Salvador Climent · Joaquim Moré · Antoni Oliver · Míriam Salvatierra · Imma Sànchez · Mercè Vàzquez (Barcelona)

Tecnologies de la traducció per a la gestió de la doble oferta docent en català i castellà a la UOC

Introducció

La Universitat Oberta de Catalunya (UOC)* és una universitat plenament virtual que actualment, deu anys després de ser fundada, l'any 1994, ofereix un total de 17 titulacions homologades en català. L'any 2000 la UOC va començar a fer docència universitària també en castellà, amb la inauguració del campus iberoamericà, el qual aplega actualment uns cinc mil estudiants distribuïts en 14 titulacions i unes 400 aules i assignatures.

En un gran nombre de casos, les aules i les assignatures del campus iberoamericà tenen una correspondència directa amb les seves homòlogues del campus principal (el campus en català), és a dir, s'hi imparteixen els mateixos continguts estructurats de la mateixa manera, la planificació del curs i la gestió del temps és equiparable i les pràctiques, les proves i els exàmens són idèntics. En conseqüència, en moltes assignatures, una part important del conjunt de documentació que s'utilitza a l'aula d'un dels campus és bàsicament una traducció del que s'utilitza a la seva aula homòloga, en l'altre campus.

Concretament, per a cada quadrimestre cada aula virtual incorpora els documents següents: un pla docent,[1] de tres a sis guies d'estudi,[2] de tres a sis documents d'enunciat i plantejament de proves d'avaluació contínua (PAC),[3] tres models d'enunciat d'examen i tres models de prova de valida-

* Aquest treball ha estat finançat en part pel projecte A MEDIDA (Programa PROFIT, FIT 350201-2004-6)

1 Pla docent: document que defineix l'assignatura i el procediment docent que s'emprarà per a la seva impartició – continguts, objectius, metodologia, recursos que s'utilitzaran, forma d'avaluació, dates clau, etc.

2 Guia d'estudi: document que recull les pautes de treball de cada una de les proves d'avaluació contínua – objectius, continguts, conceptes més importants, etc. Hi ha una guia d'estudi per a cada prova d'avaluació contínua.

3 PAC: prova d'avaluació contínua, que és avaluable; en molts casos superar les PAC eximeix de fer exàmens.

Zeitschrift für Katalanistik 18 (2005), 31–57
ISSN 0932-2221
https://doi.org/10.46586/ZfK.2005.31-57

ció,[4] amb la qual cosa cada assignatura pot arribar a tenir un total de 19 documents vinculats.

Des de l'obertura del campus iberoamericà fins a l'actualitat les necessitats de traducció de documentació de les aules han anat creixent. Per aquest motiu, l'any 2000 el servei lingüístic de la Universitat va decidir de fer servir el sistema de traducció automatitzada (TA), que la Universitat ja tenia instal·lat des de l'any 1997, per als documents de l'aula que havien de ser traduïts. Per a fer-ho, es van aplicar els mateixos processos de correcció i de traducció que es feien servir per a altres documents docents acadèmics i administratius amb l'objectiu d'augmentar la productivitat del servei substituint la feina de traduir manualment per la de posteditar (revisar) –menys costosa en temps– l'esborrany que resulta d'una traducció automatitzada.

El sistema de traducció automatitzada que es va adoptar fou Comprendium (vegeu Comprendium, 2004), ja que avaluacions prèvies fetes del sistema i l'experiència en la traducció de materials didàctics ja havien determinat que era el que obtenia millors resultats en el parell de llengües català–castellà (en ambdós sentits de la traducció).

Ara bé, aquesta incorporació del sistema de TA no va solucionar el problema, atès que les necessitats de traducció van continuar creixent amb el temps de manera exponencial. Així mateix, l'explotació del sistema de TA presenta diverses mancances, essent les principals les següents: (a) el sistema de TA és de propòsit general i, per tant, no està adaptat a les especificitats lingüístiques i temàtiques dels documents que s'han de traduir; (b) malgrat que molts dels documents del conjunt total són similars i que cada document és molt similar al seu anàleg que s'utilitzarà en la mateixa aula i assignatura en el quadrimestre següent, el sistema de TA, òbviament, repeteix els mateixos errors cada cop que tradueix. Això fa que s'hagi hagut de mantenir i incrementar progressivament l'esforç en la correcció de documentació, un treball de Sísif, ja que la correcció es fa en molts casos sobre els mateixos errors.

Per aquestes raons, la UOC ha posat en marxa a mitjan 2004 un programa de treball, *Traducció Automatitzada a Campus Virtual* (TAaCV), destinat a integrar sistemes de correcció automatitzats i tècniques de traducció automatitzada i de traducció assistida per al català i el castellà primerament

4 Prova de validació: prova final que en moltes assignatures substitueix l'examen per a aquells estudiants que han presentat i aprovat les PAC.

a les aules virtuals i, més endavant, a la totalitat del Campus Virtual i també en múltiples processos interns de treball de la Universitat.

Cal situar com a antecedent d'aquest programa de treball el projecte *Interlingua* (vegeu Interlingua, 2004), en què el mateix equip que actualment treballa en el projecte *TAaCV* que aquí presentem va abordar la problemàtica de la traducció automatitzada no supervisada de missatges de correu electrònic dins el campus de la UOC.

En aquest article farem, en primer lloc, una petita revisió de la problemàtica i les experiències d'aplicació de les tecnologies lingüístiques a la gestió del bilingüisme en organismes i entitats en els territoris de parla catalana (apartat 1), seguidament presentarem el disseny del projecte *TAaCV* (apartat 2) i, finalment, comentarem el treball que hem portat a terme durant el primer any de durada del projecte (apartat 3).

1 Experiències i problemàtica d'aplicació massiva de tecnologies de la traducció en entorns corporatius

En aquest apartat presentarem les experiències de traducció automatitzada en els territoris de parla catalana i veurem com les tecnologies del llenguatge han contribuït en la tasca de normalització lingüística del català. La tasca de normalització comporta un nivell d'exigència de qualitat en les traduccions que, en el cas concret de la traducció automatitzada, depèn molt de la qualitat dels documents originals i d'un treball intens de revisió de les traduccions. Així mateix, exposarem la problemàtica d'aquesta exigència.

1.1 Experiències de traducció automatitzada en entorns corporatius

La situació sociolingüística de la llengua catalana ha contribuït bastant a l'aplicació de tecnologies de la traducció als Països Catalans. La promoció d'aquestes tecnologies està molt relacionada amb la Llei 1/1998, de 7 de gener, de política lingüística, que va substituir la Llei 7/1983, de 18 d'abril, de normalització lingüística a Catalunya. La llei de política lingüística estableix dues obligacions a les institucions i empreses públiques: D'una banda, han d'emprendre mesures perquè el català tingui una posició d'igualtat respecte al castellà en els mitjans de comunicació, la informàtica, la publicitat, el món laboral o empresarial, el sistema educatiu, etc. De l'altra, han de garantir el dret constitucional de tot ciutadà a ser atès tant en català com en castellà i a usar tots dos idiomes.

El compromís de preservar el dret de tot ciutadà a usar el català i a ser informat i atès en català o castellà ha fet que les institucions públiques tinguin un volum important de documentació oficial que tradueixen en les dues llengües. Per exemple, el Diari Oficial de la Generalitat de Catalunya (DOGC) i el Butlletí Oficial de l'Estat (BOE) es publiquen diàriament en català i castellà. Això ha provocat que les institucions esdevinguin usuàries dels sistemes de traducció automatitzada. L'any 1999 la Generalitat, des del Comissionat per a la Societat de la Informació, per mitjà del pla estratègic *Catalunya en Xarxa*, acorda un conveni amb Incyta,[5] empresa formada per integrants de l'equip que a Barcelona desenvolupà el mòdul espanyol del sistema METAL de Siemens i que, gràcies a un conveni amb la Generalitat i la Universitat Autònoma de Barcelona, desenvolupà les direccions català–castellà.[6] Segons el conveni de l'any 1999, el Comissionat adquireix els traductors automàtics desenvolupats per als parells de llengües català–castellà i català–anglès per a ús intern, i també la traducció automatitzada (amb revisió posterior de correctors humans) del BOE al català.

La llei de política lingüística, concretament en l'article 29, que fa referència a les indústries de la llengua i la informàtica, també estableix que les institucions públiques han de promoure i difondre els sistemes de traducció automatitzada que inclouen el català.[7] Per això, la Direcció General de Política Lingüística (DGPL) va donar suport a la iniciativa del Consell dels Il·lustres Col·legis d'Advocats de Catalunya (CICAC) de portar a terme l'any 2001 una prova pilot,[8] que va consistir a posar a disposició dels advocats, de forma gratuïta, els serveis de traducció automatitzada de l'empresa Sail Labs, la qual va adquirir els sistemes d'Incyta.

Fora del Principat, les tecnologies de la traducció també han estat promogudes per l'Administració. L'any 2000 la DGPL va cedir al Govern Balear una llicència del traductor automàtic català–castellà d'Incyta. I, per la seva banda, la Conselleria de Cultura de la Generalitat Valenciana ha desenvolupat el sistema Salt (traductor del castellà–valencià) i ha finançat

5 La descripció del conveni està disponible a: <http://www10.gencat.net/dursi/generados/catala/departament/recurs/doc/3_que_hem_fet.pdf> (pàg. 31).

6 Amb la notació Llengua1–Llengua2 indiquem que el sistema tradueix en les dues direccions.

7 Vegeu <http://www6.gencat.net/llengcat/legis/cat_llei.htm> (versió per a imprimir), <http://www.gencat.net/lleicat/cindex.htm> (versió web).

8 La presentació d'aquesta iniciativa en la III Jornada sobre Llenguatge Jurídic i Administratiu Universitari, que tingués lloc el 4 de juliol del 2002 a Tarragona, està disponible a <http://www.ub.es/slc/cilaj/pagines/jornades/cicac_03.ppt>.

l'empresa Autotrad en la construcció del programa Ara,[9] que és un traductor castellà–català.

Ara bé, en l'empresa privada no hi ha un interès per l'ús de sistemes de traducció automatitzada equiparable al de les institucions, encara que hi ha experiències molt importants com la publicació de la versió catalana del diari *El Periódico*, feta amb traducció automatitzada a partir de l'edició revisada i corregida en castellà. Les empreses situades en els territoris de parla catalana no solen considerar les despeses de traducció com a necessàries, donat que la immensa majoria de catalans són bilingües i que el seu principal mercat és a l'Estat espanyol. Moltes d'aquestes empreses han pres la decisió estratègica d'usar el castellà com la seva llengua corporativa. Una mostra d'això és l'estudi de la WICCAC (Webmàsters Independents en Català, de Cultura i Àmbits Cívics) sobre l'ús del català a Internet fet durant els mesos de juny i juliol del 2004 (vegeu WICCAC, 2004). La majoria dels webs corporatius de més de 400 empreses que operen o són radicats en territori de parla catalana estan en castellà i, si estan traduïts, ho estan en anglès. En canvi, en els webs vinculats a l'Administració publica o semipública la situació és "força correcta" a Catalunya i millorable a les Balears. La impressió més negativa la fa la Comunitat Valenciana.

Amb tot, la internacionalització de l'economia sembla que podria ser un estímul per a activar la traducció automatitzada de la documentació, pàgines web corporatives, etc. a altres idiomes (anglès, francès, etc.), però en aquest marc tampoc no té gaire aplicació la TA, ja que els empresaris, malgrat tenir molt bones referències de la qualitat dels traductors català–castellà, no tenen encara gaire confiança en els traductors automàtics la llengua de destinació dels quals no és el català o el castellà (anglès, alemany, etc.). Per una altra banda, les empreses de serveis de traducció situades als Països Catalans (Incyta, AutomaticTrans) s'han concentrat sobretot en la traducció català–castellà. Només Incyta, que s'ha convertit en representant de l'empresa internacional Comprendium, la qual va adquirir els motors de la desapareguda Sail Labs, pot oferir un servei de traducció o vendre motors en altres parells de llengües.

En aquest sentit, l'Administració és la que aposta clarament per explotar la traducció automatitzada al servei de l'empresa. L'any 2001 va entrar en funcionament el servidor de traduccions automatitzades de la Generalitat de Catalunya, disponible per mitjà de la seva intranet. Amb aquest servei es pot enviar per correu electrònic el text que s'ha de traduir i es rep un

9 <http://www.ara-autotrad.com/>.

esborrany de la traducció en català o castellà. Aquest servei també s'ha encarregat de la traducció al català, al castellà i a l'anglès de webs corporatius. En el cas de les traduccions en català–anglès, es vol obtenir en temps real una traducció de les pàgines en anglès visitades i afavorir les consultes dels continguts en català a la xarxa per persones d'arreu del món.

El servei de lliurament de traduccions és la manera que han trobat els propietaris de motors de traducció automatitzada com Incyta o AutomaticTrans per a treure'n un rendiment econòmic i per a activar el sector de les empreses de traducció i consultoria lingüística. Per exemple, AutomaticTrans té com a clients l'Agència EFE –la qual utilitza el motor de traducció per a generar la versió catalana dels seus butlletins–, l'entitat bancària Caixa de Catalunya o l'Agència Europa Press, entre altres. Ara bé, hi ha una oferta important d'obtenció de traduccions de manera gratuïta: hi ha la possibilitat de baixar d'Internet el traductor Salt i també és possible obtenir traduccions gratuïtes per mitjà del sistema InterNostrum. InterNostrum és un sistema de traducció automatitzada elaborat per la Universitat d'Alacant per encàrrec d'una institució privada, la Caja de Ahorros del Mediterráneo (CAM). InterNostrum funciona com un servidor gratuït d'Internet per als empleats de la CAM, per als de la Universitat d'Alacant i per al públic en general. No és un servei gratuït i obert de l'estil que hom troba en portals d'empreses internacionals de traducció com Systran. Permet obtenir la traducció de grans volums, tot i que, de moment, aquest servei és restringit als empleats de la CAM, als membres de la Universitat d'Alacant i als avaluadors externs del sistema. L'objectiu, però, és convertir el motor en un programari de codi obert, amb la subvenció del Ministeri d'Indústria, Turisme i Comerç de l'Estat espanyol per mitjà del programa PROFIT.

Els creadors i propietaris dels motors també venen el motor i les llicències d'explotació econòmica a les institucions no governamentals i empreses que decideixen gestionar les seves traduccions, perquè els surt més a compte econòmicament que mantenir una relació comercial proveïdor-client. Esmentarem algunes de les experiències més destacables, com ara la de *El Periódico*, que fa servir el traductor d'AutomaticTrans, i la dels serveis lingüístics de la Universitat Autònoma de Barcelona, la Universitat de Girona, la Universitat Oberta de Catalunya i la Universitat Politècnica de Catalunya amb l'adquisició del sistema InfoStore de l'empresa Comprendium, que duu integrat un paquet de memòries de traducció (MT).

Malgrat que en altres països les experiències en l'autogestió de les traduccions amb tecnologies de la llengua són nombroses (vegeu Sprung,

2000), als Països Catalans l'autogestió de traduccions fent servir alhora la combinació de diverses de tecnologies de suport a la llengua (TA + MT + lèxics automatitzats) és una experiència molt recent i, en el cas d'algunes universitats com la UOC, encara està en fase experimental.

1.2 Problemàtica en l'aplicació de la traducció automatitzada

Tothom és conscient que fent servir traducció automatitzada és difícil arribar a un nivell òptim de qualitat. Generalment, qualsevol empresa, institució, etc. que tradueix de manera automatitzada la seva documentació ha d'assumir una despesa en la correcció del text traduït. La qüestió fonamental és saber si, malgrat els costos de correcció del text, l'aplicació de la traducció automatitzada surt més a compte econòmicament que pagar un traductor o un equip de traductors humans perquè facin la traducció manualment.

L'experiència ha demostrat que els costos de correcció disminueixen quan els continguts dels textos que s'han de traduir estan expressats amb el que anomenem *llenguatge controlat*. Un llenguatge controlat és un subconjunt d'un llenguatge natural les gramàtiques i els lèxics dels quals s'han restringit per a reduir o eliminar l'ambigüitat i la complexitat. El paradigma d'aquest tipus de sistemes és Taum-Meteo (Thouin, 1982), dedicat a la traducció de l'anglès al francès de butlletins meteorològics a Montréal des de l'any 1976.

Els llenguatges controlats no sorgeixen amb les aplicacions de la traducció automatitzada. De fet, responen a la necessitat de les empreses i institucions de ser clars en la documentació que han de difondre als clients, als visitants virtuals, etc. Per exemple, l'European Association of Aerospace Industries va decidir escriure la seva documentació en l'anomenat *Simplified English*.[10] En aquest llenguatge controlat el lèxic es redueix a les paraules més freqüents, que són les més conegudes pels emissors no nadius. A més, dels possibles sentits d'una paraula se selecciona el "sentit primer". Pel que fa a la sintaxi, la longitud de les frases es redueix, les estructures són simples, les oracions passives no són freqüents i s'evita tant com es pot l'ús de pronoms.

La comprensibilitat i claredat de les traduccions de textos controlats amb un llenguatge simplificat són altes. Això repercuteix en els costos de correcció de les traduccions. Per això, l'escriptura de documents amb un

10 Per a l'espanyol també hi ha hagut algunes iniciatives (vegeu Cascales / Sutcliffe, 2003: 35).

llenguatge simplificat ha estat una mesura adoptada per empreses, corporacions, etc. que utilitzen sistemes de traducció automatitzada.[11] L'ús de llenguatges simplificats i altres mesures de control de l'*input* comporta sovint que el sistema de traducció automatitzada s'especialitzi en una tasca, com és el cas del Taum-Météo.

En aquest context, la situació de la traducció automatitzada als Països Catalans és prou "especial". Com hem vist, la traducció automatitzada s'aplica principalment en tres àmbits, que són l'administració, els mitjans de comunicació i les institucions acadèmiques. En aquests àmbits és difícil "controlar" el llenguatge dels textos originals, i els sistemes usats no són sistemes especialitzats. Per exemple, el sistema de *El Periódico* no és especialitzat, ben al contrari. Ha de traduir al català notícies d'economia, l'horòscop, ressenyes de llibres, necrològiques o crítiques teatrals. Ara bé, el nivell d'exigència és molt alt: l'error de traduir el nom d'una ministre de la primera legislatura del govern de José Maria Aznar, que es deia Isabel Tocino, com a *Isabel Cansalada* va tenir tanta repercussió pública que, en cas que s'haguessin comès més errors com aquest, la imatge del diari hauria pogut quedar compromesa.

Així mateix, en una aplicació només factible amb un sistema de traducció automatitzada com és la traducció de missatges de correu electrònic dels estudiants i docents de la UOC, el llenguatge dels comunicants és qualsevol cosa menys controlat. El temor de l'equip encarregat d'aquest projecte (projecte *Interlingua*) fou que els errors de traducció produïts per un mal *input* poguessin provocar la incomunicació entre l'emissor i el receptor. Per aquest motiu, van construir mòduls de preedició automatitzada per a evitar-ho.

Quant a l'administració i als organismes legislatius i judicials, malgrat poder-se estudiar si la documentació que generen es pot expressar en un *català simplificat*, ens sembla que tal proposta no tindria gaire èxit. Creiem que en una situació com l'actual, en la qual el català no està encara del tot normalitzat, la iniciativa es podria prendre de manera que l'administració i els organismes judicials contribueixen a l'empobriment de la llengua. En realitat, els serveis lingüístics de les institucions governamentals i acadèmiques han assumit la responsabilitat de donar exemple en la tasca de norma-

11 Per a conèixer algunes experiències, consulteu les actes de la conferència conjunta de la European Association of Machine Translation (EAMT) i la Controlled Language Association Workshop (CLAW) que tingué lloc a Dublín l'any 2003 (<http://www.eamt.org/eamt-claw03/>).

lització i, per això, treballen perquè les publicacions de la institució o de la universitat siguin models de correcció i riquesa expressiva del català. Ara bé, el paper de donar exemple no es limita a l'ús del català, sinó també a la qualitat de les traduccions al castellà, donada la pressió d'uns prejudicis existents dins i fora dels Països Catalans, nodrits en gran part per interessos polítics, com ara que el nivell de castellà dels catalanoparlants és molt pobre i que la política educativa en matèria lingüística no és correcta perquè els ciutadans no tenen una bona competència en cap de les dues llengües.

Sigui perquè el català no és encara una llengua normalitzada o perquè són inevitables les interferències lingüístiques entre el català i el castellà (barbarismes, calcs sintàctics i lèxics d'una llengua o de l'altra), el nivell de llengua dels textos originals en català que passen pel traductor automàtic no és prou bo. Això té sovint conseqüències greus en la traducció automatitzada. Per exemple, si un alumne que escriu un correu electrònic no accentua la paraula *bé* en la frase *espero que aquest email arribi bé*, el traductor traduirà *espero que este mail te llegue cordero*. Segons un estudi realitzat durant el projecte *Interlingua* (Climent *et al.*, 2003), les mancances en la competència lingüística dels estudiants en català provoquen el 48,1% dels errors de traducció del sistema. Lamentablement, les mancances en la competència lingüística no es presenten solament en els estudiants, sinó també en els docents i, en general, en els membres d'una generació que no va ser escolaritzada en català.

El fet paradoxal és que el llenguatge emprat per les institucions que assumeixen ser models de normalització no està normalitzat. Per exemple, en la III Jornada sobre Llenguatge Jurídic i Administratiu Universitari,[12] que tingué lloc a la Universitat Rovira i Virgili l'any 2002, es van explicar mancances en els textos de les administracions locals i les administracions de justícia que s'han de publicar en el DOGC, i també es va expressar la necessitat de normalitzar el llenguatge jurídic per a evitar la barreja de nivells d'expressió, els graus de formalitat, la manca de criteris terminològics, etc. que es manifesten en els textos que s'han de publicar.

Quan les institucions administratives i acadèmiques recorren a la traducció automatitzada, aquesta es converteix en una eina que ha de contribuir a la normalització del llenguatge jurídic, científic, tècnic, etc. Per tant, és fonamental que els organismes oficials vinculats a la normalització lin-

12 En el web <http://www.ub.es/slc/jcilaj.pdf> es poden trobar les actes d'aquestes jornades.

güística com el Termcat, un organisme finançat per la Generalitat que s'ocupa d'establir l'equivalent en català d'un terme científic, tècnic, etc. i que també té un servei de consultes terminològiques, alliberin els seus recursos lingüístics perquè puguin ser explotats directament pels traductors automàtics.

Ara bé, més que el problema de l'explotació dels recursos, la contribució de la traducció automatitzada a la normalització té com a principal obstacle la manca d'implicació personal dels autors dels documents originals en aquesta tasca. En moltes ocasions, l'especialista farceix els seus escrits amb els termes d'ús més comú (sovint en la forma anglesa o castellana) malgrat no ser acceptats pel Termcat. Un sistema que només tradueix els termes acceptats pel Termcat comporta una inversió important per part dels serveis lingüístics de les universitats i de les entitats jurídiques, governamentals, etc. en tasques de correcció. Si la correcció no es fa, els termes no acceptats es deixen sense traduir i, malgrat que molts d'ells són entesos pels lectors de la traducció perquè també els usen, és indubtable que el resultat no fomenta la tasca de normalització. Un altre obstacle important és l'excessiva confiança de l'autor en el seu nivell de competència, tant en català com en castellà. Com que el mateix autor coneix les dues llengües, sovint fa ell mateix la correcció de la traducció automatitzada. Per desgràcia, si l'autor no sap que en l'ús de la llengua de destinació hi afloren moltes incorreccions fruït de les interferències amb la llengua d'origen, l'autocorrecció no és cap garantia de la qualitat lingüística del text traduït.

Amb tot, també s'ha de reconèixer que la bona fama que tenen els traductors automàtics català–castellà quant a la seva qualitat és paradoxalment un inconvenient. Sovint l'autor d'un document cau en la temptació d'enviar a publicar la traducció automatitzada sense haver-la revisat ni ell ni cap especialista.

2 Plantejament del projecte de traducció automatitzada

En els serveis de traducció estructurats entorn d'un motor de traducció automatitzada, com els que ofereixen empreses com ara Incyta o AutomaticTrans, un grup de lingüistes i traductors experts revisen la qualitat en les traduccions de webs corporatius, fullets publicitaris, etc. La relació comercial entre el client i el servei és de dependència i no és prou flexible quan la generació de traduccions és massiva. L'empresa o la institució client d'aquests serveis no poden crear bases de dades terminològiques pròpies ni afegir nous termes si no paga al servei perquè li ho faci.

En canvi, a la UOC, com que és decisiva la generació massiva de material didàctic en dues llengües, l'única opció que hi ha és comprar el motor de traducció i integrar-lo en el flux de producció de publicacions de la Universitat i en el flux de funcionament del Campus. Així, la UOC ha d'autogestionar les traduccions de manera semblant a com ho fa una empresa multinacional, com EPSON.

Això significa que la UOC, com qualsevol entitat que integra la traducció en el seu flux productiu, ha de gestionar la creació d'MT i bases de dades terminològiques (BDT), ha d'establir uns procediments de revisió i ampliació de les MT i les BDT, i també ha d'impulsar l'automatització de les tasques de postedició (correctors gramaticals i ortogràfics, detectors d'incoherències terminològiques i estilístiques) amb la finalitat d'augmentar la producció. En aquest sentit, el fons de documentació pròpia en format electrònic és fonamental per a fer buidatges i crear MT i BDT. Així, d'aquest fons es poden extreure models d'ús de la llengua que poden ser útils per a la creació d'eines de postedició.

Així mateix, s'han d'establir procediments de gestió de projectes (traducció de grans volums) i de coordinació entre el conjunt de traductors que intervenen en un projecte. I, una cosa molt important, tots els autors dels materials han de seguir escrupolosament els procediments establerts per a evitar les traduccions fetes de manera espontània i que poden ser incorrectes.

2.1 Definició del projecte

En la definició del projecte *Traducció Automatitzada a Campus Virtual* hi han participat els equips del Servei Lingüístic de la UOC, del grup operatiu Desenvolupament d'Intranets i el professorat i investigadors dels estudis d'Humanitats i Filologia Catalana. Anteriorment aquests grups havien col·laborat en el projecte *Interlingua*, dedicat fonamentalment a la traducció de missatges de correu electrònic, i fruit de l'experiència adquirida en aquesta col·laboració és la posada en marxa del projecte *TAaCV*. La primera conseqüència –i també premissa necessària– de l'inici del projecte fou la incorporació al Servei Lingüístic de dos especialistes en lingüística computacional.

Els objectius bàsics que s'han proposat d'assolir amb el projecte són dos: en primer lloc, garantir la qualitat lingüística dels textos tractats amb el sistema de traducció automatitzada, de manera que es pugui reduir al mínim la intervenció humana necessària per a obtenir un text amb la qualitat

òptima per a ser publicat; en segon lloc, oferir al personal de la Universitat que és usuari del sistema de TA la formació i les eines informàtiques i de consulta necessàries per a facilitar-li, tant com sigui possible, la gestió i el tractament lingüístic previ (preedició) i posterior (postedició) a la traducció automatitzada.

En una segona fase del projecte s'ha plantejat el fet de generalitzar l'ús dels sistemes de traducció automatitzada en la comunicació en línia i estàtica del Campus Virtual per tal de facilitar l'entesa entre les diferents comunitats lingüístiques presents a la UOC.

Des d'un punt de vista estratègic, l'organització espera que l'assoliment d'aquests objectius comportarà diversos beneficis per a la UOC, com ara esdevenir pionera entre les universitats en la implantació de sistemes de correcció i traducció automatitzats en la gestió interna i l'acció docent; agilitar el procés de treball intern de la Universitat i de l'acció docent amb la incorporació d'eines de tractament lingüístic automatitzades; oferir prou recursos perquè la qualitat dels textos sigui la que correspon a l'àmbit universitari; obtenir un corpus terminològic bilingüe, organitzat per especialitats, susceptible de ser tractat per a altres usos, fruit de la tasca docent i de recerca que es duu a terme a la Universitat i com a culminació del treball de normalització de la llengua catalana com a llengua tècnica i d'especialitat que han comportat els material didàctics editats, i també obrir el camí de treball per a incorporar en un futur circuits paral·lels per a altres parells de llengües.

2.2 Línies d'actuació

En el conjunt d'actuacions necessàries per a dur a terme la incorporació de sistemes de correcció i traducció automatitzats als processos de treball i comunicació de la Universitat, s'han distingit dues línies d'actuació diferents: primer, l'obertura del sistema de traducció català–castellà al personal de gestió i professorat propi i, segon, la integració de sistemes de correcció i traducció automatitzats als espais de comunicació en línia del Campus. Seguidament parlarem només de la primera línia d'actuació, ja que és la que està en procés de desenvolupament; la segona és una línia de treball futur.

El procés d'obertura del sistema de traducció automatitzada al personal de gestió i professorat de la UOC s'ha fet de manera progressiva, d'acord amb les necessitats i prioritats dels diferents grups i els processos de treball de la Universitat, i també tenint en compte les premisses següents:

a) Col·lectiu d'usuaris
   - Personal de gestió
   - Professorat propi

b) Circuit d'accés al sistema de traducció automatitzada
   - Accés al sistema de traducció per bústies lògiques
   - Centralització de la gestió en la figura d'un responsable per grup o procés de treball

c) Programari necessari
   - Detector de llengua
   - Corrector de català i de castellà (automatitzat o amb intervenció) per a aplicacions i programes que no en tenen
   - Corrector del Word de català i castellà
   - Motor de traducció català–castellà

d) Formació
   - Sessions de formació al personal implicat sobre les característiques i els requisits dels sistemes de traducció automatitzada a càrrec de les tècniques del Servei Lingüístic

e) Documentació susceptible de tractar
   - Tipus: institucional, acadèmica i administrativa (externa al Campus i integrada al Campus)
   - Formats d'origen: doc, html, xls, ppt i aplicacions a mida integrades al Campus, etc.[13]
   - Registre: formal
   - Tractament previ: sense correcció

L'accés del personal de gestió i professorat propi de la Universitat al sistema ha estat precedit per unes sessions de formació sobre les característiques i els requisits del procés de treball que comporta l'ús de sistemes automatitzats de traducció i correcció, i sobre l'ús i el funcionament automatitzat de les eines que s'apliquin al procés, tant si són eines existents al mercat (corrector del Word) com si es tracta d'eines desenvolupades específicament per al projecte. El Servei Lingüístic de la UOC ha estat l'encarregat de preparar i impartir les esmentades sessions i d'assessorar el personal tècnic de la Universitat en tot moment.

13 Cal tenir en compte que el sistema de traducció només admet l'entrada dels formats doc, rtf, txt i html, per la qual cosa cal adequar el sistema perquè permeti l'entrada de textos procedents d'altres formats (com són les aplicacions de treball internes i externes al Campus).

El procés d'obertura del sistema de traducció ha estat possible gràcies al desenvolupament d'eines complementàries i específiques, com ara un programa de detecció de llengua, un corrector de català i castellà, un extractor de terminologia, un cercador d'equivalents de traducció i eines d'ajut a l'edició. Aquestes eines complementàries les presentem en l'apartat 3 d'aquest article.

3 Desenvolupament del projecte de traducció automatitzada

En el moment d'escriure aquest article, el projecte *TAaCV* fa deu mesos que funciona. La feina feta fins ara queda organitzada en cinc grans apartats: organització, eines, bases de coneixement, formació i avaluació. A continuació comentarem les tasques que s'han portat a terme en cada un d'aquests apartats.

3.1 Organització

La traducció de la documentació de la Universitat, amb el suport del sistema de traducció automatitzada, segueix un procés de treball molt pautat per a aconseguir que els textos traduïts tinguin una qualitat final òptima i puguin ser publicats. Aquest procés de treball compta amb el guiatge del Servei Lingüístic.

3.1.1 Definició del flux de treball

El Servei Lingüístic de la Universitat rep els textos originals abans de ser publicats, els quals passen sempre per un procés de correcció, que és fet per un professional de la llengua. Quan el document original ja s'ha corregit, aleshores s'envia al sistema de traducció automatitzada per mitjà d'una bústia lògica. Al cap d'uns quants minuts, el sistema de traducció torna un document traduït a l'altra llengua, que s'ha de considerar com a esborrany de traducció.

L'esborrany de traducció que torna el sistema de traducció automatitzada s'ha de revisar amb atenció tenint en compte sempre el document original corregit. El procés que s'ha de seguir és el següent: primer, comprovar que el sistema ha traduït tot el text acarant l'original i la traducció; segon, restituir el format de l'original en cas que s'hagi perdut, i, tercer, corregir el text traduït (posteditar-lo). La tasca de corregir l'esborrany de traducció la fa un professional de la llengua.

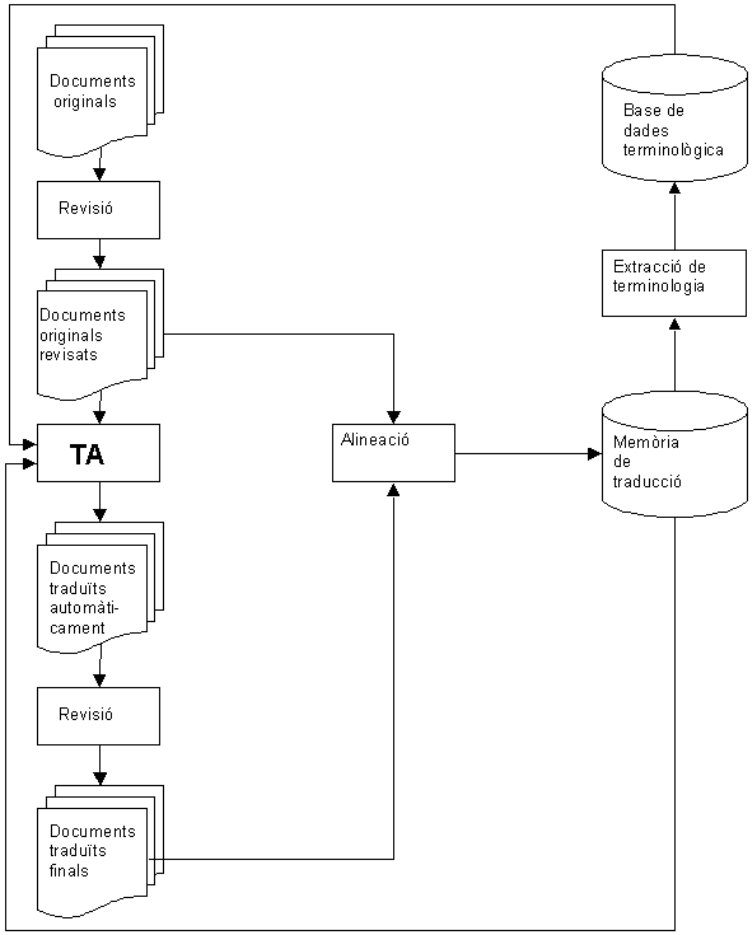

Figura 1: Esquema del flux de treball

A partir del moment en què hi ha un document original corregit i la traducció corresponent supervisada, el pas següent és el procés d'alineació de tots dos documents amb l'objectiu de crear BDT i MT que serveixin per a alimentar el sistema de traducció automatitzada. Així s'obté una resposta més afinada del sistema de traducció automatitzada i es redueix l'esforç de supervisió manual de l'esborrany de màquina. Podeu veure un esquema d'aquest flux a la figura 1.

3.1.2 Integració de la traducció automatitzada a les aules

Amb l'obertura del sistema de traducció automatitzada al professorat propi de la Universitat i en el marc del projecte *TAaCV*, s'ha volgut facilitar l'ús d'aquest sistema integrant el servei de traducció automatitzada a les aules del Campus.

Actualment, des de l'aula virtual del Campus els professors poden redactar els documents que han d'enviar als estudiants –plans docents, guies d'estudi, proves d'avaluació contínua i altres documents relacionats amb la docència– en una llengua i, quan els tenen revisats, per mitjà d'un botó que hi ha a l'aula mateix els poden enviar al sistema de traducció. Al cap d'una estona reben el text traduït en l'altra llengua. A partir d'aquí, el professor fa la revisió de l'esborrany de màquina i després envia el document traduït als estudiants.

3.2 Eines

Amb l'objectiu d'automatitzar les tasques de preedició i postedició dels documents de la Universitat i millorar el procés de traducció automatitzada, en el marc del projecte *TAaCV* s'han desenvolupat o millorat una sèrie d'eines informàtiques específiques. A continuació expliquem amb més detall cada una d'aquestes eines.

3.2.1 Detecció automàtica de llengua

El programa de detecció de llengua es va desenvolupar per a ser aplicat al sistema de traducció automatitzada de missatges de correu electrònic. Com a pas previ a la traducció d'un missatge cal determinar la llengua en què és escrit per a poder-lo enviar al corrector de la llengua original i a la direcció de traducció correcta. El programa funciona fent una estadística d'unigrams, bigrams i trigrams (combinacions d'un, dos i tres caràcters) del text d'entrada i la compara amb uns models de llengua existents. El programa desenvolupat a la UOC és una adaptació del programa TextCat desenvolupat per Gertjan van Noord,[14] que es basa en l'algorisme de Cavnar i Trenkle (1994).

14 <http://odur.let.rug.nl/~vannoord/TextCat/>.

3.2.2 Corrector ortogràfic

Hi ha diversos correctors ortogràfics de català, però aquests no sempre es poden integrar de manera senzilla a les aplicacions del Campus Virtual. Per aquest motiu es decidí desenvolupar un corrector ortogràfic de català, castellà i anglès. D'entrada, el requisit fonamental és que el corrector pugui treballar tant de manera automàtica, sense intervenció de l'usuari, com de manera interactiva. En el mode interactiu el corrector ortogràfic marca les paraules que contenen errors ortogràfics i proposa una sèrie d'opcions a l'usuari. La correcció, doncs, es deixa en mans de l'usuari, que pot acceptar o rebutjar les opcions proposades pel corrector ortogràfic. En el mode totalment automàtic el corrector ortogràfic fa la correcció sense cap intervenció de l'usuari. En aquest cas, només es corregeixen els errors ortogràfics dels quals es té una gran seguretat que són erronis i, a més, se sap quina és l'alternativa correcta.

El corrector ortogràfic compara totes les paraules del text amb una llista de paraules (flexionades en totes les seves formes) considerades correctes. Com sabem, una llista de paraules mai no es pot considerar completa, motiu pel qual hi ha un diccionari d'usuari per a afegir-hi noves formes. L'algorisme de cerca de les opcions inclou una sèrie de ponderacions que s'han calculat a partir de l'anàlisi dels errors més freqüents i de la distància entre les tecles d'un teclat estàndard.

Un exemple de funcionament del corrector ortogràfic pot ser aquest:

"Aixo és un exemple de correccio de misatges de correu electronic"

El mòdul de correcció ortogràfica de català retorna l'esquema següent:

aixo	això:aixe:ixo:així:baixo:pixo:aixà:aixol:mixo:aixa:fixo
correccio	correcció
electronic	electrònic
misatges	missatges:visatges:miratges

Així, el mòdul de correcció retorna les paraules incorrectes amb una llista de possibles opcions de paraules correctes. La primera de la llista és la més probable, de manera que en el model de correcció sense intervenció d'usuari s'agafa com a vàlida cada una de les primeres opcions. Després d'haver-hi passat el corrector, el text corregit sense intervenció de l'usuari queda de la manera següent:

"Això és un exemple de correcció de missatges de correu electrònic"

El corrector conté una llista de canvis *ad-hoc*, és a dir, una sèrie de paraules o grups de paraules incorrectes amb el seu equivalent correcte. Si el programa es troba amb alguna entrada de la llista de canvis *ad-hoc*, farà la substitució directament en mode automàtic o mostrarà la correcció en mode interactiu. Aquest procés és previ al de correcció ortogràfica a partir del diccionari de formes. A continuació mostrem una sèrie d'exemples d'entrades de la llista de canvis *ad-hoc*.

desde:des de
donguès:donés
dongués:donés
escriguent:escrivint
poguer:poder
y:i

3.2.3 Eines de tractament de la terminologia

Per a poder treballar amb els termes dels diversos àmbits d'especialitat de la Universitat s'ha desenvolupat un programari específic per a l'extracció automàtica de terminologia (el programa s'anomena *n-grams*) i la cerca automàtica d'equivalents de traducció en un corpus paral·lel (el programa s'anomena *tond*). L'objectiu d'aquests programes és extreure una sèrie de candidats a terme d'un text o conjunt de textos. El funcionament és bàsicament estadístic i requereix molt poc coneixement lingüístic per part del sistema, únicament compta amb una llista de paraules buides o *stop-words* (paraules que no poden aparèixer en primera o darrera posició d'un terme). Amb l'ajut del programa *n-grams* s'estan confeccionant glossaris terminològics específics de l'àmbit de la UOC. El programa *tond* permet trobar els equivalents de traducció dels termes extrets. El programa ha de tenir un corpus paral·lel[15] per a poder fer una sèrie de càlculs estadístics de cara a determinar quina és la paraula o grup de paraules que presenta un índex més alt de probabilitat de ser la traducció del terme que cerquem.

Actualment aquests programes estan implementats en Perl, però s'està treballant en una versió en Java que unifica tots dos programes en un de

15 Corpus de frases o segments en una determinada llengua amb la corresponent traducció a una altra llengua.

sol i que presenta una interfície gràfica d'usuari més agradable i fàcil de fer servir. Aquests programes són de lliure distribució i es poden descarregar de la pàgina web del grup de lingüística computacional de la UOC.[16]

3.2.4 Eines d'ajuda a l'edició

A la UOC s'estan creant diverses eines d'ajuda a l'edició que tenen com a finalitat facilitar la revisió de documents, ja siguin originals o traduccions. En aquest sentit, com que molts materials de la Universitat es caracteritzen per tenir un nombre significatiu d'enllaços a pàgines web i periòdicament s'han de revisar per a verificar si encara estan actius i mantenen la mateixa adreça, s'ha creat l'eina Cercador d'enllaços per a fer més àgil aquesta tasca. El Cercador d'enllaços comprova si els enllaços estan actius; si no ho estan, fa una cerca a Internet per a trobar la nova adreça de l'enllaç. Aquesta eina permet un estalvi important de temps en la revisió dels documents.

Així mateix, s'estan dissenyant eines d'edició que facilitin la revisió terminològica dels documents, mitjançant cerques automàtiques a BDT, i que detectin la presència de segments existents a les MT.

3.3 Bases de coneixement

El gran volum de traduccions que es fa a la Universitat Oberta de Catalunya fa que sigui de gran importància mantenir una sèrie de bases de dades que continguin les frases o segments originals amb les traduccions corresponents. Aquestes bases de dades, com hem dit, s'anomenen *memòries de traducció*. Per aquest motiu, s'ha automatitzat el procés de creació d'aquestes memòries i es comença a fer un projecte interuniversitari de col·laboració en la creació, manteniment i explotació conjunta d'aquests valuosos recursos.

3.3.1 Creació de memòries de traducció

El concepte de memòria de traducció és similar al de corpus paral·lel, és a dir, és una base de dades de frases o segments en una determinada llengua amb la traducció corresponent a una llengua o a més d'una. Per a crear MT a partir de documents originals i la seva traducció cal alinear-los prèviament, és a dir, relacionar cada una de les frases o segments originals amb la

16 <http://www.uoc.edu/in3/interlingua/>.

traducció corresponent. Aquesta tasca es pot fer manualment amb uns programes que proporcionen un entorn amigable, però a causa del gran volum de documentació generat a la UOC el plantejament manual d'aquesta tasca és inviable. Per aquest motiu, investiguem diferents tècniques i algorismes existents: Moore (2002) i Melamed (1996). Actualment hem aplicat amb èxit l'algorisme de Moore per a aquesta tasca i treballem en el desenvolupament d'una aplicació informàtica pròpia adaptada a les necessitats específiques de la Universitat.

El procés d'alineació que estem desenvolupant es pot dividir en els passos següents: transformació automàtica del format original del document a format text, segmentació dels documents en frases, anàlisi de la repetitivitat, alineació dels documents, creació del fitxer dels segments no alineats i creació de la memòria de traducció. El fitxer de segments no alineats es pot fer servir per a alinear els segments que no s'han pogut alinear automàticament. El percentatge de segments no alineats, tenint en compte els documents que estem tractant, és baix (de l'ordre del 3,5%).

El programa d'alineació genera MT en format text separat per tabuladors i en format TMX (Translation Memory eXchange). Aquests formats permeten l'ús de les MT tant en el sistema de traducció automatitzada Comprendium com en la majoria d'aplicacions de Traducció Assistida per Ordinador.

3.3.2 Projectes interuniversitaris

Els serveis lingüístics universitaris (SLU) en la seva qualitat d'usuaris habituals de programes de suport a la llengua es plantegen la possibilitat d'establir convenis interuniversitaris de col·laboració amb l'objectiu de fer explotacions col·lectives de corpus de documentació universitària (docent, acadèmica i administrativa) que passen pels circuits de correcció i de traducció dels serveis.

Aquests corpus contenen informació similar que és susceptible de ser tractada massivament per a obtenir MT i llistes lèxiques que, amb el vist-i-plau de professionals de la llengua, es poden incorporar als motors de traducció automatitzada per a donar més rendibilitat a l'esborrany que ofereix el sistema després de fer la traducció automatitzada.

Una altra característica d'aquests corpus és que són textos que han passat per processos de correcció, de revisió i de comprovació del lèxic d'especialitat. Si l'original que s'envia al sistema de traducció ha estat revisat pel que fa a ortografia, estructures sintàctiques i puntuació, el resultat

de la màquina és molt superior i el temps d'inversió en la postedició –procés de comprovació amb l'original i de correcció de l'esborrany de màquina– és molt inferior al temps que s'hi ha de dedicar quan l'original no s'ha revisat.

Una de les línies d'actuació que les universitats comencen a plantejar-se actualment és la possibilitat de traspassar a un dels serveis lingüístics volums importants de textos perquè pugui preparar MT i llistes lèxiques que després es puguin retornar a les altres universitats per a incorporar a cada un dels motors de traducció. Si es treballa amb un bon original i una bona traducció revisada, la creació d'MT (frase original–frase traduïda) i de lèxics amb equivalències és una tasca molt rendible.

En l'anàlisi de requisits dels projectes interuniversitaris es detecta la necessitat de classificar els corpus que s'han de tractar en quatre nivells, que respondrien a la situació del text original i del text traduït relacionat amb el procés de correcció. Vegeu-ne un exemple a la taula 1.

	document original	*document traduït*
Nivell 1	Sí corregit	Sí posteditat
Nivell 2	No corregit	Sí posteditat
Nivell 3	Sí corregit	No posteditat
Nivell 4	No corregit	No posteditat

Taula 1: Nivells de corpus

3.3.3 Classificació de les àrees de coneixement

Per a assolir uns nivells de rendibilitat més elevats dels motors de traducció és recomanable marcar les MT i les llistes lèxiques que hem esmentat per àmbits d'especialitat. Per exemple, el sistema Comprendium té una classificació temàtica incorporada que bàsicament recull tres àmbits: el de la llengua general, el de les ciències socials i el de la tecnologia. Dins els àmbits de ciències socials i de tecnologia hi ha subàmbits que permeten classificar especialitats amb l'objectiu que el motor de traducció busqui primer en el lèxic de la subespecialitat i, si no el troba, que busqui en els àmbits superiors, quan s'envia un text a traduir marcat amb un àmbit d'especialitat.

D'altra banda, en un treball coordinat des dels serveis lingüístics universitaris, s'ha fet una classificació de les àrees que corresponen a les carreres que s'ofereixen a les universitats i s'ha arribat a classificar fins a vuit àmbits diferents. Aquesta classificació de les grans matèries universitàries s'ha

adaptat a la classificació temàtica del sistema de traducció automatitzada incloent-la en els grans àmbits de ciències socials i de tecnologia. Vegeu-ne un exemple a la taula 2.

Classificació del motor de traducció	*Classificació dels SLU*
3.4. Electrical Engineering	2.5. Sector industrial
3.4.1. Semiconductor Technology	2.5.1. Enginyeria industrial
3.4.2. Electrical Engineering, User Entries	2.5.2. Indústries diverses
3.5. Mechanical Engineering	
3.5.1. Mechanical Engineering, User Entries	

Taula 2: Exemple de classificació temàtica

L'objectiu d'adaptar la classificació dels SLU a la del motor de traducció ha estat poder tractar els documents que s'envien a traduir i les MT que posteriorment es poden crear amb els mateixos ítems temàtics. Al mateix temps, l'esforç d'unificació d'àmbits temàtics universitaris ajudarà a poder tractar grans volums de documents sota els mateixos ítems i facilitarà l'ús de les MT en diferents universitats que ofereixin les mateixes carreres; per exemple, les MT de documents de Dret civil o de Sistemes de programació es poden compartir en diferents serveis lingüístics per als seus processos de correcció i de traducció.

3.4 Formació

En el procés d'integració del sistema de traducció automatitzada al Campus s'han portat a terme sessions de formació adreçades al personal del Servei Lingüístic i també al professorat de la Universitat i al personal de gestió, com a futurs usuaris del sistema.

Les sessions de formació per al personal del Servei Lingüístic s'han organitzat en tres blocs: la gestió de la terminologia, el lèxic en un sistema de traducció automatitzada i la informació lèxica en el sistema Comprendium. En aquestes sessions s'han volgut presentar els conceptes bàsics sobre gestió de la terminologia, introduir les tècniques d'extracció automàtica de terminologia, presentar les idees fonamentals sobre la informació lèxica que necessita un sistema de traducció automatitzada i presentar l'eina LexShop de Comprendium.

Pel que fa a la terminologia, s'ha treballat específicament l'organització i classificació de les BDT, la creació de glossaris específics d'un projecte, la

recopilació de terminologia i l'extracció automàtica de terminologia. Com a programari específic per a treballar la terminologia s'ha disposat del TermBase de ForeignDesk, de l'*n-grams* per a l'extracció automàtica de terminologia i del *tond* per a la cerca automàtica d'equivalents de traducció en un corpus paral·lel.

Quant al lèxic, s'han comparat les gramàtiques descriptives i normatives amb les gramàtiques formals i generatives; s'ha treballat l'aplicació computacional de les gramàtiques formals i generatives, és a dir, les gramàtiques d'unificació, i s'ha presentat la manera de codificar la informació lingüística en una gramàtica d'unificació: estructura de trets, unificació, informació lèxica i sintàctica en una única representació i el pas del lèxic a la sintaxi i viceversa.

Pel que es refereix a la informació lèxica en el sistema Comprendium, s'ha comprovat que funciona com un sistema de transferència fent servir les gramàtiques d'unificació, s'ha vist que els lèxics són essencials perquè el sistema pugui analitzar correctament les estructures de la llengua i s'ha treballat amb el LexShop, eina que emmagatzema informació lèxica d'una llengua i les relacions d'una llengua amb una altra.

Les sessions de formació adreçades al professorat i al personal de gestió de la Universitat han presentat el funcionament de la traducció automatitzada i també la manera de preparar els textos originals i com s'ha de treballar l'esborrany de màquina, és a dir, quins aspectes formals i gramaticals s'han de revisar. Pel que fa als aspectes formals s'ha explicat quin sistema de marques de colors fa servir el sistema per a identificar les paraules que no tradueix, com s'ha de revisar la traducció dels noms propis i com s'han de marcar els textos escrits en una llengua diferent de la del text original. Quant als aspectes gramaticals, s'han presentat els errors més freqüents que es poden trobar en els esborranys de màquina: construccions verbals errònies, preposicions que no corresponen a la construcció de la llengua de destinació, confusió de pronoms relatius i interrogatius i de conjuncions, pronominalització incorrecta, construccions calcades de la llengua original, entre altres.

Els objectius d'aquestes sessions han estat presentar la nova eina de suport a la traducció al conjunt de professionals de la Universitat, veure quines millores introdueix en el procés de traducció dels documents de la institució i valorar el perill que representa publicar textos sense haver-ne fet la preedició (correcció del text original) i la postedició (correcció del text traduït).

3.5 Avaluació

El procés d'avaluació del sistema de traducció automatitzada permet saber quin és el resultat inicial de les traduccions que ofereix el sistema (avaluació inicial) i quin és el grau de millora en les traduccions després d'haver-hi introduït noves entrades lèxiques, MT, etc. (avaluació final). Aquest procés es fa paral·lelament a l'explotació del sistema. Actualment ja s'ha fet l'avaluació inicial del sistema, a partir de la qual es valorarà quin percentatge d'encerts i d'errors té i es plantejarà amb quines noves entrades lèxiques i MT es va engruixint el corpus.

A l'hora d'avaluar un sistema de traducció cal decidir si convé més fer una avaluació humana manual, una avaluació automàtica o una combinació de les dues. El sistema de traducció automatitzada que fa servir la UOC és avaluat amb mesures recollides automàticament però amb una mínima intervenció humana. Aquesta opció combinada és la millor perquè és més curta, més econòmica i més objectiva que l'avaluació manual i, a més, els resultats són més fiables que no pas els de la traducció totalment automàtica (Tomàs *et al.*, 2003).

L'avaluació del sistema s'ha fet a partir d'un corpus d'originals i un corpus de referència que consta de mil frases. El criteri per a considerar una traducció de referència és el següent: traducció humana d'un original corregida ortogràficament i sense errors de picatge. La unitat de traducció que s'avalua és l'oració. La persona que fa l'avaluació indica les paraules correctes de la traducció que s'avalua i que es desvien de la traducció de referència i afegeix la nova traducció al conjunt de traduccions de referència.

Per a fer l'avaluació s'ha emprat l'eina EvalTrans, desenvolupada per Tomàs *et al.*, en la qual la persona que fa l'avaluació veu les paraules de la traducció que ha d'avaluar i que no surten en la frase de referència (detectades amb mètodes automàtics) i ha d'indicar si cada paraula és un error o forma part d'una traducció alternativa de referència. D'aquesta manera, els valors negatius de la bondat de la traducció que s'avalua aconseguits automàticament es van ajustant cada vegada més i la nova traducció de referència es pot utilitzar en posteriors avaluacions. L'eina també permet a qui fa l'avaluació puntuar del 0 al 10 la qualitat de la traducció (vegeu la figura 2). D'aquesta manera es poden establir correlacions entre els valors numèrics aconseguits automàticament i els valors subjectius de la persona que avalua.

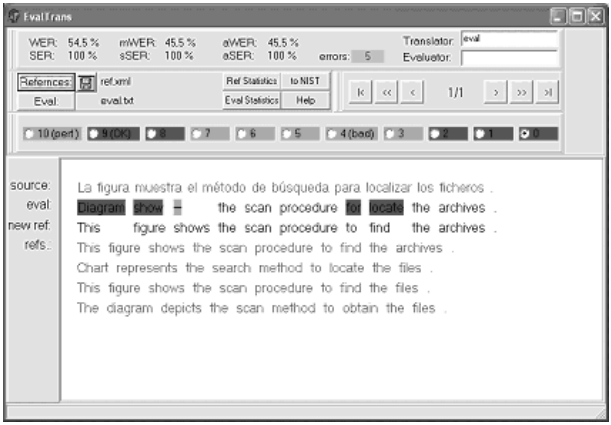

Figura 2: Exemple d'avaluació amb EvalTrans

4 Conclusions i treball futur

En aquest article hem presentat el disseny del projecte *Traducció Automatitzada a Campus Virtual* i el treball que s'hi ha realitzat fins a finals de 2004. Aquest projecte aplega un conjunt d'accions, protocols i creació i aplicació d'eines i tècniques per a l'automatització de la traducció entre català i castellà (en ambdues direccions) en el Campus Virtual de la Universitat Oberta de Catalunya. Aquest projecte s'ha posat en marxa a causa de l'augment exponencial de les necessitats de traducció de documentació docent a la UOC, ja que aquesta universitat imparteix diverses titulacions en les dues llengües amb documentació paral·lela a través del seu Campus Virtual.

En pocs mesos, s'ha dissenyat el projecte (tenint en compte els col·lectius d'usuaris implicats i les seves necessitats de formació, els tipus de documentació a tractar, les modalitats d'accés per part dels usuaris al sistema de traducció automatitzada, les necessitats de manteniment i optimització d'aquest sistema i el programari necessari per a dur-les a terme), s'ha definit el flux de treball i s'ha creat una classificació temàtica per a organitzar els textos, s'ha integrat el sistema de traducció automatitzada a la interfície d'accés que tenen els professors a les aules virtuals, s'han creat diverses eines (un detector de llengua, un corrector ortogràfic, un extractor de terminologia i un cercador d'equivalents de traducció en un corpus paral·lel), se n'estan dissenyant d'altres i s'estan creant MT per a l'alimentació del sistema integrat de traducció automatitzada i traducció assistida.

L'objectiu en els propers mesos és acabar la implementació de les eines començades, integrar-les en una *suite* per a posar-les a disposició dels professionals de la llengua de la Universitat, i sobretot, crear bases de dades de memòries de traducció i de lèxic terminològic per a tot el ventall de tipologia documental i temàtica de la docència de la UOC i alimentar amb elles el sistema de traducció automatitzada d'acord amb l'organització temàtica a fi d'obtenir traduccions cada cop més fiables i adequades a l'àmbit de coneixement al qual pertany el text. Com a pas previ, està previst realitzar una avaluació sobre un àmbit temàtic específic que determini el grau d'augment de la qualitat de la traducció un cop aplicades les millores abans indicades. Evidentment, el resultat d'aquesta avaluació pot fer reorientar l'estratègia del projecte.

D'altra banda, s'està treballant en la construcció d'un projecte conjunt amb els serveis lingüístics de diverses universitats dels territoris de parla catalana per a la creació i compartició de recursos en l'àmbit de la traducció, correcció i revisió de textos.

A més llarg termini, el projecte *TAaCV* preveu integrar sistemes de correcció i traducció automatitzades als espais de comunicació del Campus Virtual de la UOC (fòrums, missatgeria electrònica, etc.).

Referències

Cascales, Remedios Ruiz / Sutcliffe, Richard (2003): "A Specification and Validating Parser for Simplified Technical Spanish", in: *Proceedings of the EAMT-CLAW 200*, Dublin: EAMT, 35–44.

Cavnar, William B. / Trenkle, John M. (1994): "N-Gram-Based Text Categorization", in: *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*. Las Vegas: UNLV Publications / Reprographics, 11–13, 161–175.

Climent, Salvador / Moré, Joaquim / Oliver, Antoni / Salvatierra, Míriam / Sànchez, Imma / Taulés, Mariona / Vallmanya, Lluïsa (2003): "Bilingual Newsgroups in Catalonia: A Challenge for Machine Translation", *Journal of Computer-Mediated Communication* 9:1, <http://www.ascusc.org/jcmc/vol9/issue1/climent.html>.

Comprendium (2004): Pàgina web de *Comprendium España* S.L.: <http://www.comprendium.es> [data de consulta: 8 d'octubre de 2004].

Diz, Inés (2001): "The importance of MT for the survival of minority languages: Spanish-Galician MT system", in: *Proceedings of the MT Summit 2001*, Santiago de Compostel·la: EAMT, 89–92.

Interlingua (2004): Pàgina web del projecte *Interlingua*: <http://www.uoc.edu/in3/interlingua> [data de consulta: 8 d'octubre de 2004].

Melamed, Dan (1996): "A Geometric Approach to Mapping Bitext Correspondence", in: *IRCS Technical Report #96-22*, a revised version of the paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, <http://www.cs.nyu.edu/~melamed/pubs.html> [data de consulta: 8 de març de 2005].

Moore, Robert (2002): "Fast and Accurate Sentence Alignment of Bilingual Corpora", in: *Machine Translation. From Research to Real Users* (Proceedings 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California), Heidelberg: Springer, 135–244.

Sprung, Robert C. (ed. 2000): *Translating Into Success. Cutting-edge strategies for going multilingual in a global age* (ATA Scholarly Monograph Series; 11), Amsterdam / Philadelphia: John Benjamins.

Thouin, Benoît (1982): "The Meteo System"; in: Lawson, Veronica (ed.): *Practical Experience of Machine Translation*, Amsterdam: North-Holland, 39–44.

Tomás, Jesús / Mas, Josep Àngel / Casacuberta, Francisco (2003): "A Quantitative Method for Machine Translation Evaluation"; in: *Proceedings of EACL 2003 workshop on Evaluation Initiatives in Natural Language Processing*. 11th Conference of the European Chapter of the Association for Computational Linguistics, Budapest, <http://www.dcs.shef.ac.uk/~katerina/EACL03-eval/eacl-doc/Tomas.pdf> [data de consulta: 8 de març de 2005].

WICCAC (2004): *Baròmetre de l'ús del català a internet*. Pàgina web de l'organització "Webmàsters Independents en Català, de Cultura i d'Àmbits Cívics" <http://wiccac.org/webscat.html> [data de consulta: 10 de gener de 2005].