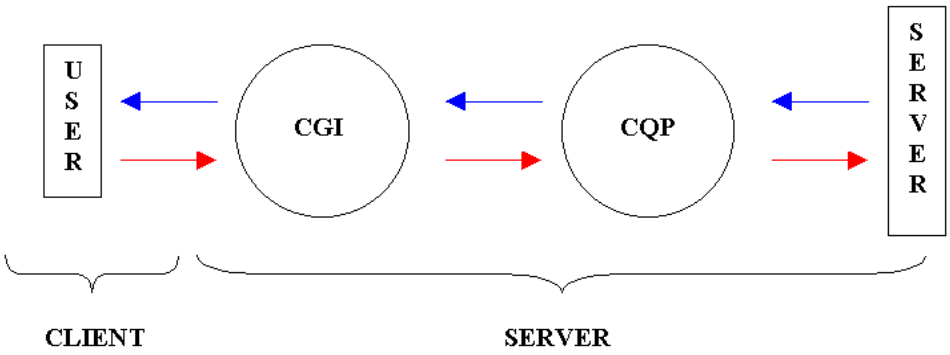

Figura 2: Arquitectura de la interfície de BancTrad (la cerca en vermell, els resultats en blau)

La interfície gràfca es va fer en HTML per assegurar-ne la llibertat d'accés i la independència de plataforma i sistema operatiu. L'EPI es va realitzar amb el llenguatge Perl, usant la CGI estàndard, un mòdul especial de formateig HTML i un mòdul de CQP dissenyat pels seus autors per tal d'efectuar la interacció amb la web.

5 Possibilitats de cerca i aplicacions

Aquest apartat exposa diferents maneres d'usar BancTrad, des de dues perspectives diferents però relacionades amb l'usuari: les possibilitats de cerca que ofereix BancTrad (ap. 5.1), aspecte relacionat amb els coneixements de l'usuari, i algunes de les aplicacions que pot tenir BancTrad (ap. 5.2), aspecte lligat al seu perfil professional o acadèmic.

5.1 Cerques

5.1.1 Tres nivells de cerca

La interfície web de BancTrad permet accedir els corpus sense tenir coneixements profunds de lingüística ni de la sintaxi de les expressions regulars o de CQP, però també permet aprofitar aquests coneixements en cas que es tinguin. Això s'ha aconseguit articulant tres nivells de cerca, en funció del domini que tingui l'usuari de la sintaxi de cerca i del nivell d'informació lingüística que vulgui fer servir.

a) **Nivell bàsic:** cerca per seqüències de paraules específiques (amb l'opció d'especificar-ne l'equivalència a la llengua d'arribada)

b) **Nivell intermedi:** cerca per seqüències de fins a cinc paraules, amb possibilitat de cercar per forma, lema, categoria i funció sintàctica (això últim només per al català)

A la Fig. 3 hi ha una captura d'una cerca en aquest nivell: per tal de cercar la traducció de construccions causatives del català a l'anglès, es cerca el lema *fer* seguit de qualsevol verb (v. subapartat següent per a la presentació de resultats).

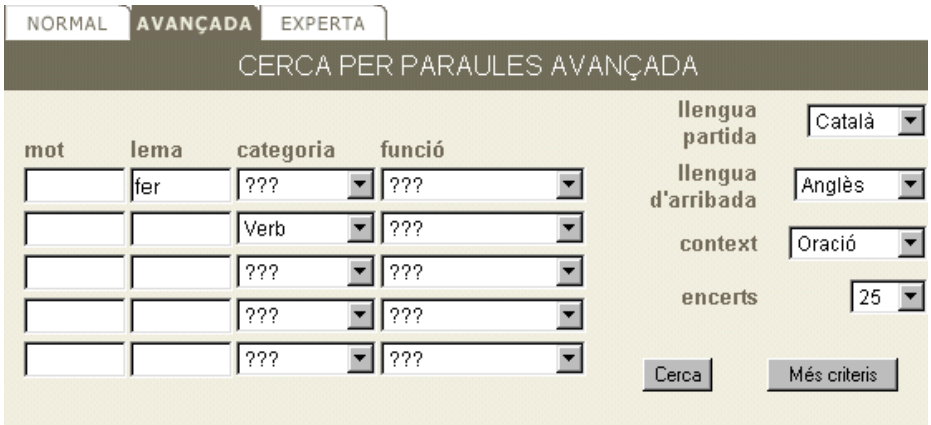

Figura 3: Cerca al nivell intermedi de BancTrad

c) **Nivell expert:** cerca en la sintaxi de CQP

5.1.2 Restriccions sobre característiques textuals i cerca de text sencer

En tots tres nivells hi ha l'opció de demanar restriccions no només sobre els mots, sinó sobre les característiques extralingüístiques dels textos explicades a l'ap. 2. Així, doncs, mitjançant l'extensió

del formulari, es pot restringir les ocurrències del mot "*banc*" a textos d'economia. Aquesta mena d'anotació possibilita una altra mena de cerques: les cerques de text sencer, per cercar textos paral·lels. La Fig. 4 exemplifica aquesta possibilitat: en aquest cas, l'usuari vol recuperar assajos sobre art en què l'original estigui en alemany i la traducció en espanyol.

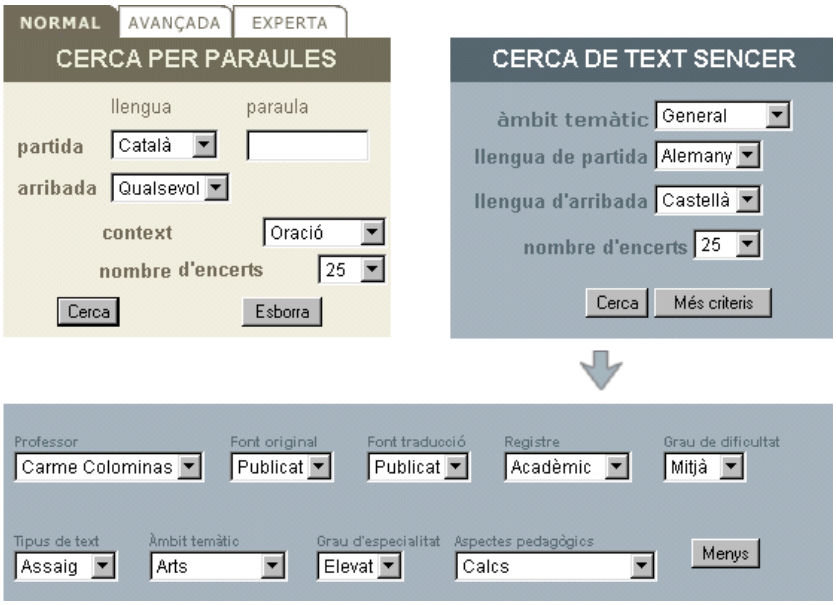

Figura 4: Cerca de text sencer a BancTrad

5.1.3 Presentació de resultats

Per defecte, els resultats es presenten en forma d'oracions alineades, tot i que està previst que l'usuari pugui determinar altres presentacions, com paràgraf sencer o un nombre determinat de paraules al voltant del mot cercat. La Fig. 5 mostra els resultats de la cerca de construccions causatives feta a l'apartat 5.1.1:

Finalment , l' any 1413 el rei Ferran I donà a la Generalitat una forma legal definitiva i esdevingué un organisme de govern , gairebé desvinculat de les Corts , autònom en la designació de els seus components , i amb funcions per **fer observar** el sistema constitucional de la Confederació .

EN: Finally , in 1413 , King Ferdinand I shaped the definitive legal form of the Generalitat ; it thus became a government body , virtually separate from the Corts , free to appoint its members , and with the authority to enforce the constitutional system of the Confederation .

La mateixa qüestió financera creà tensions amb la corona durant el regnat de Felip III (1598-1621) a causa de les contribucions que es **feien pagar** a Catalunya en profit de els interessos de la corona i que havien de ser recaptades precisament per la Generalitat .

EN: Financial problems also created conflicts with the Crown during the reign of Philip III (1598-1621) because of the taxes Catalonia was obliged to pay to the Crown . The Generalitat was , of course , charged with the collection of these taxes .

Aquests fets i les notícies sobre les actuacions de la Gran Aliança **feren esclatar** l' alçament a Catalunya a mitjan 1705 .

EN: This situation and the news of the battles undertaken by the Great Alliance led to an uprising in Catalonia in mid-1705 .

Figura 5: Resultats de la cerca exemplificada a la Figura 3

5.2 Aplicacions de BancTrad

BancTrad pot servir per diverses aplicacions. Inicialment es va pensar com a eina didàctica per als propòsits docents de la FTI, però com a mínim dues altres menes d'aplicacions s'han tingut molt en compte en procés de desenvolupament del projecte: la recerca i les aplicacions professionals.

5.2.1 Docència

Per propòsits docents (concretament, de didàctica de la traducció), tots els nivells i possibilitats de cerca esbossats a l'apartat 5.1 són potencialment adequades. La cerca de text sencer s'ha exemplificat més amunt; quant a la cerca per mots, és evident que és una possibilitat molt útil per cercar equivalències de mots en context, cosa imprescindible per a la traducció. Per exemple, es podria cercar:

a) traducció de la forma anglesa *stores* al català. Resultat: *botigues* (nom), *guarda* (verb)
b) refinant (a), es podria buscar la traducció del lema *store* al català. Resultat: *botiga*, *botigues* (nom), tot el paradigma del verb *guardar*
c) refinant (a) i (b), es podria cercar traduccions al català del lema *store* amb categoria gramatical *verb*. Resultat: tot el paradigma del verb *guardar*

Evidentment, tal com passa a les eines de cerca estàndard, es pot restringir el context d'aparició del mot que es cerca, o bé buscar combinacions de paraules o lemes. Per exemple:

d) traducció del gerundi del verb *indicate* després d'un punt
e) traducció del gerundi del verb *indicate* després d'un punt, especificant que en català no hi pot haver cap gerundi, o bé que en català no es fa servir el verb *indicar*

5.2.2 Aplicacions professionals i de recerca

Els traductors professionals poden trobar a BancTrad decisions de traducció fetes anteriorment, a més de textos paral·lels (recordem que els corpus estan constituïts per traduccions reals). Per altra banda, BancTrad pot ser útil per a investigació tant en lingüística (especialment en lingüística comparativa) com en teoria de la traducció (v. obres de Baker, M. i Teubert, W.).

Una altra possibilitat que cal tenir molt en compte és la d'utilitzar BancTrad per crear altres eines lingüístiques, com p. ex. diccionaris multilingües, chunkers, sistemes de TA, etc.

5.2.3 Un valor afegit

Finalment, volem fer notar també que un valor afegit a la interfície de BancTrad és el fet que permet incorporar altres corpus (multilingües o monolingües) amb molt poc esforç tècnic. Això permetria els usuaris de consultar no només els corpus creats expressament per a BancTrad, sinó també d'altres com el *British National Corpus* o el *Frankfurter Rundschau*, utilitzant exactament la mateixa interfície. P. ex., està previst de tenir el BNC integrat a BancTrad a l'estiu del 2002.

6 Conclusions

En aquest article hem presentat una eina que permet aprofitar els resultats obtinguts en lingüística computacional, aplicant-los un camp diferent però relacionat com és la traducció. BancTrad permet consultar corpus paral·lels que contenen informació lingüística i extralingüística. Aquesta eina integra d'altres eines, tant d'anàlisi lingüística com de processament de corpus. Es va pensar inicialment com a eina per a la didàctica de la traducció, però les seves possibilitats abarquen d'altres camps com la recerca o la traducció professional.

7 Agraïments

BancTrad ha estat finançat pel Programa d'Innovació Docent de la Universitat Pompeu Fabra (anys 2001 i 2002), així com pel Ministerio de Educación, Cultura y Deporte i el Departament d'Universitat, Recerca i Societat de la Informació de la Generalitat de Catalunya (beca 2001FI 00582).

Els autors voldrien agrair al cos docent de la FTI la seva col·laboració en el projecte, i al Stefan Bott la seva ajuda tècnica en la preparació de la interfície.

8 Bibliografia

Badia, T., À. Egea i T. Tuells. 1997 “CATMORF: Multi-two level steps for Catalan morphology”. *Demo Proceedings of the Conference on Applied Natural Language Processing*. Washington

Badia, T., Boleda, G., Bofias, E. i Quixal, M. 2001 “A modular architecture for the processing of free text”. *Proceedings of the Workshop on 'Modular Programming applied to Natural Language Processing'* at *EUROLAN 2001*. Iasi, Romania.

Christ, Oliver. 1994 “A modular and flexible architecture for an integrated corpus query system” *COMPLEX'94*. Budapest (v. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/)

Christ, Oliver, Schulze, Bruno M. i König, Esther. 1999 *Corpus Query Processor (CQP). User's Manual*, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Stuttgart

Karlsson, F. et al. 1995. *Constraint Grammar: a Language-Independent Formalism for Parsing Unrestricted Text*. Berlin/New York: Mouton De Gruyter

Schmid, Helmut. 1995 “Improvements in Part-of-Speech Tagging with an Application to German”. *Proceedings of the ACL SIGDAT-Workshop*, 47-50

Schmid, Helmut. 1997 “Probabilistic Part-of-Speech Tagging Using Decision Trees”. *New Methods in Language Processing Studies in Computational Linguistics*, Daniel Jones i Harold Somers (eds), 154-164. London: UCL Press.

Tapanainen, P. 1996 *The Constraint Grammar Parser CG-2*, Department of General Linguistics, University of Helsinki, Helsinki, Publications, 27