Llengua i ús
Revista tècnica de política lingüística
Número 42 (2008)

Instruments

Dos nous parells de llengües al servei de traducció automàtica de la Generalitat de Catalunya: català-occità-català i castellà-occità–castellà

Autors

Mònica Pereña
Secretaria de Política Lingüística

L'article descriu els diversos aspectes del traductor automàtic català-occità-català i castellà-occità-castellà de la Generalitat de Catalunya. S'hi tracten aspectes com els components lingüístics, el sistema de llicències o la qualitat de la traducció en aquesta nova eina que es posa a disposició de la ciutadania.

A Catalunya conviuen tres llengües oficials: el català, el castellà i l'occità, denominat aranès a la Vall d'Aran. A més, el català és la llengua pròpia de Catalunya i l'aranès ho és de la Vall d'Aran.

Les dues línies estratègiques del Pla de política lingüística per a la VIII legislatura, que marquen els objectius de la Secretaria de Política Lingüística per potenciar l'ús de la llengua catalana a Catalunya, són fomentar l'ús del català i fer de la política lingüística una política pública amb caràcter transversal. Es tracta de dues línies estratègiques de gran abast que defineixen les actuacions fonamentals d'aquesta legislatura i el desplegament de les polítiques necessàries per desenvolupar la reforma introduïda, en matèria d'ordenació lingüística, per l'Estatut d'autonomia de Catalunya de 2006.

L'Estatut d'autonomia de Catalunya estableix que l'occità és la llengua pròpia de l'Aran i també llengua oficial a Catalunya. Per això, el Pla incorpora el foment del coneixement i l'ús d'aquesta llengua en cadascuna de les línies estratègiques.

Concretament, l'objectiu 3 de la primera línia estratègica estableix que cal incrementar substancialment el compromís de la Generalitat de Catalunya a l'hora de dur a terme polítiques actives de foment del coneixement i l'ús de la llengua occitana, en col·laboració amb el Consell General d'Aran. Una de les accions previstes al Pla, relacionades amb aquest objectiu, és el desenvolupament d'un traductor automàtic català-occità-català i castellà-occità-castellà, difondre'n l'existència i facilitar-ne l'ús.

Amb aquest article us anunciem, doncs, que el servei de traducció automàtica de la Generalitat de Catalunya disposa, des del mes de juliol, d'aquests dos nous parells de llengües.

Sistema de traducció i implementació

El sistema de traducció automàtica per als parells de llengües català-occità-català i castellà-occità-castellà, que s'ha desenvolupat mitjançant l'adjudicació d'un concurs públic a l'empresa Taller Digital, es basa en la plataforma de traducció automàtica de codi obert Apertium (http://www.apertium.org).

El motor de traducció d'aquesta plataforma és un sistema de traducció automàtica indirecta per transferència sintàctica parcial, hereu de les tecnologies usades en el sistema de traducció espanyol-català-espanyol d'InterNOSTRUM (http://www.internostrum.com), un dels sistemes de traducció automàtica més usats a Internet, i el traductor Universia espanyol-portuguès-espanyol (http://

traductor.universia.net). Apertium és fonamentalment un sistema basat en regles amb components estadístics (com ara el mòdul de desambiguació morfosintàctica).

Aquest sistema de traducció és diferent del que utilitzen els traductors disponibles fins ara al portal de la llengua catalana, que es basen en regles i fan servir la tecnologia de Translendium. Això no obstant, la integració amb el sistema existent fa que des de la interfície d'accés no es percebi cap diferència entre els dos sistemes de traducció. Aquests dos nous parells de llengües apareixen com dues opcions més per escollir i el funcionament és exactament igual per als uns que per als altres.

Components lingüístics

Cal destacar una novetat important respecte dels altres parells de llengües que s'ofereixen des del portal de la Generalitat de Catalunya: per primera vegada és possible treballar, per separat, amb una varietat dialectal determinada, l'aranesa, i amb la varietat lingüística anomenada general.

El desenvolupament es basa en un paquet lingüístic de codi obert ja existent per al parell català-aranès-català que van desenvolupar la Universitat d'Alacant i la Universitat Pompeu Fabra en un projecte finançat per la Generalitat de Catalunya.

Evidentment, aquells recursos lingüístics s'han completat i ampliat, a fi de millorar els resultats de la traducció, i s'han utilitzat com a punt de partida per generar les dades del nou paquet lingüístic català-occità-català. Paral·lelament, s'ha aprofitat el treball de desenvolupament d'aquest paquet i altres dades lliures disponibles per al castellà per elaborar les dades lingüístiques del parell castellà-occità i castellà-aranès.

Tècnicament, els recursos lingüístics estan completament separats del motor, fan servir un format estàndard (XML) i són lliures (llicència pública general de GNU GPL). Això deixa oberta la possibilitat de fer-los servir per a altres tasques relacionades amb les tecnologies lingüístiques per a l'occità, com ara correctors ortogràfics i sintàctics, analitzadors morfològics, lematitzadors, conjugadors verbals, altres parells de llengües de traducció, etc.

Un dels problemes més difícils que es van plantejar a l'hora de construir dades lingüístiques per a la traducció automàtica de l'aranès al català i del català a l'aranès va ser la inexistència de dades lingüístiques llegibles per l'ordinador. Fins i tot ha estat difícil aconseguir textos en quantitats suficients per crear un corpus, eina fonamental per a la tasca de desenvolupament de les dades lingüístiques d'aquest traductor automàtic.

Pel que fa a l'occità general, cal remarcar que encara no hi ha un estàndard definit per a l'occità que sigui acceptat en tot el domini lingüístic, encara que hi ha un cert consens en una forma de base llenguadociana i en versions estandarditzades i convergents de cada dialecte. Al llarg de la història han sorgit diverses propostes normativitzadores de la llengua, però la majoria es refereixen només a varietats dialectals concretes. La voluntat de fixar una varietat comuna per a tot l'occità encara no ha donat lloc a una norma comunament acceptada, cosa que palesen els diversos models de llengua que, per exemple, s'ensenyen a les *calandretes* (escoles occitanes). La solució que sembla més acceptada és una variant general de l'estàndard amb adaptacions regionals que tinguin en compte trets dialectals típics però que conservin una gran convergència i una concepció unitària.

Com que en el domini lingüístic occità no existia, en el moment d'iniciar el desenvolupament d'aquests traductors, cap institució que fes les funcions d'acadèmia de la llengua reconeguda de manera oficial per les institucions de tot el territori, ni tampoc cap entitat admesa com a referència en l'àmbit de la codificació de l'occità pel conjunt de la comunitat lingüística, la Secretaria de Política Lingüística va impulsar la constitució de la Comissió Lingüística per al Traductor Automàtic Occità-Català i Occità-Castellà. Han format part d'aquesta comissió lingüistes de prestigi reconegut de la Vall d'Aran i de diferents regions del domini occità de l'Estat francès i del Piemont italià, els quals han estat nomenats representants de les seves respectives regions amb l'objectiu de fixar les formes i les estructures generals de la llengua occitana que apareixen en el primer traductor automàtic

occità-català-occità i occità-castellà-occità. Per designació de la Secretaria de Política Lingüística, el professor Aitor Carrera, de la Universitat de Lleida, ha estat l'encarregat de dur a terme les tasques de direcció d'aquesta Comissió Lingüística.

La feina de la Comissió ha consistit en la fixació d'un corpus textual inicial que ha servit de base per a la confecció del traductor i en l'establiment de les formes i estructures concretes de la llengua occitana que apareixen al traductor. Això vol dir que ha hagut de determinar les formes i estructures occitanes que produeix el traductor quan es tradueix un text català o castellà, però també el conjunt de formes occitanes que el traductor és capaç d'interpretar com a equivalents d'una única forma catalana o castellana, fet molt important en una llengua que, com l'occità, té una gran variació geogràfica.

Entre les nombroses tasques lingüístiques que ha dut a terme la Comissió, val la pena de destacar les tries gràfiques en casos en què no hi havia una solució unitària o admesa per la comunitat científica; les seleccions fonètiques, morfològiques o lexicals en situacions en què la diversitat dialectal era notable i es feia imperatiu de trobar una solució única (ja que el traductor automàtic ha de donar una única versió occitana de cada element català o castellà, malgrat que n'hi hagi diversos d'igualment vàlids i admissibles des d'un punt de vista normatiu); les nombroses estructures sintàctiques en què el català i el castellà i l'occità difereixen, i d'altres en què la diferència pot arribar a donar-se entre l'occità general i el de la Vall d'Aran. Cal valorar també la feina de verificació del lèxic del traductor que han dut a terme els membres de la Comissió. En bona part de les jornades de treball, a més, la Comissió Lingüística ha tingut la possibilitat de corregir els errors que produïa fins llavors el traductor automàtic a través de textos de mostra, i per tant, d'anar millorant ostensiblement la qualitat del resultat lingüístic final del traductor.

Paral·lelament a les jornades de treball de la Comissió, el director científic ha fet una tasca d'assessorament al personal tècnic de l'empresa encarregada de la realització del traductor automàtic que ha consistit des de la simple resolució dels dubtes episòdics que podien anar presentant-se durant la introducció de dades fins a la confecció de documents lingüístics en què s'especificaven les principals diferències entre l'aranès i l'occità general, així com la recerca i la documentació bibliogràfiques sobre alguns punts conflictius de sintaxi occitana.

Tanmateix, hi ha un bon nombre de decisions de la Comissió que no s'han pres en funció de criteris estrictament lingüístics i específicament relatius a la codificació de la llengua, sinó en funció del grau d'operativitat que tenia triar una solució determinada en comptes d'una altra. Hi ha seleccions de la Comissió, doncs, que es basen en la freqüència d'una determinada solució en el corpus o en les possibilitats d'evitar ambigüitats o problemes d'interpretació en els mecanismes del traductor automàtic. Això implica que bona part dels resultats de la Comissió no es poden considerar una descripció gramatical d'un hipotètic occità estàndard, sinó decisions lingüístiques preses en un context en el qual es prioritza que els parlants es reconeguin en la llengua produïda pel traductor automàtic, i en què es vol generar el màxim nombre de resultats acceptables.

La codificació de la llengua occitana és actualment incompleta i això implica que el procés d'introducció de dades sigui més feixuc o complicat que en llengües normalitzades. Les dades lingüístiques que s'han desenvolupat per a les direccions occità-català i occità-castellà han procurat donar la màxima cobertura a les variants dialectals de l'estàndard occità elegit, per tal de fer el traductor més robust i més útil.

Desenvolupament com a codi obert amb llicències lliures

Una de les característiques diferencials d'aquest sistema de traducció respecte dels que s'utilitzen en els altres parells de llengües del portal Gencat.cat és que el desenvolupament, tant del programari com de les dades lingüístiques i de la documentació associada, s'ha fet en codi obert amb llicències lliures del tipus *copyleft*.

Les llicències de distribució no afecten la propietat. De fet, les dades que Taller Digital ha desenvolupat per als parells occità-català i castellà-occità estan basades en dades obertes ja existents que la Generalitat ha decidit que tinguin llicència lliure, però, com a propietària, conserva els drets a llicenciar-les com consideri convenient i a qui consideri convenient en cada moment.

Qualitat de traducció

Els parells de llengües català-occità i castellà-occità tenen més de 15.000 correspondències de lemes en tots els sentits de traducció. Amb aquest volum d'entrades bilingües, les dades d'avaluació de traducció fetes amb textos d'àmbit general mostren que el sistema té una cobertura (percentatge de paraules d'un text que el sistema tradueix perquè són als diccionaris) aproximada del 94 % en tots els sentits de la traducció. Només 6 de cada 100 paraules queden sense traduir. Quant al percentatge de text que s'ha de corregir a la sortida del sistema de traducció, les xifres són un 10 % per a la variant aranesa i un 23 % per a la variant referencial, més allunyada del català i del castellà.

Com qualsevol sistema de traducció automàtica, aquest també aporta rapidesa i abaratiment de costos pel que fa a la feina de traducció i de revisió de la qualitat lingüística del text traduït. Tot i això, continua sent necessària la revisió feta per un corrector o correctora professional si el text s'ha de publicar o bé es vol obtenir la mateixa qualitat que amb una traducció professional. Ara com ara, doncs, la tecnologia de la traducció automàtica no permet generar traduccions publicables.

Pel que fa a la traducció en línia de pàgines web, tot i que la qualitat lingüística sempre és d'un nivell esborrany, el valor que té és que ens permet entendre el contingut de webs elaborats en altres llengües que desconeixem, perquè els podem llegir en la nostra llengua, i facilita que altres persones que no coneixen la nostra llengua entenguin els continguts que difonem a través dels nostres webs.

Conclusions

Aquests són, doncs, els nous parells de llengües amb què s'incrementa el servei de traducció automàtica de la Generalitat (*http://www.gencat.cat/traductor*), el principal objectiu del qual és cobrir les necessitats de traducció automàtica que tenen els diferents departaments de la Generalitat i oferir un servei gratuït de traducció automàtica per a les empreses i la ciutadania que permeti traduir en línia textos, documents i pàgines web, i ofereixi la possibilitat que qui administri un web pugui incorporar-hi un botó de traducció.

La possibilitat d'incorporar els botons de traducció en pàgines web és molt important per a la internacionalització dels continguts, serveis i productes en català i en occità que hi ha disponibles a Internet i per facilitar la comprensió dels continguts en les altres llengües. L'impacte que ha tingut aquesta funcionalitat fins ara ens demostra que els usuaris li han trobat utilitat i li han atorgat valor. En aquest cas pensem que proporciona als usuaris de l'occità la possibilitat de difondre la seva llengua d'una manera ràpida i fàcil, cosa que contribueix a la promoció de textos i pàgines web en aquesta llengua.

■