

Projecte RESTAD. Recursos de suport a la traducció automatitzada aplicats a la docència

Assessorament i Terminologia

Introducció

Les universitats catalanes, actualment, estan en procés de reformar els seus estudis per a adaptar-los al marc de l'espai europeu d'ensenyament superior (EEES). Aquest nou marc representa un canvi de paradigma en el model d'ensenyament-aprenentatge de nivell superior, en els aspectes següents: homologació de les titulacions reconegudes entre els diferents països; increment de la mobilitat d'estudiants i de professors durant el grau i el postgrau; importància de la docència virtual i dels materials elaborats o recopilats pel mateix professor; augment de la relació interpersonal professor-alumne; recerca personal de l'alumne com a part primordial del procés d'aprenentatge.

Aquests canvis incidiran clarament en els usos lingüístics docents i discents. La llengua oral de la docència, que fins ara era l'indicador clau per a determinar la situació sociolingüística a les universitats, ja no serà el factor determinant, sinó que el procés d'aprenentatge esdevindrà més complex, hi intervindran més elements, cosa que provocarà que l'alumne esdevingui un agent que triarà la llengua que li resulti més convenient per a comunicar-se amb el professor i amb la resta d'alumnes de l'aula i per a accedir als materials de la xarxa.

Les TIC i les tecnologies aplicades al processament del llenguatge han de permetre preservar un espai per al català com a llengua preferent de la docència i aprofitar la conjuntura del marc del nou disseny metodològic europeu com una oportunitat per a la llengua catalana a la universitat, perquè continuï sent la llengua de comunicació a l'aula i per a promocionar una nova manera de treballar dins de cada universitat i entre universitats.

En aquest context sorgeix el projecte RESTAD, un projecte interuniversitari desenvolupat pels serveis lingüístics de la Universitat Autònoma de Catalunya, la Universitat Politècnica de Catalunya, la Universitat de Girona, i la Universitat Oberta de Catalunya i finançat el 2006 pel

Departament d'Universitats, Recerca i Societat de la Informació (actualment Departament d'Innovació, Universitats i Empresa) de la Generalitat de Catalunya. Des del primer moment les universitats han contribuït a impulsar, de manera decidida, la col·laboració interuniversitària en l'àmbit del suport lingüístic, des del convenciment que la gestió del multilingüisme necessita ajuntar esforços, compartir coneixements i continguts per a assolir uns resultats que siguin garantia de qualitat i ajudin a afrontar els reptes de l'EEES.

El projecte ha consistit a desenvolupar recursos que facilitin la traducció automatitzada de documents docents, especialment la documentació que el professorat posa a disposició de l'alumnat en el campus virtual, tant si es tracta d'assignatures presencials com d'assignatures virtuals i, també, de documents acadèmics i administratius, els quals cada cop més s'hauran de poder oferir en dues o tres llengües. Aquests recursos tenen dos grans objectius globals:

- Sistematitzar i reaprofitar la informació lingüística que es treballa en els serveis lingüístics universitaris amb la finalitat de fer més fàcil la traducció de materials docents al català i d'assegurar la qualitat dels textos docents en aquesta llengua amb la creació de memòries de traducció i de gestió de terminologia a gran escala.
- Desenvolupar aplicacions informàtiques que permetin optimitzar els programes de suport a la traducció de què ja disposen els serveis lingüístics, com ara alineador automàtic de textos, gestor de memòries de traducció i extractor de lèxic a partir dels textos de la memòria de traducció (entrades en català amb els equivalents en altres llengües).

El projecte s'ha concebut en clau de programari lliure que es posa a disposició de totes les universitats, institucions, enti-

Autores

Marta de Blas
Servei de Llengües i Terminologia, Universitat Politècnica de Catalunya
Marta Estella
Servei de Llengües, Universitat Autònoma de Barcelona
Meius Ferrés
Servei de Llengües Modernes, Universitat de Girona
Imma Sánchez
Servei Lingüístic, Universitat Oberta de Catalunya

Aproximació al projecte Restad i als treballs que es desenvolupen en aquest marc d'investigació pel que fa a la traducció automàtica.

tats i particulars interessats a fer-ne ús, amb la qual cosa facilita que es puguin fer adaptacions a altres necessitats.

Les eines

En el marc del projecte RESTAD, el desenvolupament de les eines informàtiques i la formació per a conèixer-les adreçada als tècnics lingüístes han anat a càrrec del Servei Lingüístic de la UOC i les proves pilot d'aplicació i ús de les eines per a l'explotació de continguts han anat a càrrec del Servei de Llengües de la UAB, Servei de Llengües i Terminologia de la UPC, Servei de Llengües Modernes de la UdG i, també, del Servei Lingüístic de la UOC.

Les eines s'han desenvolupat íntegrament en llenguatge Perl i, per tant, són programes multiplataforma. Totes les eines que es descriuen són de codi obert i de distribució lliure. Es distribueix el codi font i la versió executable per a Windows i es poden consultar al web del Servei Lingüístic de la UOC, espai RESTAD.

(<http://www.uoc.edu/serveilinguistic/home/restad/restad.html>).

Aquestes eines donen suport a la traducció amb explotació de continguts, alineador de textos (AlinUOC) i extractor de lèxic d'especialitat (Stem-LES), i a l'edició i suport a la revisió de documents: cercador d'enllaços (CREN).

AlinUOC

AlinUOC és un paquet de programes, basat en l'eina d'alineació bilingüe de frases, el Bilingual Sentence Aligner de Robert C. Moore, que serveix per a alinear documents. Si tenim dos documents A i B, en què B és la traducció d'A, l'alineació és el procés d'aparellar cada frase del document A amb la seva frase equivalent en el document B. L'objectiu principal de l'eina és obtenir documents alineats que estiguin preparats per a ser convertits en una memòria de traducció o en un corpus a partir del qual es pot fer extracció de lèxic d'especialitat amb equivalents de traducció. El paquet AlinUOC conté els següents programes: DOC2SNT-Multi i Stem-Aligner-Multi. El programa DOC2SNT-Multi s'ocupa del pretractament dels documents que s'han d'alinejar i el programa Stem-Aligner-Multi s'ocupa de l'alineació pròpiament dita.

Els documents per a alinear han de passar un procés de pretractament per a convertir-los en text pla, ja que no poden tenir cap marca de format. A més, com que l'alineador aparella les frases d'un document original amb les frases corresponents del document traduït, és convenient segmentar els documents per frases. Cal que cada document per a alinear que no tingui text pla passi per un procés de segmentació i de conversió a un fitxer en format SNT. Si els documents per a alinear són en format TXT però no estan segmentats, també s'han de convertir en fitxers SNT. La conversió i segmentació es fa amb l'eina DOC2SNT-Multi, que està preparada per a convertir a format SNT documents HTML, DOC, RTF o TXT.

El programa Stem-Aligner-Multi alinea documents en format SNT. El programa pot alinear un conjunt de documents originals amb les seves traduccions. Per a dur a terme aquesta operació, cal que el nom dels documents, original i traduït, contingui la informació necessària perquè l'alineador identifiqui i aparelli correctament els segments del document original amb els segments de la seva traducció.

L'alineador acara els documents que coincideixen en el nom del fitxer, excepte en el distintiu del codi de llengua. Els distintius de llengua segueixen la codificació de llengües segons l'ISO-639. Si es volen alinear dos documents DOC en català amb les seves traduccions a l'espanyol (per exemple, el document original 1 és la memòria econòmica d'una institució i el document original 2 conté els estatuts de la mateixa institució) aquests documents han de tenir un nom en què hi hagi, per una banda, un nom descriptiu del contingut i, per una altra banda, el codi de llengua en la qual són escrits. Tots dos elements s'han de distingir separant-los amb un guió baix. Els codis de llengua poden ser amb majúscules o minúscules. Així, els documents originals 1 i 2 es podrien dir *memòria-econòmica_CA.doc* i *estatuts_CA.doc*. Perquè l'alineador aparelli els segments dels documents en català amb els segments de la seva traducció a l'espanyol, el codi de llengua dels documents de les traduccions seria ES i tindrien els noms següents: *memòria-econòmica_ES.doc* i *estatuts_ES.doc*.

L'alineador aparella una frase F1 del document 1 amb una frase F2 del docu-

ment 2, segons la probabilitat que F2 sigui l'equivalent de F1 en la llengua de destinació. El valor de probabilitat va de 0 a 1. Si el valor és 0, F2 no és l'equivalent de F1. Si el valor és 1, llavors F2 és l'equivalent de F1. Malgrat tot, els resultats del càlcul de probabilitat sovint són entre 0 i 1 i, per tant, l'usuari ha d'escollir el llindar a partir del qual l'alineador decidirà acceptar o rebutjar F2 com a equivalent de F1. L'Stem-Aligner-Multi té un llindar (*threshold*), per defecte, de 0,6. Ara bé, l'usuari el pot modificar, però ha de tenir en compte que, si posa un llindar baix, és molt probable que F1 s'identifiqui amb una frase que no és el seu equivalent. En canvi, si el posa alt corre el risc de trobar poques frases alineades perquè és molt exigent.

El resultat de l'alineació és un fitxer de text amb els segments alineats i separats per un tabulador. Aquest fitxer es crea en la carpeta que ha seleccionat l'usuari per a col·locar-lo. A més, de la mateixa manera que hem desat el fitxer de segments alineats, podem desar un fitxer amb els segments que no s'han pogut alinear. Això permetrà, si volem, revisar el resultat o alinear-los manualment.

El fitxer resultat de l'alineació està preparat per a ser operatiu en sistemes de traducció assistida que poden consultar memòries de traducció amb el format de text tabulat. De totes maneres, per a garantir la possibilitat que qualsevol traductor pugui utilitzar la memòria, independentment de l'eina de traducció assistida que utilitzi, s'ha desenvolupat una eina, anomenada TXT2TMX, que transforma el fitxer de text tabulat al format TMX, que és el format estàndard per a treballar les memòries de traducció.

Stem-LES

Stem-LES (*Lexical Extraction Suite*) és un programa que crea llistes monolingües i bilingües de lèxic rellevant que són útils per a la traducció automàtica i assistida. Gràcies a l'aplicació de mètodes estadístics, la creació d'aquestes llistes es fa d'una manera ràpida i eficient.

L'objectiu principal de l'extracció de lèxic rellevant és obtenir candidats a terme. També es poden recollir combinacions de paraules considerades unitats lèxiques amb una traducció específica, sense ser considerades estrictament termes.

El programa Stem-LES permet crear llistes bilingües de lèxic rellevant perquè detecta amb mètodes estadístics la traducció més probable d'una determinada unitat lèxica a partir d'un corpus paral·lel. Aquest corpus paral·lel es pot obtenir mitjançant l'alineació de documents amb les seves respectives traduccions amb el paquet AlinUOC descrit en l'apartat anterior.

Stem-LES determina quina paraula o conjunt de paraules d'un segment traduït correspon a la traducció de la unitat lèxica en qüestió. El procés estadístic no sempre encerta la selecció; per aquest motiu, el programa mostra més d'un candidat possible perquè l'usuari pugui triar el que consideri correcte. El programa també permet veure el context d'aparició en la llengua de destinació del candidat a equivalent de traducció.

Stem-LES fa servir una tècnica estadística per a l'extracció automàtica de lèxic rellevant basada en el càlcul de tots els *n-grams* de paraules. Normalment des de $n = 2$ paraules fins a una n determinada per l'usuari; per defecte, s'ha determinat que sigui 3. És a dir, es calculen totes les combinacions de dues paraules, de tres paraules, etc. que hi ha en un document o en un conjunt de documents. Entre aquestes combinacions hi ha les unitats lèxiques rellevants que trobem en el text, però també hi ha moltes altres combinacions no rellevants.

Per a reduir notablement el nombre de candidats no rellevants, és imprescindible filtrar els resultats amb una llista de paraules buides (*stop words*). Les paraules buides són les que habitualment no són en la posició extrema (la primera o l'última paraula) d'una unitat lèxica rellevant. Normalment les llistes de paraules buides són llistes de paraules funcionals (preposicions, conjuncions, pronoms relatius, etc.). Per tant, si tenim el text *Descripció de la tecnologia virtual*, combinacions com *Descripció de*, *Descripció de la* o *de la tecnologia* no es consideren unitats lèxiques rellevants, mentre que *tecnologia virtual* sí. Els resultats del procés d'extracció dependran de la qualitat de la llista de paraules buides, que l'usuari podrà mantenir i millorar simplement modificant-la en un editor de textos.

Un cop eliminats tots els *n-grams* que comencen o acaben amb una paraula buida, disposem d'una llista de combina-

cions de paraules ordenades per ordre de freqüència, com ara l'exemple següent:

Taula1. Freqüència i combinació

Freqüència	Combinació
24	societat del coneixement
23	net art
12	value Web
13	història
13	universitat
10	llengües minoritàries
9	vida artificial
8	realitat virtual
8	interfície
7	difusió del patrimoni
7	interactivitat
6	procés textual

S'estableix un criteri d'aparició per a la selecció de candidats a terme basat en la freqüència d'aparició. Les freqüències baixes en un corpus textual de gran volum no són rellevants. La llista de candidats sempre ha de ser revisada, ja que no totes les combinacions seran realment rellevants. Si el revisor detecta combinacions amb paraules extremes que s'han de considerar buides, cal que les afegeixi a la llista de paraules buides.

Per a ajudar a la revisió, Stem-LES permet a l'usuari consultar els contextos en què apareix la combinació de paraules, de manera que el context d'aparició li serveix per a decidir si la considera una unitat lèxica rellevant o no.

CREN

L'eina CREN (cerca i revisió d'enllaços) revisa automàticament els enllaços a pàgines web d'un o més documents DOC o HTML, fins i tot una pàgina web sencera. El tret distintiu d'aquesta aplicació és que quan troba un enllaç inactiu busca la nova adreça de la pàgina web a què l'enllaç inactiu fa referència.

La detecció de la nova adreça es pot fer comprovant si hi ha una diferència en l'extensió HTML o HTM que apareix en l'enllaç o bé fent una cerca automàtica de pàgines web de la xarxa que tinguin com a títol el mateix text ancorat de l'enllaç o un títol que contingui aquest text. La cerca automàtica es fa mitjançant crides a l'API del cercador Google. El paràmetre d'entrada és la cadena de paraules que s'ha de cercar, que és el text ancorat de

l'enllaç. La sortida és una relació de pàgines web, amb els URL respectius, que contenen el text ancorat. Entre aquestes pàgines identifiquem l'URL de les pàgines els títols de les quals coincideixen amb el text ancorat o el contenen.

A continuació presentem un exemple de localització de la nova adreça d'una pàgina web seguint l'estratègia de cerca automàtica. En un document HTML en què apareix una llista d'editorials que publiquen llibres en català apareix l'enllaç *Alfonduco Edicions* amb l'adreça següent: <http://www.menorcaweb.net/alfonduco/>. CREN va detectar que aquest enllaç era inactiu, per la qual cosa, després d'haver fracassat la primera estratègia, va buscar a la xarxa una pàgina web que tingués el títol «Alfonduco Edicions» i la va trobar en aquesta adreça: <http://www.alfonducoedicions.com/>. Tenint en compte que és possible que l'usuari no trobi en aquesta pàgina la informació a què es fa referència, CREN també busca l'adreça de pàgines web els títols de les quals contenen el text ancorat. Per exemple, des de <http://www.alfonducoedicions.com/poesia/> i <http://www.alfonducoedicions.com/conte/>, etc. com a enllaços alternatius.

El resultat de la revisió és un fitxer HTML en què, per cada enllaç inactiu, es dona la informació següent: el text ancorat, l'enllaç inactiu, la llista d'enllaços suggerits i el context en què apareix. L'usuari pot fer clic en un enllaç i accedir a la pàgina a què es fa referència per a comprovar que és la pàgina que busca. La llista és ordenada segons el grau de coincidència entre el text ancorat i el títol de la pàgina, i la primera opció és la pàgina el títol de la qual coincideix exactament amb el text ancorat. El context indica a l'usuari on és l'enllaç que ha d'actualitzar per a accedir-hi de manera àgil. També serveix per a triar un enllaç suggerit alternatiu, en cas que l'usuari consideri que l'enllaç presentat com a primera opció no s'ajusta al context en què hauria d'aparèixer en la versió actualitzada.

Com a prova pilot es van analitzar tots els enllaços de l'espai digital Lletra (<http://www.uoc.edu/lletra/>), dedicat a la literatura catalana. El total d'enllaços que es va revisar va ser de 14.000, aproximadament. El cercador d'enllaços CREN va detectar 2.408 enllaços inactius i en va presentar 1967 d'alternatius. El temps

emprat per a fer la cerca va ser d'unes quatre hores.

Els continguts

En el marc del projecte RESTAD, cada servei universitari ha treballat algun dels continguts següents: memòries de traducció i extracció de lèxic d'especialitat per a treballs de terminologia. Els serveis han considerat adequat per al projecte RESTAD alinear documents que poguessin ser útils a totes les universitats de parla catalana i no tan sols a les universitats participants en el projecte. Respecte a la preparació dels documents per a ali-

ció de memòries de traducció s'ha fet a partir de textos revisats als serveis lingüístics, com ara programes d'assignatures, documents docents i documents normatius.

Extracció de lèxic especialitzat

Amb les tècniques de cerca automàtica d'equivalències es poden localitzar els termes equivalents en documents traduïts i obtenir llistes lèxiques bilingües i, per tant, bases de dades de lèxic d'especialitat que, amb la preparació adequada, es poden fer servir en sistemes de traducció assistida i de traducció automatitzada. Aquestes llistes bilingües són de gran

Taula 2. Còmput per àmbits

Àmbits d'especialitat	Tipus de documents	Paraules	Llengües
Ciències socials (dret, economia, filologia, periodisme, política, psicologia) Ciències de la natura (física, geografia, geologia, medicina, química, veterinària) Especialitats tècniques (enginyeria, informàtica, optometria, telecomunicació)	Noms i programes de les assignatures d'estudis de 1r i 2n cicles Exàmens Apunts Pràctiques Problemes Exercicis Descriptors dels nous programes oficials de postgrau màsters propis Descriptors de les fitxes de les assignatures que l'alumnat pot consultar als web de les universitats Descriptors d'assignatures que es publiquen al BOE	704.565	català>espanyol
Art digital i humanitats	Articles de revistes científiques digitals	700.000	català>espanyol català>anglès
Gestió administrativa i acadèmica	Estatuts, noms de departaments i unitats estructurals de les universitats	31.702	català>espanyol català>anglès

near, s'han dut a terme els treballs de descripció dels àmbits d'especialitat, recopilació dels fitxers per a tractar, reanomenament dels fitxers d'acord amb els criteris consensuats de nomenclatura de fitxers aplicable també a la nomenclatura de les memòries de traducció i recompte del nombre de paraules dels documents originals i traduïts.

Memòries de traducció

L'alineació de documents i posterior crea-

ajuda en els processos de postedició de documents traduïts amb el suport d'un motor de traducció per a assegurar la coherència del lèxic d'especialitat i per a la creació de glossaris bilingües especialitzats, siguin en format paper o en format electrònic.

A partir de les sol·licituds de traducció a l'espanyol i a l'anglès que arribaven als serveis lingüístics, s'ha extret lèxic d'especialitat que s'ha considerat productiu per a futures traduccions en algunes àrees de

Taula 3. Entrades

Àmbits d'especialitat	Entrades en català	Entrades en castellà	Entrades en anglès
Intel·ligència artificial	253	—	253
Qualitat	80	80	80
Glossari de l'EEES	226	237	239
Ciències socials	671	460	870
Tecnologia	1348	1.569	2.141
Òptica i optometria	825	816	654

coneixement, com ara les que anotem en la taula següent.

Aquests recursos també poden ajudar els autors de materials docents i documents administratius i acadèmics a millorar els processos de redacció, de correcció d'originals i de preparació dels originals (preedició) que s'han d'enviar al motor de traducció automatitzada, ja que es poden publicar les llistes de lèxic, un cop revisades.

Conclusions

Els grans volums d'informació i documentació amb què es treballa avui dia, juntament amb el reciclatge de parts d'aquests documents (un mateix text pot aparèixer en productes finals diferents), fan possible i necessari alhora potenciar el desenvolupament d'eines que permetin reaprofitar al màxim les feines de traducció i de revisió lingüística. El desenvolupament d'aquestes eines és imprescindible per a garantir que el català mantingui una presència en un context de comunicació global.

L'experiència del projecte RESTAD ha estat molt satisfactòria pel model de treball innovador que ha representat: a) desenvolupar eines informàtiques de suport a la correcció, a la traducció i a l'edició en el si mateix dels serveis lingüístics universitaris i b) explotar els mateixos corpus textuais que es revisen i tradueixen als serveis.

Les proves pilot que s'han fet amb l'aplicació de les eines i l'explotació de continguts demostren que les eines de suport a la traducció i la correcció poden ajudar a relacionar parells de llengües amb més facilitat, estalviar costos i temps en els flu-

xos de treball de correccions i traduccions que fan els serveis lingüístics universitaris i el mateix professorat i reaprofitar continguts a gran escala.

Els resultats del projecte són fruit d'una recerca totalment pràctica i aplicada, ja que en la mesura que ha estat possible s'han incorporat les eines creades en el marc del projecte als fluxos de treball dels tècnics dels serveis lingüístics.

Aquests mètodes nous de treball ajudaran a afrontar alguns dels reptes de futur que l'EEES proposa a les universitats, perquè serviran per a gestionar el multilingüisme de manera més eficaç, contribuir a la normalització del català i garantir-ne la presència en aquest nou marc, fomentar l'autonomia del professorat en l'elaboració de documents, millorar la qualitat final dels documents destinats als estudiants, formar professorat i tècnics en l'ús d'eines informàtiques i donar continuïtat a la política de codi lliure.

Confiam que en el futur el treball cooperatiu que hem iniciat sigui una realitat imparabile i que els òrgans de govern de les universitats facin una aposta decidida per la consolidació d'equips especialitzats en tecnologies lingüístiques (lingüistes i enginyers) que, en un tàndem indissoluble i solidari, treballin junts per fer aportacions innovadores per als mètodes de treball de professionals de la llengua, de redactors i, especialment, de professorat universitari.¹



1. Les autores volen expressar el seu agraïment a Antoni Oliver González, doctor en enginyeria tècnica de telecomunicacions, a Jordi Atserias Batalla i a Albert Romero Sánchez, pel desenvolupament tecnològic aportat.