



Instruments

La integració de les tecnologies de la llengua en el flux de treball del Servei Lingüístic de la UOC

Autors

Lluís Villarejo
Alba Corral
David Cullen
Universitat Oberta de Catalunya

L'article aborda les tecnologies de la llengua que es fan servir al Servei Lingüístic de la UOC i de quina manera s'integren en cadascun dels diferents processos productius del Servei. En particular, veurem com la traducció automàtica, les memòries de traducció, l'extracció de lèxic i la creació de glossaris tenen un paper important en el desenvolupament de la tasca del Servei.

Introducció

Les grans institucions, com ara les universitats, tenen grans necessitats de tractament lingüístic (premsa, màrqueting, portal...). Si a això hi afegim el caràcter virtual d'universitats com la Universitat Oberta de Catalunya (UOC), la llengua escrita passa a ocupar un paper primordial en el dia a dia de la universitat. I, si a les dimensions de la institució i al seu caràcter virtual hi sumem el context geogràfic multilingüe en què es desenvolupa la seva activitat, les necessitats de tractament lingüístic es multipliquen. Aquesta és una situació generalitzada que afavoreix la creació d'iniciatives universitàries conjuntes, com el projecte RESTAD [referències bibliogràfiques 1 i 2], per al tractament massiu del multilingüisme.

En concret, l'any 2008 el Servei Lingüístic de la Universitat Oberta de Catalunya (UOC) va servir comandes de correcció i traducció per un volum total de 17.000 pàgines de text (4.590.000 paraules) en català, espanyol o anglès i va cobrir d'aquesta manera les necessitats lingüístiques de més de 40 unitats operatives de la UOC. Davant la quantitat ingent de documents docents i administratius que el Servei Lingüístic ha de processar, sigui per corregir o per traduir, l'automatització de part del tractament lingüístic esdevé gairebé una obligació per optimitzar processos i arribar a satisfer la demanda. És en aquest punt que una anàlisi dels diferents fluxos de treball que es duen a terme diàriament al Servei Lingüístic permet identificar processos o sub processos susceptibles de ser automatitzats que reaprofitin la informació generada, que millorin l'homogeneïtat del tractament lingüístic i que optimitzin el temps dels tècnics d'assessorament alliberant-los de tasques repetitives i de poc valor afegit. Alguns d'aquests processos es poden dur a terme fent servir eines lingüístiques ja desenvolupades, mentre que d'altres necessiten un desenvolupament a mida.

En aquest article presentem, en l'apartat segon, el flux de treball del Servei Lingüístic sense l'aplicació de tecnologies de la llengua. A continuació, en l'apartat tercer, presentem un seguit d'eines que optimitzen part d'aquests processos dins el flux de treball del Servei Lingüístic. L'apartat quart descriu, amb diversos casos pràctics il·lustratius, el flux de treball comptant amb les eines descrites en l'apartat corresponent. I, finalment, en les conclusions enumerem els avantatges de la integració de les tecnologies de la llengua en el principal flux de treball del Servei Lingüístic: l'acompliment de comandes de correcció i traducció.

Flux de treball sense tecnologies de la llengua

Com s'ha dit en la introducció, al Servei Lingüístic de la UOC arriben comandes de tots els departaments de la Universitat, des de premsa, comunicació interna, passant per atenció a l'estudiant i fins als estudis i el rectorat. Malgrat la diversitat de documents que aquesta varietat d'òrgans comporta, gran part dels textos, comunicats interns, comunicats als estudiants, notícies, memòries, textos de màrqueting, són de contingut institucional i repetitiu. Generalitzant, podem dir que hi ha dos grans tipus de documents, els articles, en què hi ha molta terminologia científica, i els documents de funcionament intern de la Universitat, textos institucionals o adreçats a l'estudiant, en què es repeteixen estructures, terminologia i gran quantitat de text.

Les llengües principals amb què tracta el Servei Lingüístic són el català, l'espanyol, l'anglès i, en menor mesura, el francès. Fins ara hi havia només el Campus Virtual (l'entorn de suport i aprenentatge des d'on s'accedeix als recursos i s'interactua amb la comunitat universitària) en català i espanyol, però, amb l'entrada a l'espai europeu d'ensenyament superior, se'n crea un de nou en anglès, que es preveu ampliar a altres idiomes. Això fa que les comandes d'anglès hagin augmentat considerablement aquest últim any i que molts textos ja es demanin directament en tots tres idiomes.

La gestió de comandes al Servei Lingüístic es fa amb el correu electrònic i una base de dades en què es van creant automàticament els registres de les comandes que entren. Quan entra una comanda s'ha d'avaluar què demana l'usuari (correcció, traducció o supervisió) i si el termini que ha posat és adequat, per a la qual cosa es fa el recompte de paraules del document i, a partir d'unes mesures que hi ha establertes, es decideix quants dies es necessiten realment per fer la feina. Si la comanda consisteix en fitxers en format DOC, el recompte no implica cap dificultat. Però en el cas, per exemple, d'un conjunt de fitxers que continguin formats XML i HTML, el recompte no és trivial. El recompte de paraules i la data de lliurament final es registren a la base de dades per tenir un control de les comandes. Els documents que arriben normalment són en format DOC, però també n'arriben en ODT, XLS, HTML i XML, PPT, PDF o InDesign en cas de galerades, TXT o DAT.

La major part de les comandes que entren al Servei s'externalitzen a algun dels col·laboradors que té disponibles. Normalment es mira de posar un marge de temps per poder fer una revisió ràpida de la feina lliurada pel col·laborador, però no sempre és possible. Si es revisa i es considera necessari, se li envia un retorn de feina, normalment de recordatori de criteris propis.

Finalment, es lliura el document o documents a l'usuari que ha fet la comanda i s'acaba de tancar el registre que s'ha obert. El Servei Lingüístic intenta oferir qualitat segons el termini que té. La qualitat dels originals també és un factor decisiu, sobretot si se'n demana directament la traducció, per la qual cosa primer es procura fer la correcció del document original, encara que de vegades no hi ha prou temps.

Eines

El flux de treball presentat en l'apartat anterior consta de diverses tasques que poden ser automatitzades, o semiautomatitzades, mitjançant l'ús d'eines de tecnologies de la llengua, de manera que els tècnics de llengua disposin de més temps per dedicar-se a la seva feina principal: l'elaboració de criteris, la traducció, la correcció i la revisió de documents. En aquest sentit, el Servei Lingüístic de la UOC incorpora tecnologies de la llengua de desenvolupament propi i aliè, sempre posant èmfasi en el programari lliure,¹ que miren d'optimitzar els processos productius que implica l'acompliment de comandes.

Les eines que es fan servir al Servei Lingüístic de la UOC són de distribució lliure i cost zero. Originalment es feien servir productes de propietat, però hi ha hagut un canvi de filosofia en els últims anys tant en l'àmbit universitari com en l'àmbit dels serveis lingüístics, i les solucions de distribució lliure, i preferentment les de codi obert, tenen un pes cada vegada més important en els fluxos de les institucions. Les eines que s'han creat al Servei Lingüístic es posen a disposició de tothom mitjançant llicències de codi obert i han estat desenvolupades íntegrament en llenguatge Perl,² i, per tant, són programes multiplataforma, la qual cosa vol dir que es poden fer servir tant en sistemes operatius Windows com en Linux.

SALI, alineador de segments

L'alineador SALI, desenvolupat al Servei Lingüístic, genera com a resultat un fitxer bilingüe paral·lelitzat a partir del qual es pot fer extracció de lèxic d'especialitat amb equivalents de traducció o que pot ser convertit en memòria de traducció, per ser aprofitada dins un programa de traducció assistida o automàtica.

L'alineador rep com a entrada fitxers en text pla. Típicament aquests fitxers són un conjunt de documents originals, corregits en una llengua i la seva traducció en una segona llengua. A partir d'aquest conjunt de documents, l'alineador genera, de manera automàtica, documents alineats en segments. Un segment és una frase o part d'una frase, i un segment alineat és un segment en la llengua d'origen relacionat amb un segment en la llengua de destinació. Aquests segments són alineats mitjançant tècniques estadístiques de coocurrència independents del parell de llengües amb què es treballa.

El resultat de l'alineació són tres fitxers en text pla. El primer és un fitxer bilingüe paral·lelitzat que conté els segments alineats. El segon i el tercer són fitxers monolingües que contenen els segments, de la llengua d'origen i destinació respectivament, per als quals l'alineador no ha tingut prou indicis per aparellar-los. L'usuari, segons les seves necessitats, pot modificar el paràmetre del llindar de confiança. El llindar de confiança és el valor resultant de l'aplicació de les tècniques de coocurrència per sobre del qual el parell de segments avaluat és considerat alineable. Si l'usuari no vol obtenir segments mal alineats, haurà d'apujar el nivell de confiança i el programa només alineirà els segments per als quals tingui molts indicis per aparellar-los. Si l'usuari vol obtenir el màxim de segments alineats, encara que no totes les alineacions siguin correctes, haurà d'abaixar el llindar de confiança.

Lexterm, extractor de lèxic per a terminologia i traducció

L'extractor de lèxic Lexterm [referència bibliogràfica 3], desenvolupat al Servei Lingüístic, genera, a partir d'un corpus bilingüe paral·lelitzat com el que es pot obtenir amb l'alineador, llistes bilingües de lèxic d'especialitat que són útils per a la traducció automàtica i la traducció assistida per ordinador.

L'extractor rep com a entrada un corpus monolingüe o bilingüe. Si el corpus és bilingüe, un cop feta l'extracció de lèxic es pot fer una cerca automàtica d'equivalents de traducció. A més del corpus, l'extractor pot rebre com a entrada una llista de paraules "buides" (funcionals) que es fa

1. http://ca.wikipedia.org/wiki/Programari_Lliure

2. <http://www.perl.org>

servir per descartar unitats no rellevants; per exemple, no ens interessen unitats lèxiques que continguin en la primera o darrera posició una preposició o un article. Un cop es tenen el corpus i la llista de paraules buides, l'extractor fa servir tècniques estadístiques per generar una llista de lèxic a partir del corpus inicial. En aquest punt, i si el corpus de treball és bilingüe, l'extractor ofereix per a cada unitat lèxica els candidats de traducció més probables mitjançant l'aplicació de tècniques de coocurrència. Entre els candidats de traducció que l'extractor ofereix, l'usuari pot deixar el que l'extractor considera com a més probable o triar manualment el que consideri més convenient.

El resultat de l'extracció és una llista d'unitats lèxiques monolingüe o bilingüe, segons el corpus de treball, en text pla. Si la llista és monolingüe, constitueix una font d'informació valuosa per tenir una idea de les unitats lèxiques que es fan servir en el corpus de treball. Si la llista és bilingüe, es pot fer servir per alimentar el lèxic del motor de traducció o com a glossari en un programa de traducció assistida per ordinador (TAO).

Apertium, motor de traducció automàtica

Apertium [referència bibliogràfica 4] és un motor de traducció automàtica de codi lliure en què el Servei Lingüístic de la UOC, juntament amb l'empresa Prompsit (www.prompsit.com), ha invertit esforços per donar suport a la diversitat de formats amb què es troba en les comandes de traducció.

El motor de traducció fa servir un mecanisme de transferència superficial que processa el text d'entrada en etapes: desformatatge, anàlisi morfològica, desambiguació categorial, transferència estructural superficial, transferència lèxica, generació morfològica i reformatatge. El motor tradueix documents en els parells de llengües següents: català <> espanyol, català <> anglès i català <> francès. Els resultats de la traducció automàtica milloren considerablement quan les llengües implicades en la traducció estan emparentades. En aquest cas, les afinitats morfològiques, sintàctiques i semàntiques simplifiquen el disseny d'aquests sistemes i permeten arribar fàcilment a traduccions en les quals únicament el 5 % del text és incorrecte. En qualsevol cas, si es vol difondre el text, s'ha de fer una revisió de la traducció lliurada pel sistema de traducció automàtica.

Els formats amb què treballa el motor són els següents: text pla, RTF, ODT, HTML, XHTML, ODS, PPT, PPTX, DOC, DOCX, XLS, XLSX. A més de la traducció d'aquests tipus de documents, el motor ofereix diverses prestacions, com traduir directament els continguts de qualsevol lloc web o rebre documents comprimits i traduir el seu contingut respectant l'estructura interna de carpetes. Un dels desenvolupaments a mida que s'ha fet és el que ofereix la possibilitat de fer servir memòries de traducció, de manera que el motor incorpora els segments ja traduïts presents a la memòria de traducció i només tradueix el que no troba a la memòria de traducció.

Recompte de paraules

La diversitat de formats amb què tracta el Servei obliga a fer servir diverses eines per fer el recompte de paraules de les comandes. En aquest sentit, el Servei Lingüístic fa servir diverses eines de codi lliure que possibiliten el recompte de tots els formats amb què tracta. Entre les eines més utilitzades hi ha les comandes d'Unix³ per a la manipulació de fitxers. Aquestes eines permeten manipular els fitxers i aplicar una sèrie de filtres molt útils a l'hora de destriar les paraules que s'han de tractar de

3. http://research.microsoft.com/users/church/wwwfiles/tutorials/unix_for_poets.ps

la informació de format. A més de les comandes d'Unix també es fan servir altres aplicacions de distribució lliure com RecFly⁴ i CountAnything,⁵ que ens ajuden a completar el ventall de formats en què podem fer el recompte de paraules.

Tractament de formats

La conversió de formats és indispensable per poder processar els fitxers i compartir recursos; per aquest motiu, el Servei Lingüístic ha desenvolupat eines que permeten fer aquestes conversions de manera automàtica. Aquestes eines permeten la conversió dels formats HTML, DOC i PDF a TXT per possibilitar l'alineació de documents, i la conversió de TXT a TMX per possibilitar la creació de memòries de traducció.

Olifant

Olifant és una eina de codi lliure que permet la manipulació i el manteniment de memòries de traducció en diferents formats, com TMX, format estàndard per a l'intercanvi de memòries de traducció (Translation Memory eXchange),⁶ o text pla delimitat per tabuladors. Un cop carregades les memòries de traducció, per mitjà d'una interfície gràfica es poden editar les unitats de traducció, els seus atributs i qualsevol altra dada associada. Entre les prestacions que ofereix l'eina hi ha la cerca d'entrades duplicades, la cerca i reemplaçament de text mitjançant expressions regulars, el filtratge d'entrades fent servir diversos criteris, la unificació de dues memòries en una i l'exportació total o parcial de les memòries a diversos formats compatibles amb els principals programes de TAO.

OmegaT

OmegaT⁷ és un entorn de TAO de codi lliure que permet la traducció de documents mitjançant la recuperació eficient de segments presents en memòries de traducció. Entre les principals característiques de l'eina hi ha l'ús simultani de diverses memòries de traducció, l'aplicació de coincidències parcials, la propagació de coincidències, l'ús de glossaris externs, el tractament d'un ampli conjunt de formats i la compatibilitat amb Unicode (UTF-8) per al tractament d'alfabets no llatins.

Flux de treball amb tecnologies de la llengua

Cadascuna de les eines descrites en l'apartat anterior encaixa en algun dels punts del procés productiu del Servei Lingüístic, com es pot veure en la figura 1. Algunes eines es fan servir només en determinats processos i per a determinats parells de llengües, com és el cas del motor de traducció, que es fa servir en el procés de traducció de documents del català a l'espanyol i viceversa, mentre que altres eines, com les de recompte de paraules, necessàries per saber el volum de la comanda i establir el termini de servei de la comanda, es fan servir per a totes les comandes independentment del parell de llengües. Per exemplificar la integració de les eines en els fluxos de treball del

4. <http://www.traduit.com/cas/recfly.html>

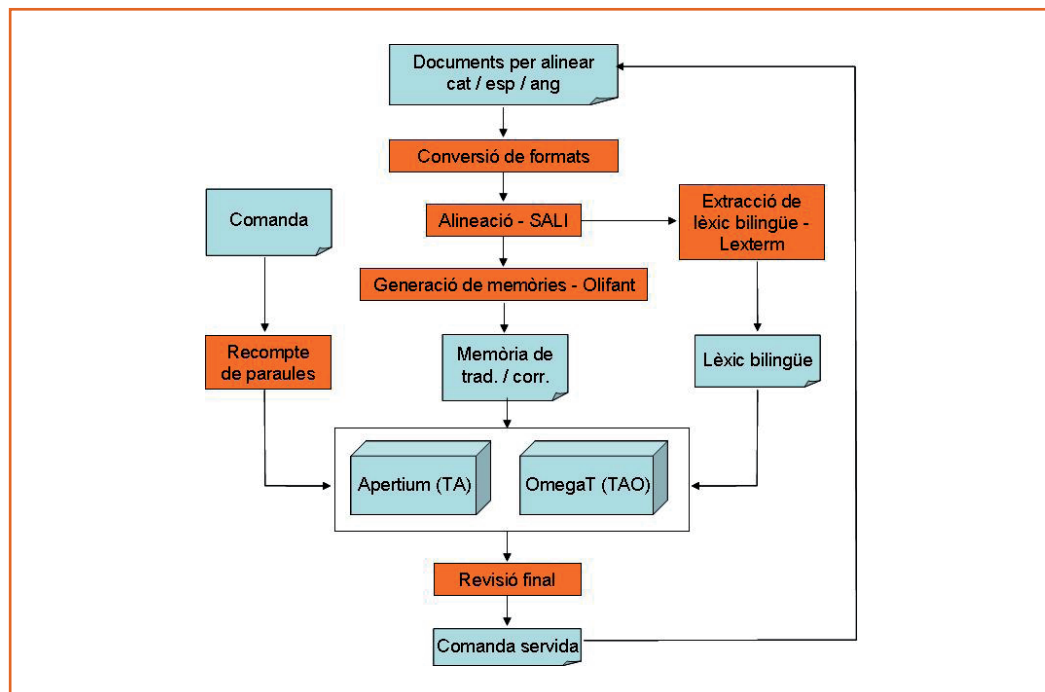
5. <http://ginstrom.com/CountAnything/>

6. <http://www.lisa.org/Translation-Memory-e.34.0.html>

7. <http://www.omegat.org>

Servei Lingüístic ens centrarem en tres casos pràctics: la traducció del català a l'espanyol i viceversa mitjançant la traducció automàtica, la traducció del català a l'anglès i viceversa mitjançant les memòries de traducció i, per acabar, la correcció de documents en català mitjançant les memòries de correcció.

Figura 1. Flux de treball amb tecnologies de la llengua.



La traducció automàtica: la traducció català <> espanyol

En el moment en què arriba una comanda de traducció que implica el català i l'espanyol, sigui quina sigui la direcció, l'original es passa pel motor de traducció automàtica. El procés és molt senzill. El motor de traducció té una interfície web que permet pujar el fitxer que es vol traduir, en què s'especifica el parell de llengües involucrat en la traducció, i el motor de traducció retorna un fitxer en el mateix format que l'original. A més a més, el motor de traducció automàtica Apertium incorpora la possibilitat de fer servir una memòria de traducció per pretraduir el document abans que el motor actuï. L'execució d'aquesta pretraducció és transparent per a l'usuari, ja que només ha d'especificar la memòria de traducció que vol fer servir en el moment d'enviar el fitxer al motor.

Com s'ha esmentat abans, el resultat de la traducció automàtica millora considerablement quan es tracta de llengües emparentades. Per al cas del català <> espanyol, les estimacions del rendiment de la traducció situen en el 5 % el percentatge del text que és incorrecte, encara que depèn del tipus de document. Aquests percentatges no tenen un impacte directe en el nombre de pàgines que el tècnic de llengua ha de revisar per tancar l'esborrany de traducció. És a dir, el 5 % no vol dir que en un document de 100 pàgines el tècnic de llengua només s'hagi de mirar 5 pàgines en total. Ara bé, sí que donen una idea del temps que ha d'invertir en la postedició per tancar l'esborrany en relació amb el temps que representaria fer la traducció des de zero. Així, doncs, l'esborrany que s'obté automàticament proporciona un punt de partida que estalvia temps al tècnic de llengua i costos en cas d'enviar la comanda a un col·laborador extern, ja que se li dona part de la feina feta.

Un cop es disposa de la traducció posteditada final, s'enceta el procés per crear una memòria de traducció. Per fer això, és necessari tenir els documents en format TXT i alineats. La gran majoria dels documents que arriben al Servei tenen format DOC. Per tant, es fa servir l'eina de conversió de DOC a TXT per passar-los al format que llegeix SALI. El fitxer que surt de SALI és un TXT (que s'ha de desar amb la codificació UTF-8 fent servir un editor de text). Aquest fitxer és la memòria de traducció, però, abans de fer-la servir, s'ha de netejar i passar al format estàndard per a memòries de traducció, TMX. Per fer això, s'importa des d'Olifant. Poden sorgir problemes amb la importació a causa d'errors en la codificació, cas en què s'ha de fer servir un editor de text (com ara el Notepad++) per trobar i eliminar aquests errors. Aquestes memòries són reaprofitades en futures comandes en què el contingut que s'hagi de traduir sigui semblant al de la traducció feta. Així, les memòries, com que són codificades en el format estàndard de memòries de traducció, es poden muntar en un projecte i enviar als col·laboradors perquè les facin servir amb OmegaT o qualsevol de les eines de traducció assistida per ordinador disponibles al mercat.

Les memòries de traducció: la traducció català <> anglès

Ara per ara, l'ús de la traducció automàtica per a traduccions del català i espanyol a l'anglès i vice-versa no és factible. Els costos de postedició són tan elevats que no resulta viable. El rendiment dels motors és cada vegada millor, però encara queda camí. S'han fet algunes proves inicials amb alguns dels motors disponibles (Translendum, Apertium i Google Translate), però el resultat ha estat que el seu rendiment no justifica la seva incorporació al flux de treball per a les traduccions que involucren l'anglès. En aquest sentit, el Servei ha encetat projectes de col·laboració per millorar la qualitat lingüística del motor Apertium per a tots els parells de llengua que s'utilitzen.

En canvi, les memòries de traducció sí que donen resultats prou significatius. Com s'ha vist en l'apartat corresponent, les comandes que arriben al Servei presenten un alt grau de repetitivitat, tant en la terminologia com en els continguts dels documents que s'han de tractar. En certs casos, la documentació d'un any a l'altre és bàsicament la mateixa; per exemple, els manuals que es lliuren abans de començar el curs al professorat, les memòries anuals o les llistes de cursos oferts per la Universitat. Igualment, a vegades, el Servei ha de gestionar projectes molt grans, com ara la traducció del Campus Virtual. Amb projectes d'aquestes dimensions les memòries de traducció poden arribar a reduir els costos i terminis d'una manera important. Tanmateix, el seu ús en el dia a dia també ajuda a mantenir la coherència del vocabulari que fan servir els diversos traductors que treballen amb el Servei.

Hi ha dos casos comuns en l'ús de les memòries al Servei. El primer és l'alimentació i aprofitament d'una memòria general i el segon, la creació expressa d'una memòria per a un projecte concret. Des de fa un temps el Servei ha anat alimentant memòries amb totes les comandes que són sotmeses a un control de qualitat. Tota comanda que arriba de l'oficina de premsa de la Universitat acaba a la memòria pel fet que les notes de premsa tenen un format estàndard que es repeteix en cada document. També s'ha creat una memòria amb totes les comandes, ja que en alguns casos una nota de premsa tractarà una notícia que també tindrà ressò en documents creats per altres àrees de la Universitat, com poden ser l'institut de recerca, el rectorat o vicerectorats, o els estudis.

En el cas d'afrontar un projecte de grans dimensions, sempre es demana als responsables material anterior que s'hagi traduït per poder crear una memòria i extreure lèxic. En el cas de no disposar d'aquest material, les memòries també poden tenir un paper clau en el procés. El procediment, en aquest cas, és construir la memòria sobre la base dels lliuraments parcials que arriben dels traductors durant el desenvolupament del projecte en si. Així, a mesura que es va avançant, la memòria va creixent, i l'aprofitament de la feina feta és cada vegada més gran i millor. Si no es disposa de lèxic específic per a la comanda, es pot fer servir Lxterm per fer una extracció de lèxic bilingüe sobre el

corpus proporcionat pels responsables. Aquesta extracció de lèxic és transformada en un glossari que pot ser carregat tant a Apertium com a OmegaT, els quals, cada cop que detecten una paraula present en el glossari, mostren automàticament la traducció corresponent.

Les memòries de correcció: la correcció del català

Una part important de les comandes que rep el Servei Lingüístic són de correcció de textos en català. En aquest tipus de comandes les tecnologies de la llengua tenen la seva incidència per mitjà de la traducció assistida per ordinador. Quan una comanda de correcció de català es tanca, es crea una memòria de correcció fent servir el procés de creació d'una memòria de traducció descrit anteriorment. En comptes de fer servir un document en una llengua i la seva traducció, es fan servir l'original i la correcció de l'original. D'aquesta manera, es disposa d'una memòria de correcció en què hi ha relacionats segments incorrectes lingüísticament i la correcció corresponent. Així, el tècnic de llengua pot treballar els documents que s'han de corregir dins OmegaT, l'entorn de traducció assistida, que identificarà segments erronis del document per corregir que ja van ser corregits en comandes anteriors (o anteriorment en el document actual per corregir).

Conclusions

Com a resultat d'una anàlisi dels processos productius del Servei Lingüístic, l'experiència dels tècnics de llengua i la identificació i disponibilitat de tecnologies de la llengua, s'ha dissenyat una integració de les tecnologies de la llengua en el procés de servei de comandes del Servei Lingüístic que no solament consisteix en la simple aplicació d'eines, sinó en la modificació en si dels fluxos de treball. Un exemple d'aquesta modificació de fluxos de treball és el fet que, abans de l'aplicació de les tecnologies de la llengua, la comanda es donava per tancada quan es disposava del document final, fos una traducció o una correcció. Ara, amb l'aplicació de les eines, el flux de treball no s'acaba en aquest punt, sinó que s'amplia amb la generació de memòries de traducció o correcció i l'extracció de lèxic per capitalitzar el coneixement lingüístic de la feina desenvolupada en el procés de producció i alimentar les eines de traducció automàtica i assistida.

Com s'ha vist en l'apartat anterior, la integració de les tecnologies de la llengua en els processos productius del Servei Lingüístic és d'una profunditat considerable. D'una banda, la disponibilitat d'un motor de traducció amb capacitat per tractar un ampli ventall de formats i una taxa de rendiment elevada per a llengües properes representa un estalvi de temps considerable a l'hora de traduir documents. D'altra banda, l'explotació dels recursos propis, mitjançant l'aplicació de les eines, facilita el reaprofitament de la informació de traducció i correcció, cosa que permet estalviar temps i costos a més de proporcionar els mitjans automàtics per aportar homogeneïtat a la traducció i la correcció.

Per acabar, l'aplicació d'aquestes tecnologies, i la seva integració en els processos productius, ha suscitat un gran interès en institucions alienes a la UOC. És el cas del Centre de Terminologia Termcat, amb el qual es va encetar el 2007 un conveni de col·laboració per a portar terme el projecte TALAIA⁸ [referències bibliogràfiques 5 i 6], un observatori de terminologia de la societat del coneixement que recull, a partir d'un buidatge semiautomatitzat i l'aplicació d'algunes de les eines descrites en aquest article, la terminologia en català, espanyol i anglès utilitzada en els articles de les revistes acadèmiques⁹ de la UOC.

8. <http://www.termcat.cat/dicci/talaia>

9. http://www.uoc.edu/portals/catala/difusio_i_publicacions/revistes_digital/index.html

Bibliografia

DE BLAS, M. [et al.]. "Projecte RESTAD. Recursos de suport a la traducció automatitzada aplicats a la docència". *Llengua i Ús* (abril 2007), núm. 38, p. 41-46. ISSN 2013-052X.

VILLAREJO, I.; MORÉ, J.; VÁZQUEZ, M. "Projecte RESTAD: Eines de codi lliure per a la traducció i postedició de documents". *Congrés Internacional de Sistemes de Programari Lliure i Codi Obert (FLOSS)* (març 2007), p. 262-271.

MORÉ, J.; VÁZQUEZ, M.; VILLAREJO, L. "Lexterm, an Open Source Tool for Lexical Extraction". *4t Seminari Internacional Slovko* (octubre 2007), p. 183-189.

RAMÍREZ-SÁNCHEZ, G. [et al.]. "Opentrad Apertium open-source machine translation system: an opportunity for business and research". *Actes de la Conferència Internacional Translating and the Computer 28* (novembre 2006).

MONTES, D.; VÁZQUEZ, M. "L'observatori de terminologia Talaia: un exemple d'innovació per mitjà de la cooperació". *Llengua i Ús* (abril 2009), núm. 44, p. 26-35. ISSN 2013-052X.

MORÉ, J.; VÁZQUEZ, M.; RIUS, L.; VILLAREJO, L. "L'observatori de terminologia Talaia: mètode i processos". *Revista Tradumàtica: Terminologia i Traducció* (desembre 2008). ISSN 1578-7559.

