El corpus paral·lel del Diari Oficial de la Generalitat de Catalunya: compilació, anàlisi i exemples d'ús

Antoni Oliver (Barcelona)

Summary: In this paper the process of compilation of the parallel corpus from the Official Diary of the Catalan Government (DOGC) is presented. It describes the downloading process, the tools and processes for the treatment and linguistic analysis. The final result is a big parallel corpus that is freely available in several formats and with several annotation levels. This corpus is a very valuable resource for different applications. As example, three possible fields of application are described: as a translation memory to be used in a Computer-Assisted Translation tool; for terminology extraction and query and for training statistical machine translation systems.

Received: 17-03-2016 · Accepted: 09-06-2016

1 Introducció

1.1 Els corpus lingüístics monolingües i bilingües paral·lels

La lingüística de corpus

Els corpus lingüístics són col·leccions de mostres d'ús de les llengües emmagatzemades electrònicament (Hunston, 2006). Sinclair (1996: 27) afegeix alguns conceptes a la definició i diu:

> Un corpus és una col·lecció de fragments de la llengua que han estat seleccionats i ordenats d'acord amb un criteri lingüístic explícit amb l'objectiu de ser utilitzat com a mostra de la llengua. (Traducció de l'autor)

Aquestes mostres poden ser tant orals com escrites i poden estar formades per mostres en una sola llengua o en diverses llengües. Els exemples d'ús poden ser representatius del llenguatge general o bé centrar-se en algun tipus de comunicació especialitzada o en alguna variant de la llengua.

Els corpus bilingües o plurilingües estan formats per textos en més d'una llengua. En el cas que els textos siguin la traducció dels textos en una de les llengües, parlem de corpus *paral·lels*. Si no són la traducció, però formen part del mateix tipus de text, parlem de corpus *comparables*.

El corpus del Diari Oficial de la Generalitat de Catalunya que presentem en aquest article és un corpus paral·lel bilingüe català–castellà format per textos legislatius i administratius de la Generalitat de Catalunya.

Corpus lingüístics per al català

El català compta amb una sèrie de corpus lingüístics, tant monolingües com bilingües amb diferents nivells d'anotació. En aquest apartat exposarem els corpus textuals sincrònics més significatius, amb les seves característiques bàsiques de tipologia, accés, mida i nivell d'anotació.

Corpus Textual Informatitzat de la Llengua Catalana (CTILC)[1]

Es tracta d'un corpus consultable a través d'Internet creat per l'Institut d'Estudis Catalans, a partir d'una gran varietat de textos d'entre els anys 1832 i 1988. Presenta anotació morfosintàctica i de lema i té una mida superior a 52 milions de paraules.

AnCora-CA corpus[2]

La versió catalana del corpus AnCora (Taulé / Martí / Recasens, 2008) és un corpus descarregable amb una mida de 488.389 paraules. Disposa de diferents nivells d'anotació:

- lema i categoria morfosintàctica
- constituents i funcions sintàctiques
- estructura argumental i papers temàtics
- classe semàntica verbal
- tipus denotatiu dels noms deverbals
- sentits de WordNet nominals
- entitats nombrades
- relacions de coreferència

La part anotada sintàcticament constitueix per si mateixa el corpus CESS-CAT.

1 <http://ctilc.iec.cat>.
2 <http://clic.ub.edu/corpus>.

Wikicorpus[3]
El Wikicorpus (Reese et al., 2010) està format per les entrades de la Viki-pèdia en català corresponent a un *dump* de l'any 2006 i està enriquit amb informació morfosintàctica i semàntica, fent servir l'analitzador Freeling i UKB. La versió catalana (hi ha també una versió castellana i anglesa) té una mida aproximada de 50 milions de paraules.

Corpus paral·lel català–castellà de El Periódico[4]
Aquest corpus està format per l'alineació de 10 anys d'articles bilingües de *El Periódico de Catalunya* i conté més de 100 milions de paraules. El corpus es distribueix sota pagament i està format per text pla sense cap mena d'anotació.

Els corpus paral·lels de l'Opus Corpus[5]

Corpus	Paraules cat	Paraules spa	Paraules eng
EUBookshop cat spa	75.145	77.562	—
EUBookshop cat eng	88.487	—	82.714
GNOME cat spa	79.748	79.714	—
GNOME cat eng	182.896	—	165.406
KDE4 cat spa	1.503.907	1.397.487	—
KDE4 cat eng	1.521.353	—	1.201.719
Open Subtitles 2012 cat spa	633.836	607.223	—
Open Subtitles 2012 cat eng	552.584	—	593.275
Open Subtitles 2013 cat spa	738.997	720.565	—
Open Subtitles 2013 cat eng	818.676	—	870.086
Open Subtitles 2016 cat spa	2.473.502	2.436.822	—
Open Subtitles 2016 cat eng	2.385.132	—	2.516.392
Taoeba cat spa	15.423	15.012	—
Taoeba cat eng	6.595	—	6.775
Ubuntu cat spa	34.563	33.069	—
Ubuntu cat eng	33.404	—	27.320
DOGC cat spa (versió 2013)	162.981.769	150.435.197	—
Books cat eng	73.463	—	68.625

Taula 1. Corpus paral·lels anglès–català i anglès–castellà.

3 <http://www.cs.upc.edu/~nlp/wikicorpus>.
4 <http://catalog.elra.info/product_info.php?products_id=1122>.
5 <http://opus.lingfil.uu.se>.

La iniciativa OPUS Corpus (Tiedemann, 2012) recopila corpus paral·lels en diverses llengües i tots ells són descarregables lliurement. Per al català podem trobar principalment corpus paral·lels amb el castellà i l'anglès. A la taula 1 (pàgina anterior) podem observar el nombre de paraules per llengua de cada un dels corpus disponibles.

Com podem veure, ja hi ha la versió prèvia del corpus DOGC, que comprenia els anys de 1998 al 2013 i que ha estat recopilat pel mateix autor d'aquest article.

La versió 2013 del Corpus DOGC ja era el corpus paral·lel del català de major mida amb més de 150 milions de paraules per llengua. A la taula 4 de l'apartat 2.8 podem observar la mida de la nova versió del corpus.

1.2 El Diari Oficial de la Generalitat de Catalunya

El Diari Oficial de la Generalitat de Catalunya (DOGC) és el mitjà de publicació oficial de les lleis, normes, acords, resolucions, edictes, notificacions i anuncis de l'Administració i del Govern de Catalunya.

El DOGC apareixia originàriament en paper i des del 2007 es va substituir íntegrament per una versió electrònica d'accés lliure.[6] Els textos provinents dels documents del DOGC tenen una llicència lliure i es poden distribuir i processar sense cap limitació legal. Molts dels textos del DOGC apareixen publicats tant en català com en castellà.

2 Descàrrega i processament del corpus

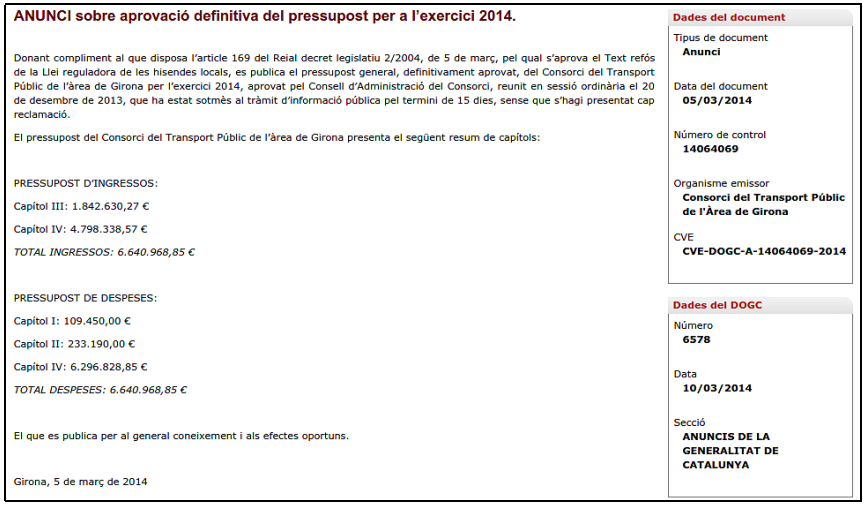
En aquesta secció expliquem les eines, tècniques, dificultats i solucions que hem emprat per descarregar tots els articles de tots els números del DOGC des del 1977 fins el 2015 i processar-los fins a obtenir un corpus paral·lel català–castellà amb el contingut publicat en les dues llengües. També descrivim els formats finals en els que es pot obtenir aquest corpus. Tot el procés de descàrrega i processament s'ha dut a terme mitjançant programes desenvolupats en llenguatge de programació Python[7] i fent servir programari lliure. A la taula 2 es pot observar les diferents etapes del processament del corpus i el resultat que s'obté de cada etapa.

6 Es pot accedir al DOGC des de l'enllaç <http://dogc.gencat.cat>.
7 <http://www.python.org>.

Procés	Resultat
Descàrrega	Corpus html
Classificació automàtica per anys	Corpus html per a cada any
Conversió de html a text	Corpus text per a cada any
Verificació de llengua	
Segmentació dels textos amb llengua verificada	Corpus segmentat per a cada any
Alineació	
Neteja	Corpus paral·lel en format de text tabulat
Eliminació segments repetits	Corpus paral·lel en format de text tabulat Corpus paral·lel en format Moses Memòria de traducció en format TMX

Taula 2. Etapes del processament del corpus i resultat de cada etapa.

2.1 Descàrrega del corpus



ANUNCI sobre aprovació definitiva del pressupost per a l'exercici 2014.

Donant compliment al que disposa l'article 169 del Reial decret legislatiu 2/2004, de 5 de març, pel qual s'aprova el Text refós de la Llei reguladora de les hisendes locals, es publica el pressupost general, definitivament aprovat, del Consorci del Transport Públic de l'àrea de Girona per l'exercici 2014, aprovat pel Consell d'Administració del Consorci, reunit en sessió ordinària el 20 de desembre de 2013, que ha estat sotmès al tràmit d'informació pública pel termini de 15 dies, sense que s'hagi presentat cap reclamació.

El pressupost del Consorci del Transport Públic de l'àrea de Girona presenta el següent resum de capítols:

PRESSUPOST D'INGRESSOS:

Capítol III: 1.842.630,27 €

Capítol IV: 4.798.338,57 €

TOTAL INGRESSOS: 6.640.968,85 €

PRESSUPOST DE DESPESES:

Capítol I: 109.450,00 €

Capítol II: 233.190,00 €

Capítol IV: 6.296.828,85 €

TOTAL DESPESES: 6.640.968,85 €

El que es publica per al general coneixement i als efectes oportuns.

Girona, 5 de març de 2014

Dades del document

Tipus de document
Anunci

Data del document
05/03/2014

Número de control
14064069

Organisme emissor
Consorci del Transport Públic de l'Àrea de Girona

CVE
CVE-DOGC-A-14064069-2014

Dades del DOGC

Número
6578

Data
10/03/2014

Secció
ANUNCIS DE LA GENERALITAT DE CATALUNYA

Figura 1. Exemple de plana del DOGC en català.

Tots els documents del DOGC s'identifiquen per un identificador (*documentId*) i una llengua (*language*) que pot ser català (*ca_ES*) o castellà (*es_ES*). Per exemple, el següent enllaç:

http://dogc.gencat.cat/ca/pdogc_canals_interns/pdogc_resultats_fitxa/?action=fitxa&documentId=**656789**&language=**ca_ES**

ens porta a la plana que podem observar a la figura 1 (pàgina anterior). Canviant el paràmetre *language* a *es_ES* i modificant l'enllaç com es mostra a continuació:

http://dogc.gencat.cat/es/pdogc_canals_interns/pdogc_resultats_fitxa/index.html?action=fitxa&documentId=**656789**&language=es_ES&newLang=es_ES

que ens porta al mateix document, però corresponent a la versió castellana.

Sobre la publicació dels articles cal tenir en compte alguns aspectes que fan que no sempre sigui possible obtenir tot el text en les dues llengües, i fins i tot que no sigui possible obtenir-lo ni tant sols en una de les llengües:

- En alguns casos, especialment en números anteriors al 2007, en la web només apareix el títol de l'article, però el cos de l'article es presenta com a un enllaç a un pdf corresponent al escanneig de la plana del DOGC on apareix l'article.
- No tots els articles estan en les dues llengües. Hi ha una sèrie d'articles que estan en català tant en la versió catalana com en la castellana, o que estan en castellà en les dues versions. En aquests casos l'únic contingut que està en les dues llengües és el títol de l'article.

Per obtenir tot el DOGC des del 1977 fins al 2015 s'han descarregat tots els arxius html corresponents a documents amb identificadors des de l'1 fins al 715000 (que ja inclou documents corresponents al 2016).

2.2 Classificació automàtica per anys

Si ens fixem en la figura 1, veurem dos requadres a la dreta de la plana. El requadre inferior conté el número de DOGC a que es correspon l'article, la data i la secció. Aprofitarem el camp data per classificar automàticament per anys els arxius html descarregats. Cal tenir en compte que la relació any–número de document no és purament incremental i podem trobar documents d'un determinat any amb número de document superior a alguns d'anys posteriors.

2.3 Conversió d'html a text

En aquest pas pretenem extreure de l'arxiu html el text corresponent a l'article del DOGC, sense extreure altres elements textuals comuns, com opcions de menú, enllaços, etc. La principal dificultat d'aquest pas és que l'estructura dels arxius html han anat canviant amb el temps i s'han de desenvolupar estratègies d'extracció de text lleugerament diferents. També s'ha hagut de tenir en compte que en alguns períodes de temps els textos presentaven salts de línia manuals per justificar el text en pantalla. Aquests salts s'han hagut d'eliminar per poder dur a terme el pas de segmentació del text, que expliquem en el següent apartat.

La conversió de html a text es fa amb un algorisme propi basat en el paquet de Beautiful Soup[8] de Python.

2.4 Verificació de llengua

Com ja hem comentat, no tots els documents del DOGC estan en les dues llengües i podem trobar-nos documents en castellà en la versió catalana del DOGC i a la inversa. Per evitar alinear documents en la mateixa llengua hem dut un procés de detecció automàtica de llengua sobre tots els arxius en format text. Hem fet servir la llibreria de Python anomenada *langdetect*, que és una adaptació directa de la llibreria en Java *language detection*.[9] Aquesta detecció automàtica ens permet separar els documents que estan en una llengua que no correspon. La resta de passos els farem sobre els documents amb la llengua verificada.

2.5 Segmentació

Quan extraiem el text dels arxius html descarregats obtenim documents que estan organitzats en paràgrafs. Per a obtenir el corpus paral·lel ens interessa tenir el corpus segmentat, és a dir, dividit en segments de mida aproximada a una oració. Aquesta segmentació es du a terme identificant els caràcters que habitualment constitueixen un final d'oració, com el punt (.), l'interrogant (?) i l'exclamació (!). Ara bé, el procés de segmentació presenta dificultats ja que els punts es fan servir també en les abreviatures i les sigles. S'ha desenvolupat un algorisme de segmentació propi basat en

8 <http://www.crummy.com/software/BeautifulSoup>.
9 <https://code.google.com/p/language-detection>.

expressions regulars i que té en compte tant les abreviatures habituals en català i castellà, com les sigles més habituals emprades en el DOGC, que s'han extret prèviament mitjançant un algorisme desenvolupat especialment per a aquesta tasca.

2.6 Alineació

Un cop segmentat els textos tant en català com en castellà es du a terme un procés d'alineació automàtica. L'alineació consisteix a relacionar els segments d'un determinat arxiu en català amb els corresponents segments l'arxiu corresponent en castellà. La principal dificultat d'aquesta tasca és que no sempre la relació entre segments en català i segments en castellà és 1 a 1. Pot passar que a un determinat segment en català li corresponguin dos segments en castellà (relació 1 a 2), o a l'inversa, que dos segments en català es corresponguin a un sol segment en castellà (relació 2 a 1). En alguns casos pot ser que un determinat segment en català no tingui correspondència en castellà (relació 1 a 0) o que apareguin nous segments en castellà (relació 0 a 1).

Hi ha dos factors que faciliten la tasca d'alineació en el corpus que estem creant:

- Les dues llengües són properes i les traduccions acostumen a tenir una estructura similar pel que fa a segments.
- Les traduccions dels textos del DOGC acostumen a ser molt fidels i acostumen a mantenir l'estructura pel que fa al nombre de segments.

Així doncs, la majoria de casos de discordança en l'estructura en segments del català i el castellà no han estat provocats per l'estil de la traducció, sinó per discrepàncies en el procés de segmentació automàtica.

Per dur a terme l'alineació automàtica s'ha fet servir Hunalign (Varga et al., 2005). Hunalign funciona millor si es proporciona un diccionari bilingüe entre les dues llengües de treball. En el nostre procés d'alineació hem fet servir un diccionari creat a partir del diccionari de transferència català–castellà del sistema de traducció automàtica Apertium (Forcada, 2010).

2.7 Neteja

Un cop alineats els arxius s'ha fet un procés de neteja per eliminar certs segments problemàtics. Concretament, s'han dut a terme les següents accions:

- S'ha fet una normalització del caràcter corresponent a l'apòstrof, ja que alguns documents contenien accents tancats o altres caràcters.

- S'han eliminat tots els segments que contenen únicament xifres o altres símbols però que no contenen cap paraula (considerant com a paraula qualsevol cadena de cinc o més lletres).
- S'han eliminat els segments massa llargs, ja que amb molta probabilitat provenen d'errors de segmentació. Per determinar què vol dir un segment massa llarg s'ha calculat la mida mitjana en caràcters dels segments i la desviació estàndard i s'han eliminat tots els segments amb una mida superior a la mitjana més dues vegades la desviació estàndard. Donat que la mida en caràcters pot variar entre el català i el castellà, el càlcul s'ha dut a terme de manera separada per a cada una de les llengües i s'ha eliminat el segment en les dues llengües si s'acomplia el criteri almenys en una llengua. A la taula 3 podem observar aquests valors.

Mitjana cat	Desviació Estàndard cat	Eliminar cat	Mitjana spa	Desviació Estàndard spa	Eliminar spa
233.58	212.98	1299	239.77	219.25	1336

Taula 3. Dades estadístiques de la mida en caràcters dels segments alineats.

A partir d'aquesta darrera versió s'eliminen tots els segments repetits i es presenten els resultats endreçats alfabèticament.

D'aquesta manera tenim dues versions disponibles del corpus:

- Completa: inclou tots els segments resultants de la neteja descrita anteriorment. Es manté la informació de número de document i any de publicació,
- Sense repeticions (únic): S'han endreçat els segments alfabèticament i s'han eliminat els segments repetits. No es manté la informació de número de document ni l'any de publicació. Aquesta versió pot ser útil per a la creació de memòries de traducció i models estadístics de traducció, així com per a tasques d'extracció automàtica de terminologia.

2.8 Arxius presents a la distribució

El corpus es pot descarregar des de la plana web del *Language Processing Group* de la Universitat Oberta de Catalunya.[10] La distribució consta de diversos arxius i directoris que contenen els arxius html, els arxius de text

10 <http://lpg.uoc.edu/corpus-DOGC>.

resultant de la conversió de html a text, els arxius segmentats i els arxius alineats en diversos formats, tant una versió global com a una per a cada any. Presentem a continuació els arxius presents a cada directori i algunes estadístiques de nombre d'arxius, segments i paraules. Tots els arxius es presenten comprimits en format zip.

Directori principal:
En aquest directori presentem els arxius alineats conjunts de tots els anys en diferents formats:

- **DOGC-info-cat-spa.txt.zip**: Arxiu paral·lel separat per tabuladors que conté els segments catalans alineats amb els castellans amb informació de número de document de procedència, any de publicació i el valor de fiabilitat de l'alineació donat per Hunalign.
- **DOGC-cat-spa.txt.zip**: Arxiu paral·lel separat per tabuladors que conté els segments catalans alineats amb els castellans.
- **DOGC-unic-cat-spa.txt**: Arxiu paral·lel separat per tabuladors que conté els segments catalans alineats amb els castellans endreçats alfabèticament i sense repeticions.
- **DOGC-unic-cat-spa.tmx.zip**: Les alineacions sense repeticions en el format estàndard per a l'intercanvi de memòries de traducció (TMX: *Translation Memory eXchange*).
- **DOGC-unic.ca-es.ca.zip** i **DOGC-unic.ca-es.es.zip**: Alineacions sense repeticions en format Moses (Koehn et al., 2007), és a dir les alineacions sense repeticions en dos fitxers de text separats, un per al català i altre per al castellà.

A la taula 4 podem observar el nombre de segments i paraules en castellà i català del corpus sencer i del corpus sense repeticions.

	Segments	Paraules català	Paraules castellà
TOT	8.074.284	188.908.522	197.991.183
UNIC	5.026.847	142.502.123	149.339.268

Taula 4. Mida del corpus en la seva versió completa
com en la seva versió sense segments repetits.

Si comparem la mida de la versió actual amb la de la versió 2013 observem que tenim un increment del 15.9% en paraules en català. Aquest augment podria haver estat major, ja que hem dut a terme un procés de neteja

que ha fet esborrar una sèrie de segments. Gràcies a aquest procés de neteja, però, obtenim un corpus paral·lel de molta més qualitat.

Directori anys:
Es presenten els mateixos arxius del directori principal però separats per anys. Fixem-nos que hi ha arxius des de l'any 1998 fins el 2015. Tot i que s'han descarregat tots els documents del DOGC alguns d'ells no estaven en les dues llengües i alguns altres estaven com a enllaços a arxius PDFs resultants de l'escanejat del diari en paper. Per aquest motiu no hi ha arxius paral·lels des del 1977 fins el 1997.

Directori html:
Es presenten tots els arxius html descarregats distribuïts per anys. En aquest directori hi ha arxius des de l'any 1977 fins al 2015. Els arxius estan classificats per anys i llengua en comprimits en arxius zip. Cal tenir en compte que la llengua és la corresponent a la del document publicat en la versió de DOGC en aquesta llengua. Com que alguns articles en la versió castellana apareixen en català i alguns de la catalana en castellà, la classificació per llengües no sempre correspon a la llengua real en què està escrit en document.

Directori txt:
Es presenten els arxius resultants de l'extracció del text dels html descarregats, també distribuïts per anys.

Directori segment:
Es presenten els arxius de text segmentats distribuïts per anys.

3 Anàlisi lingüística

El corpus del DOGC es distribueix també analitzat a nivell morfosintàctic i a nivell semàntic mitjançant una anàlisi automàtica amb Freeling (Padró / Stanilovsky, 2012) i UKB (Padró et al., 2010). El resultat de l'anàlisi del segment català:

> S'ha intentat notificar personalment, però no s'ha pogut dur a terme.

seria la següent:

> S'|es|P0300000|0.999814|- ha|haver|VAIP3S0|0.999848| 02655135-v:0.0128289/ 02603699-v:0.0125509 intentat|intentar|VMP00SM|1|02530167-v:0.0238802 notificar|notificar|VMN0000|1|00873682-v:0.00881258/00870213-v:0.008339/

00873469-v:0.00807487 personalment|personalment|RG|1|-|,|,|Fc|1|-|però|però|CC|0.894616|-|no|no|RN|0.982564|-|s'|es|P0300000|0.999814|-|ha|haver|VAIP3S0|0.999848|02655135-v:0.0128289/02603699-v:0.0125509 pogut|poder|VMP00SM|1|-|dur_a_terme|dur_a_terme|VMN0000|0.648839|01712704-v:0.0064428/02561995-v:0.00630817/02568672-v:0.00622783/02560767-v:0.00602251 .|.|Fp|1|-

i la del corresponent segment castellà:

Se ha intentado notificar personalmente, pero no se ha podido llevar a cabo.

seria:

Se|se|P00CN000|0.465639|-|ha|haber|VAIP3S0|0.999255|02655135-v:0.0128315/02603699-v:0.0125436 intentado|intentar|VMP00SM|1|02530167-v:0.0239024 notificar|notificar|VMN0000|1|01438681-v:0.0230758 personalmente|personalmente|RG|1|00366266-r:0.00782766/00366393-r:0.00724163/00132322-r:0.00722484 ,|,|Fc|1|-|pero|pero|CC|0.999764|-|no|no|RN|0.998134|-|se|se|P00CN000|0.465639|-|ha|haber|VAIP3S0|0.999255|02655135-v:0.0128315/02603699-v:0.0125436 podido|poder|VMP00SM|1|02402825-v:0.0275664 llevar_a_cabo|llevar_a_cabo|VMN0000|1|02568672-v:0.00830142/00251463-v:0.0081989/01641545-v:0.00721123 .|.|Fp|1|-

En les anàlisis cada *token* es separa per espai i dins de cada *token* tenim la següent informació:

- forma
- lema
- etiqueta morfosintàctica
- probabilitat de l'etiqueta morfosintàctica
- tots els possibles sentits de la paraula expressats com a *synset* de WordNet i la seva probabilitat endreçats de més a menys probables.

3.1 Anàlisi morfosintàctica

Si ens fixem en les anàlisis anteriors veurem que l'anàlisi morfosintàctica s'expressa mitjançant una sèrie d'etiquetes basades en les proposades pel grup EAGLES.[11] Les etiquetes emprades per al català[12] i per al castellà[13] són pràcticament idèntiques. A la taula 5 podem observar les categories principals d'aquests etiquetaris.

11 <http://www.ilc.cnr.it/EAGLES96/home.html>.
12 <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-ca.html>.
13 <http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html>.

Categoria	Etiqueta
Adjectius	A
Adverbis	R
Determinats	D
Noms	N
Verbs	V
Pronoms	P
Conjuncions	C
Interjeccions	I
Preposicions	S
Puntuació	F
Numeral	Z
Data i hora	W

Taula 5. Categories principals dels etiquetaris emprats per al català i castellà.

Freeling du a terme una desambiguació estadística i per a cada etiqueta ens proporciona també una probabilitat. Si ens fixem en l'anàlisi catalana les paraules *intentar*, *notificar*, *personalment* i *pogut* tenen una probabilitat d'1, perquè no presenten ambigüitat morfosintàctica. En canvi, la resta de paraules tenen una probabilitat menor d'1 perquè són paraules ambigües des del punt de vista morfosintàctic.

3.2 Anàlisi semàntica

L'anàlisi semàntica duta a terme amb el mòdul d'UKB de Freeling proporciona els sentits de les paraules mitjançant *synsets* de WordNet. L'algorisme proporciona tots els possibles sentits de cada paraula de categoria oberta endreçats per probabilitat. WordNet (Fellbaum, 1998) és una base de dades de coneixement lèxic on les paraules de les categories obertes (substantius, verbs, adjectius i adverbis) s'organitzen en conjunts de sinònims que reben el nom de *synsets*. Cada *synset* representa un concepte lexicalitzat en anglès, i es connecta amb els altres *synsets* mitjançant relacions semàntiques. Els *synsets* s'expressen mitjançant un número seguit d'un guió i una categoria gramatical. Així, si seguim l'exemple d'anàlisi del català, la paraula *notificar* té un únic sentit expressat pel *synset* 00873682-v. Aquest *synset* té una definició en anglès que és: *inform (somebody) of something* i té una sèrie de sinònims en català: *advertir*, *avisar*, *informar* i *notificar*.

4 Exemples d'ús

4.1 Com a memòria de traducció

Un ús directe d'aquest corpus paral·lel és fer-lo servir com a memòria de traducció en una eina de traducció assistida. D'aquesta manera, quan traduïm un nou segment amb l'eina, ens apareixeran els segments més semblants del corpus amb la corresponent traducció. A la figura 2 podem observar l'ús del corpus com a memòria a l'eina de traducció assistida OmegaT. En la pantalla superior tenim el segment que hem de traduir (*El termini màxim per fer les actuacions esmentades és de 12 mesos des de la data d'aquesta resolució*) i en la pantalla inferior els segments més similars al que estem traduint i les traduccions que es troben a la memòria de traducció. En aquest cas el segment més similar és

> El termini per dur a terme les instal·lacions i la seva posada en servei serà de 2 anys, comptats des de la data de publicació d'aquesta Resolució.

En aquest exemple la similitud no és suficient per incorporar directament la traducció i fer modificacions, però en molts casos podem trobar segments iguals o molt semblants.

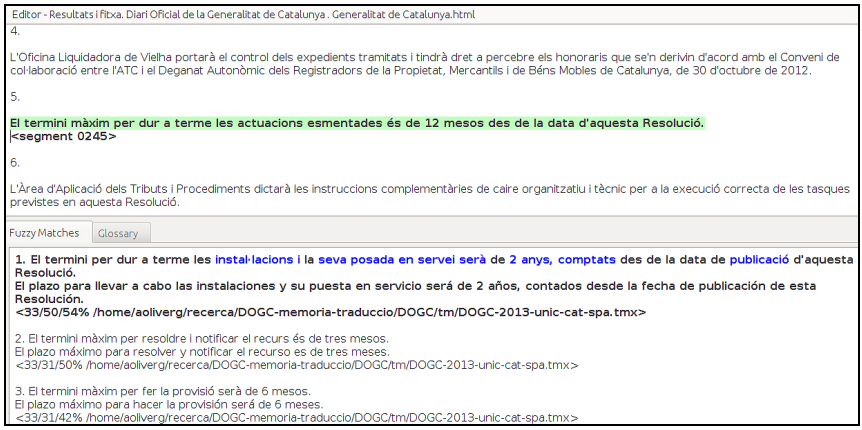

Editor - Resultats i fitxa. Diari Oficial de la Generalitat de Catalunya . Generalitat de Catalunya.html
4.
L'Oficina Liquidadora de Vielha portarà el control dels expedients tramitats i tindrà dret a percebre els honoraris que se'n derivin d'acord amb el Conveni de col·laboració entre l'ATC i el Deganat Autonòmic dels Registradors de la Propietat, Mercantils i de Béns Mobles de Catalunya, de 30 d'octubre de 2012.
5.
El termini màxim per dur a terme les actuacions esmentades és de 12 mesos des de la data d'aquesta Resolució.
|<segment 0245>
6.
L'Àrea d'Aplicació dels Tributs i Procediments dictarà les instruccions complementàries de caire organitzatiu i tècnic per a la execució correcta de les tasques previstes en aquesta Resolució.
Fuzzy Matches
Glossary
1. El termini per dur a terme les instal·lacions i la seva posada en servei serà de 2 anys, comptats des de la data de publicació d'aquesta Resolució.
El plazo para llevar a cabo las instalaciones y su puesta en servicio será de 2 años, contados desde la fecha de publicación de esta Resolución.
<33/50/54% /home/aoliverg/recerca/DOGC-memoria-traduccio/DOGC/tm/DOGC-2013-unic-cat-spa.tmx>
2. El termini màxim per resoldre i notificar el recurs és de tres mesos.
El plazo máximo para resolver y notificar el recurso es de tres meses.
<33/31/50% /home/aoliverg/recerca/DOGC-memoria-traduccio/DOGC/tm/DOGC-2013-unic-cat-spa.tmx>
3. El termini màxim per fer la provisió serà de 6 mesos.
El plazo máximo para hacer la provisión será de 6 meses.
<33/31/42% /home/aoliverg/recerca/DOGC-memoria-traduccio/DOGC/tm/DOGC-2013-unic-cat-spa.tmx>

Figura 2. El corpus DOGC com a memòria de traducció amb l'eina OmegaT.

L'ús d'aquest corpus com a memòria de traducció és directe, ja que es distribueix directament en format TMX (*Translation Memory eXchange*), que és el format estàndard admès a la immensa majoria d'eines de traducció assistida.

4.2 Consulta i extracció automàtica de terminologia

El corpus DOGC recull en els seus textos una gran quantitat de termes especialitzats en l'àmbit administratiu i legislatiu, tant en català com en castellà. Per aquest motiu el corpus es pot fer servir en un gran ventall de recerques terminològiques.

Candidats català		Candidats castellà	
3663	Generalitat de Catalunya	3088	régimen jurídico
3103	règim jurídic	3007	administraciones públicas
3066	terme municipal	2614	término municipal
3001	administracions públiques	2535	DOGC núm
2695	DOGC núm	2238	Generalidad de Cataluña
2011	procediment administratiu comú	1987	día siguiente
1957	decret legislatiu	1958	procedimiento administrativo común
1741	direcció general	1921	decreto legislativo
1679	Comissió Territorial	1733	dirección general
1661	dur a terme	1672	comisión territorial
1566	text refós	1614	información pública
1446	contingut íntegre	1529	texto refundido
1375	persones interessades	1390	contenido íntegro
1283	serveis territorials	1264	servicios territoriales
1250	diari oficial	1226	real decreto
1237	reial decret	1177	Entidad adjudicadora
1184	objecte del contracte	1177	objeto del contrato
1161	Entitat adjudicadora	1164	Generalitat de Catalunya
1146	llocs de treball	1096	puestos de trabajo
1045	públiques de Catalunya	1087	Departamento de Agricultura
1039	BOE núm	1036	Instituto Catalán
977	Territori i Sostenibilitat	985	tramita el expediente
942	dependència que tramita	970	expedientes sancionadores
848	informació pública	932	Dependencia que tramita
799	Departament de Territori	920	BOE núm

Taula 6. Llista dels 25 candidats a terme més freqüents per al català i castellà per al subcorpus corresponent a l'any 2015 fent servir estratègia estadística.

Els corpus paral·lels poden veure's també com a dos corpus monolingües. A partir del corpus en una de les llengües i aplicant tècniques d'extracció automàtica de terminologia podem extreure una llista de candidats a terme. L'eina TBXTools (Oliver, 2014) permet dur a terme aquesta tasca fent servir les dues principals estratègies:

- estratègia estadística
- estratègia lingüística

En l'estratègia estadística es calculen n-grames, es a dir, combinacions d'n paraules presents en el corpus. Posteriorment aquests n-grames es filtren per una llista de paraules buides, és a dir, paraules que no poden estar a ni a l'inici ni al final d'un terme. La llista d'n-grames resultants és la llista de candidats a terme que el terminòleg haurà de validar. Aquesta llista es presenta endreçada per freqüència d'aparició, ja que els candidats més freqüents tenen més probabilitat de ser termes rellevants. A la taula 6 (pàgina anterior) podem observar la llista de candidats a terme per al català i castellà fent servir el subcorpus corresponent a l'any 2015 i calculant bigrams (n=2) i trigrams (n=3).

En l'estratègia lingüística es cerquen certs patrons d'etiquetes morfosintàctiques que són típicament terminològiques. Per al català i castellà, com a exemples, tenim aquests patrons típicament terminològics:

Nom Nom
Adjectiu Nom
Nom Preposició Nom

Per a poder cercar aquests patrons necessitem disposar del corpus etiquetat morfosintàcticament. Aprofitem la versió etiquetada del DOGC i fem servir l'eina TBXTools, obtenint els resultats que es poden observar a la taula 7 (pàgina següent).

Però també podem fer servir el corpus paral·lel per a dur a termes tasques relacionades amb les dues llengües. Un exemple de tasca concreta seria consultar com es diu en castellà un determinat terme en català, per exemple, *recurs d'alçada*. Amb una simple cerca dels segments castellans corresponents a segments catalans que contenen el terme a cercar, podrà determinar-se de forma manual quina és la denominació castellana, Veiem-ho a continuació, on presentem 5 segments castellans dels 46.416 segment que contenen el terme cercat.

Candidats català		Candidats castellà	
2761	dogc número	2512	dogc núm.
1453	lloc de treball	1373	puesto de trabajo
1329	decret legislatiu	905	boe núm.
1280	persona interessada	894	número de expediente
1117	reial decret	738	contrato de servicios
1031	boe número	733	anuncio de notificación
875	administració_pública de Catalunya	692	entidad adjudicador
		628	convocatoria de provisión
835	número d' expedient	600	presente anuncio
739	anunci de notificació	597	anuncio de licitación
738	contracte de serveis	597	provisión núm.
631	convocatòria de provisió	585	concesión de subvenciones
614	anunci de licitació	572	declaración de impacto
601	provisió número	542	provisión de puestos
582	educatiu privat	522	forma de adjudicación
577	declaració d' impacte	496	anuncio de información
570	concessió de subvencions	494	cuerpo de Mozos_de_Escuadra
552	proposta de resolució		
544	provisió de llocs	485	convocatoria de concurso
528	forma d' adjudicació	481	real decreto
493	cos de Mossos_d'_Esquadra	460	cataluña en_materia_de función
492	anunci d' informació	454	propuesta de resolución
479	convocatòria de concurs	442	fecha de publicación
460	catalunya en_matèria_de funció	430	plazo de ejecución
458	data de publicació	414	informe propuesta
437	termini d' execució	398	plan de ordenación

Taula 7. Llista dels 25 candidats a terme més freqüents per al català i castellà per al subcorpus corresponent a l'any 2015 fent servir estratègia lingüística.

10.4 Las personas candidatas podrán presentar **recurso de alzada** contra la resolución ...
El **recurso de alzada** se tiene que presentar en el centro o sede donde actúe ...
EDICTO de 21 de octubre de 2015, de notificación de la Resolución dictada en el expediente núm. 298/2015 relativo a un **recurso de alzada** interpuesto contra el acuerdo ...
La Dirección General de Administración Local ha dictado Resolución en el expediente número 298/2015 del **recurso de alzada** interpuesto por la señora ...
¿Contra estas resoluciones, que no agotan la vía administrativa, se puede interponer **recurso de alzada**, de conformidad con lo que prevén los artículos 76 ...
EDICTO de 23 de febrero de 2015, por el que se notifica la resolución del **recurso de alzada** interpuesto contra la resolución de un expediente sancionador...

Aquesta mateixa cerca també es pot dur a terme amb algunes eines d'extracció automàtica de terminologia, per exemple TBXTools.

La cerca es pot fer en la direcció contrària, és a dir, a partir de termes castellans obtenir els equivalents catalans. Aprofitant aquesta possibilitat, la Universitat Oberta de Catalunya i el TermCat han iniciat un projecte de col·laboració amb la idea de desenvolupar una part de la base de dades terminològica IATE (*InterActive Terminology for Europe*)[14] per al català. Concretament tenim intenció de treballar amb les següents àrees temàtiques:

4. POLITICS
8. INTERNATIONAL RELATIONS
10. EUROPEAN COMMUNITIES
12. LAW

Per exemple, el terme amb codi IATE 3550697 té la denominació anglesa *scope of the register* i la denominació castellana *ámbito de aplicación*. Aquesta denominació castellana apareix 4.861 al corpus DOGC i podrem determinar que la denominació catalana és *àmbit d'aplicació*. A més, disposarem de les denominacions en la resta de llengües presents al IATE. A la taula 8 podem observar el nombre de termes amb denominació anglesa i castellana per a cada una d'aquestes àrees. També podem veure quants dels corresponents termes en castellà apareixen en el corpus DOGC amb una freqüència d'aparició de 5 o superior i que són susceptibles de ser detectats automàticament. També s'indica el nombre de termes que hi apareixen almenys una vegada.

Àrea	**Termes IATE eng i spa**	**Termes en DOGC freq>=5**	**Termes en DOGC freq>=1**
4. POLITICS	21.232	4.332	5.713
8. INTERNATIONAL RELATIONS	11.291	1.377	1.830
10. EUROPEAN COMMUNITIES	21.148	3.300	4.470
12. LAW	17.747	3.671	5.089

Taula 8. Nombre de termes amb denominació anglesa i castellana de les àrees seleccionades del IATE.

14 <http://iate.europa.eu>.

En el procés de creació del IATE en català fent servir el corpus DOGC un terminòleg revisarà que totes les denominacions catalanes siguin correctes. Un cop trobada la denominació catalana, com que estarà relacionada amb la castellana i per tant també amb l'anglesa, es podran obtenir també la relació amb les denominacions en la resta de llengües presents al IATE.

4.3 Entrenament de sistemes de traducció automàtica

Els corpus paral·lels es poden fer servir directament per a l'entrenament de sistemes de traducció automàtica estadístics. Aquest tipus de sistemes funcionen a partir de càlculs de probabilitats i treballen principalment amb dues probabilitats:

- la probabilitat de què una oració en la llengua d'arribada sigui la traducció de una determinada oració en la llengua de partida.
- la probabilitat de què una oració en la llengua d'arribada sigui una oració gramatical en aquesta llengua.

Aquestes dues probabilitats es poden calcular a partir de corpus paral·lels. Per provar la utilitat del corpus que hem compilat, hem entrenat un sistema de traducció automàtica estadística bàsic fent servir Moses (Koehn et al., 2007). Les característiques de l'entrenament són:

- Corpus corresponent a l'any 2012
- Model de llengua: 5-grames
- Tunejat del sistema a partir de 1000 segments agafats aleatòriament provinents d'articles del DOGC de 2015
- L'avaluació del sistema s'ha dut a terme fent servir 1000 segments presos aleatòriament (diferents dels anteriors) també provinents d'articles del DOGC de 2015

Sistema	**Bleu**
Moses (entrenat amb corpus DOGC)	90.46
Apertium	79.99
Google Translate	82.28
Microsoft Bing Translator	39.53

Taula 9. Comparació de la qualitat mesurada amb l'índex Bleu per al sistema Moses entrenat amb el corpus DOGC i altres sistemes de traducció automàtica.

Amb aquesta configuració hem assolit un valor de Bleu (Papineni et al., 2002) de 90.46. Per tenir una idea de la qualitat del sistema a la taula 9 (pàgina anterior) la comparem amb els valors assolits per tres coneguts sistemes de TA per a la mateixa tasca. Com podem observar, el sistema entrenat amb el corpus DOGC corresponent a l'any 2012 obté els millors resultats amb una diferència notable amb el segon, Google Translate. Cal tenir en compte però, que el nostre sistema ha estat entrenat amb textos del mateix àmbit que els que s'han fet servir per a l'avaluació (tot i que no els mateixos textos). Així, com a conclusió, fent servir un fragment del corpus DOGC per entrenar un sistema de traducció automàtica estadística molt bàsic obtenim un sistema molt adequat per a traduir textos d'aquest mateix àmbit.

A continuació podem veure el resultat de traduir el segment

> L'Àrea d'Aplicació dels Tributs i Procediments dictarà les instruccions complementàries de caire organitzatiu i tècnic per a la execució correcta de les tasques previstes en aquesta Resolució.

És el següent:

> El Área de Aplicación de los Tributos y Procedimientos dictará las instrucciones complementarias de carácter organizativo y técnico para la correcta ejecución de las tareas previstas en esta Resolución.

En un treball futur tenim la intenció de fer servir el corpus DOGC per desenvolupar un sistema de traducció automàtica híbrid, és a dir, fent servir l'estratègia de transferència sintàctica i estadística. La idea és combinar Apertium (Corbí-Bellot et al., 2005) amb Moses. El sistema resultant funcionaria segons l'esquema bàsic presentat a la figura 3.

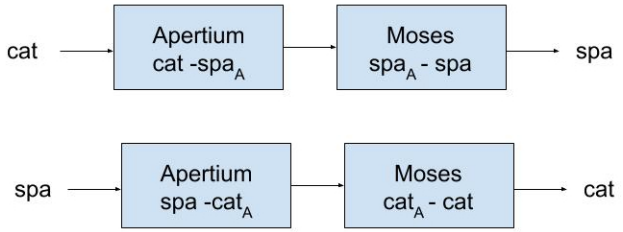


Figura 3. Esquema de funcionament del sistema de traducció automàtica híbrid Apertium–Moses.

A partir d'un text d'entrada en català, per exemple, s'obtindria una primera traducció automàtica amb Apertium (que a l'esquema anomenem spa_A). Aquesta primera traducció automàtica es passaria per un sistema Moses per a obtenir la traducció final. Aquest sistema Moses estaria entrenat de la següent manera: A partir del corpus DOGC català castellà es traduiria tota la part catalana per a obtenir un corpus paral·lel spa_A (provinent de la traducció automàtica amb Apertium)–spa (amb les traduccions reals al castellà). Es faria el mateix procés per a la direcció castellà–català. S'espera que amb aquesta hibridació el sistema de traducció automàtica assoleixi una més alta qualitat que qualsevol dels dos sistemes per separat.

5 Obtenció del corpus

El corpus DOGC es distribueix lliurement i es pot descarregar de la plana web del grup de recerca *Language Processing Group* de la Universitat Oberta de Catalunya.[15]

Els textos del DOGC es poden distribuir i reutilitzar sense cap mena de limitació amb l'obligació de citar la font, de no alterar ni desnaturalitzar la informació i especificar la data de la darrera actualització.[16]

6 Conclusions

En aquest article hem presentat el procés de descàrrega, processament i anàlisi per a la creació de la nova versió del corpus del Diari Oficial de la Generalitat de Catalunya. Aquest procés s'ha dut a terme amb la intenció d'actualitzar aquest corpus, ja que la versió anterior contenia documents fins l'any 2013. També es pretenia dur a terme un procés de neteja molt més exhaustiu, ja que s'havien detectat alguns errors, com segments sense traduir, segments que contenien només xifres i caràcters erronis.

Bibliografia

Corbí-Bellot, Antonio M. et al. (2005): "An open-source shallow-transfer machine translation engine for the romance languages of Spain", in: *Proceedings of the European Association for Machine Translation, 10th Annual*

15 <http://lpg.uoc.edu/corpus-DOGC>.

16 <http://web.gencat.cat/ca/menu-ajuda/ajuda/avis_legal/>.

Conference (Budapest, Hungary, 30–31.05.2005), Budapest: European Association for Machine Translation, 79–86.

Fellbaum, Christiane (1998): *WordNet: An Electronic Lexical Database and some of its Applications*, Cambridge MA: MIT Press.

Forcada, Mikel L. (2010): *Free/open-source machine translation: the Apertium platform*, Berlin: Translingual Europe.

Hunston, Susan (2006): "Corpus linguistics", in: Brown, Keith (ed.): *Encyclopedia of Language & Linguistics. Second Edition*, vol. 3, Amsterdam: Elsevier, 234–248.

Koehn, Philipp et al. (2007): "Moses: Open Source Toolkit for Statistical Machine Translation", *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague: ACL.

Oliver, Antoni / Vàzquez, Mercè (2015): "TBXTools: A free, fast and flexible tool for automatic terminology extraction", in: *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2015)*, Hissar: Bulgarian Academy of Sciences et al., 473–479.

Padró, Lluís et al. (2010): "Semantic services in FreeLing 2.1: WordNet and UKB", in: Bhattacharyya, Pushpak / Fellbaum, Christiane / Vossen, Piek (eds.): *Principles, Construction, and Application of Multilingual Wordnets, Global Wordnet Conference 2010*, Mumbai: Narosa Publishing House, 99–105.

— / Stanilovsky, Evgeny (2012): "FreeLing 3.0: Towards wider multilinguality", in: *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul: ELRA, 2473–2479.

Papineni, Kishore et al. (2002): "BLEU: a method for automatic evaluation of machine translation", in: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, : Association for Computational Linguistics, 311–318.

Reese, Samuel et al. (2010): "Wikicorpus: word-sense disambiguated multilingual Wikipedia corpus"; in: *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Malta: ELRA, 1418–1421.

Sinclair, John (1996): *EAGLES preliminary recommendations on corpus typology. EAGLES document TCWG-CTYP/P*, Pisa: ILC-CNR.

Taulé, Mariona / Martí, Maria Antònia / Recasens, Marta (2008): "AnCora: Multilevel annotated corpora for Catalan and Spanish"; in: *Proceedings of the 6th Language Resources and Evaluation Conference*, Marrakech: ELRA, 96–101.

Tiedemann, Jörg (2012): "Parallel data, tools and interfaces in OPUS", in: *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul: ELRA, 2214–2218.

Varga, Dániel et al. (2005): "Parallel corpora for medium density languages", in: *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets: IPP / Bulgarian Academy of Sciences / Association for Computational Linguistics, 590–596.

Antoni Oliver, Universitat Oberta de Catalunya, Estudis d'Arts i Humanitats, Avda. Tibidabo, 39–43, E-08035 Barcelona, <aoliverg@uoc.edu>.