Apertium: traducció automàtica de codi obert per a les llengües romàniques

Mikel L. Forcada
Universitat d'Alacant / Prompsit Language Engineering
mlf@ua.es

Resum

Es descriu breument la plataforma de traducció automàtica Apertium (www.apertium.org). Apertium és programari de codi obert, és a dir, programari lliure, que serveix per a construir sistemes de traducció automàtica, que funciona especialment bé en el cas de llengües emparentades com les romàniques, i que està disponible des de 2005. Després d'una breu introducció a la traducció automàtica i a les especials característiques de la traducció automàtica de codi obert, s'expliquen els principis de disseny de la plataforma Apertium, se'n fa una breu descripció tecnològica, es descriu la comunitat de desenvolupadors que s'hi ha format al voltant i es dóna notícia de la recerca realitzada sobre aquesta plataforma. Més avant s'explica el compromís d'Apertium amb les llengües de la Romània, des dels inicis amb els parells espanyol-català i espanyol-gallec fins a la situació actual, amb molts altres parells de llengües romàniques disponibles i en desenvolupament, il·lustrant-lo amb l'aplicació de la plataforma a la llengua occitana.

1. Introducció

Aquest article descriu breument la plataforma de traducció automàtica Apertium (www.apertium.org). Apertium és programari de codi obert, és a dir, programari lliure, que serveix per a construir sistemes de traducció automàtica, que funciona especialment bé en el cas de llengües emparentades com les romàniques, i que està disponible des de 2005. Després d'una breu introducció a la traducció automàtica i a les especials característiques de la traducció automàtica de codi obert (secció 2), s'expliquen els principis de disseny de la plataforma Apertium, se'n fa una breu descripció tecnològica, es descriu la comunitat de desenvolupadors que s'hi ha format al voltant i es dóna notícia de la recerca realitzada sobre aquesta plataforma (secció 3). La secció 4 explica el compromís d'Apertium amb les llengües de la Romània, des dels inicis amb els parells espanyol-català i espanyol-gallec fins a la situació actual, amb molts altres parells de llengües romàniques disponibles i en desenvolupament, il·lustrant-lo amb l'aplicació d'aquesta a la llengua occitana. La secció 5 tanca l'article amb uns comentaris finals.

2. Traducció automàtica de codi obert

2.1 Traducció automàtica

2.1.1 Qué és

La traducció automàtica tracta amb textos escrits, i, en particular, amb **textos informatitzats**, és a dir, amb documents de text emmagatzemats en un mitjà informàtic, documents com els que es poden generar o editar amb processadors de textos. És *automàtica* perquè la realitzen *sistemes informàtics*, és a dir, ordinadors amb el programari adequat instal·lat. Entenem per **traducció automàtica** la transformació, usant un sistema informàtic, d'un text informatitzat escrit en la *llengua origen,* en un altre text informatitzat escrit en la *llengua meta,* que anomenarem *traducció en brut*.

2.1.2 Limitacions de la traducció automàtica

La traducció automàtica (TA) té **limitacions**. En general, les traduccions en brut produïdes pels sistemes de traducció automàtica solen ser molt diferents a les produïdes pels professionals de la traducció i poden no ser adequades per a alguns propòsits comunicatius. Aquesta inadequació està causada per diversos factors, entre els quals podem comptar l'*ambigüitat* dels textos humans (que contenen moltíssims mots amb més d'un sentit o frases amb més d'una estructura sintàctica), les divergències sintàctiques entre la llengua origen i la llengua meta, etc. Aquestos problemes s'aborden amb mètodes que, en general, fan simplificacions bastant radicals del procés de traducció. Aquestes simplificacions, d'una banda, permeten la formulació de regles mecàniques senzilles per a poder construir sistemes de traducció automàtica ràpids i compactes en un temps raonable, però, d'una altra, fa que les solucions estiguen lluny de ser òptimes.

2.1.3 Què podem esperar de la traducció automàtica?

En vista d'aquestes limitacions podem esperar que un bon sistema de TA ens allibere de la part més mecànica (o "mecanitzable") de la tasca de traducció, però, per bo que siga, no podem esperar que comprenga el text, resolga sempre les ambigüitats correctament i produïsca textos en una variant genuïna de la llengua meta.

2.1.4 Aplicacions

Hi ha **dos grans grups d'aplicacions** de la traducció automàtica. El primer grup el formen les aplicacions per a l'**assimilació,** és a dir, l'ús de la traducció automàtica per a comprendre el sentit general de documents (per exemple, textos publicats en Internet) escrits en altra llengua. Un altre exemple de traducció automàtica per a l'assimilació és la traducció de converses en un *xat* o *chat*, de manera que cada persona que hi participa pot usar la seua llengua i llegir les contribucions dels altres participants traduïdes també a la seua llengua. En aquest tipus d'aplicacions la traducció automàtica ha de ser molt ràpida, idealment instantània, i s'usa directament, en brut; hi ha vegades que ni tan sols es llig completament, i normalment no es conserva ni guarda després d'haver-la llegit. Aquesta aplicació de la traducció automàtica no està relacionada amb la traducció professional.

En el segon grup, hi ha les aplicacions per a la **disseminació**. Es diuen així perquè comporten l'ús de la traducció automàtica com a pas intermedi en la producció d'un document en la llengua meta que serà publicat o disseminat; per tant, la traducció en brut es conserva perquè l'ha de revisar i corregir, o com se sol dir, *posteditar,* una persona especialitzada. Simplificant, podem dir que la traducció automàtica seguida de postedició constituirà una alternativa a la traducció professional només si el seu cost conjunt és menor que el de la traducció professional tradicional. De vegades, per a estalviar postedició (especialment quan es tradueix a més d'una llengua meta) es pot fer una miqueta de *preedició* del text original que es traduirà automàticament, evitant problemes coneguts del sistema de traducció automàtica concret que s'estiga usant. Una alternativa a la preedició en el cas que s'han de crear i després traduir molts documents de naturalesa similar és que els autors usen *llenguatges controlats*, és a dir, que escriguen evitant lèxic i construccions que haurien estat posteditades.

2.1.5 Dos grans grups de tecnologies de traducció

També hi ha **dos grans grups de tecnologies de traducció**. Des dels primers intents de fa uns 50 anys fins al decenni dels noranta, l'aproximació dominant a la traducció automàtica ha sigut l'anomenada *traducció automàtica basada en regles*: equips amb informàtics i experts en traducció compilen diccionaris en forma electrònica, programen analitzadors morfològics i sintàctics, definixen regles de transformació gramatical, etc. Des de principis dels noranta assistim a un creixement de l'anomenada *traducció automàtica basada en corpus* (de text): els programes de traducció automàtica "aprenen a traduir" (per exemple usant complexos models estadístics) a partir d'enormes corpus de textos bilingües on centenars de milers de frases en una llengua s'han alineat amb la seua traducció en l'altra llengua.

Aquest article presenta **Apertium**, un sistema de traducció automàtica basada en regles.

2.2 Què és el programari de codi obert?

Revisem breument el concepte de *programari de codi obert* (*open-source software*), o, si usem el seu nom històric encara en ús, *programari lliure (free software*). El programari lliure (podeu trobar una definició a http://www.gnu.org/philosophy/free-sw.html) és programari que (a) pot ser usat lliurement amb qualsevol propòsit, (b) pot ser examinat lliurement per veure com funciona i pot ser modificat lliurement per adaptar-lo a una necessitat nova o a una nova aplicació (per això, el codi font ha de ser disponible, d'ací el nom alternatiu *de codi obert*), (c) pot ser redistribuït lliurement a qualsevol, i (d) pot ser millorat lliurement i alliberat al públic de manera que la comunitat sencera d'usuaris se'n beneficie (el codi font ha de ser disponible per a això també). La *Open Source Initiative* («Iniciativa de codi obert») estableix una definició (http://www.opensource.org/docs/definition.php) que és més o menys equivalent per als propòsits d'aquest article. En aquest article, use la denominació *codi obert* perquè el meu grup ho ha fet tradicionalment, i no perquè, com altres, vulga evitar les connotacions polítiques o ètiques

associades a la denominació *lliure*, les quals compartisc.

2.3 Programari de traducció automàtica: obert o tancat?

2.3.1 Peculiaritats del programari de traducció automàtica

El programari de traducció automàtica (TA) és especial perquè depén fortament de les dades. La traducció automàtica basada en regles (TABR) depèn de dades lingüístiques com ara diccionaris morfològics, diccionaris bilingües, gramàtiques i arxius de regles de transferència estructural; la traducció automàtica basada en corpus (com ara la traducció automàtica estadística, per exemple) depén, directament o indirectament, de la disponibilitat de text paral·lel alineat frase a frase. En els dos casos, s'hi poden distingir tres components: un *motor* (descodificador, recombinador, etc.), *dades* (dades lingüístiques o corpus paral·lels), i, opcionalment, *eines* per mantenir aquestes dades i convertir-los en un format adequat perquè els use el motor.

2.3.2 La traducció automàtica comercial, normalment tancada

HhhLa majoria dels sistemes de traducció automàtica comercials són basats en regles (tot i que han començat a aparéixer sistemes de traducció automàtica amb un fort component basat en corpus[1]). La majoria dels sistemes de TABR usen motors amb tecnologies privatives o de propietat (*proprietary*) que no es revelen completament (de fet, la majoria de les empreses consideren aquestes tecnologies de propietat com el seu principal avantatge competitiu). Les dades lingüístiques no són plenament modificables tampoc; en la majoria de casos, la persona usuària només pot afegir paraules noves o els seus glossaris als diccionaris del sistema, i potser afegir-hi algunes regles senzilles, però no és possible construir un conjunt complet de dades lingüístiques per a un parell de llengües nou i utilitzar-lo amb el motor.

Que un sistema es puga usar en Internet no vol dir que siga obert. Per exemple, hi ha sistemes de TA en la xarxa que poden ser utilitzats lliurement (amb algunes restriccions); alguns són versions de prova de sistemes comercials, mentre que alguns altres sistemes lliurement disponibles, però tancats, no són ni tan sols comercials.[2]

2.3.3 Traducció automàtica de codi obert

D'una banda, perquè un sistema de traducció automàtica basat en regles siga de «codi obert», el codi font del motor i de les eines han de ser distribuïts així com el «codi font» de les dades lingüístiques pels parells de llengües desitjats. És més facil que les persones usuàries de la traducció automàtica de codi obert canvien les dades lingüístiques que que modifiquen el motor de traducció automàtica; a més, perquè les dades lingüístiques millorades puguen ser utilitzades amb el motor, les eines per mantenir-les també haurien de ser accessibles. D'altra banda, si el sistema de traducció automàtica és estadístic, el codi font tant dels programes que aprenen els models estadístics de traducció a partir del text paral·lel així com dels descodificadors que utilitzen aquests models de llengua per generar les traduccions més probables de frases noves haurien de distribuir-se *conjuntament amb els corresponents textos paral·lels alineats frase a frase.*

Recentment, han començat a aparéixer sistemes de traducció automàtica de codi obert. El sistema Apertium que es descriu en aquest article és un d'ells. Es dóna el cas que fins i tot una empresa que es dedicava al negoci de la TA comercial ha començat a distribuir els seus productes com a codi obert.[3]

2.3.4 Avantatges de la TA de codi obert

Els sistemes de TA de codi obert tenen avantatges específics sobre els sistemes comercials de codi tancat. En particular, m'agradaria destacar-ne dos:

1. **Increment de la perícia i dels recursos lingüístics.** Quan s'intenta construir un sistema de traducció automàtica de codi obert per un parell de llengües nou, cal un procés de reflexió sobre les llengües implicades que porta a l'explicitació i a la subsegüent fixació i codificació de coneixement monolingüe i

1 AutomaticTrans (http://www.automatictrans.es), Language Weaver (http://www.languageweaver.com), i, més recentment, Google Translate (http://translate.google.com).

2 Aquest és el cas, per exemple, de dos sistemes de traducció automàtica no comercials però lliurement disponibles entre espanyol i català: interNOSTRUM (http://www.internostrum.com), el qual té milers d'usuaris diaris, i un sistema menys conegut però molt potent anomenat SisHiTra (González et al. 2006).

3 LOGOS ha alliberat recentment el codi font del seu sistema de TA, ara OpenLogos (www.logos-os.dfki.de).

bilingüe. Així doncs, d'una banda, la perícia lingüística resultant, en un escenari de codi obert, queda disponible per a les comunitats lingüístiques interessades. D'altra banda, es generen recursos nous, disponibles de manera oberta per a la comunitat de parlants de les llengües implicades, i que poden ser usats per a nous parells de llengües, o fins i tot per a altres aplicacions de tecnologia lingüística a més de la traducció automàtica.

2. **Augment de la independència.** Un efecte secundari interessant és que la disseminació de coneixement obert i programari de codi obert fa que els usuaris de les comunitats lingüístiques corresponents siguen menys dependents d'un proveïdor comercial particular de programari de codi tancat, no només quant a tecnologies de traducció, sinó potser també quant a d'altres aplicacions de tecnologia lingüística que se'n podrien derivar.

La secció 3.2.3 explica amb més detall les raons per les quals la plataforma de traducció automàtica Apertium es desenvolupa i distribueix com a codi obert.

2.3.5 Reptes de la TA de codi obert

Per a poder gaudir d'aquests avantatges, les comunitats lingüístiques implicades han de fer front a una sèrie de reptes:

1. **Neutralització de les actituds «tecnofòbiques».** Moltes vegades, els experts que podrien ajudar a crear nous sistemes de traducció automàtica desconfien de les tecnologies, potser a causa de la seua visió idealitzada de la llengua i la comunicació humana, i de la seua poca estima pels usos no formals o no literaris.[4] També hi poden intervenir *barreres afectives* que interferisquen amb l'aprenentatge i la subsegüent adopció de les tecnologies de la llengua.
2. **Organització del desenvolupament comunitari.** És comú, i desitjable, que el desenvolupament de programari de codi obert es produïsca de manera comunitària, al voltant del que normalment s'anomena un *projecte*. Per organitzar un projecte, cal, d'una banda, un punt comú d'encontre, un servidor en el qual els desenvolupadors puguen millorar el programari o contribuir dades lingüístiques noves i que permeta els usuaris de la comunitat lingüística implicada descarregar o executar l'última versió del sistema. Però, d'altra banda, calen estructures de coordinació (administradors del projecte, coordinadors de cada parell de llengües, coordinadors del motor de traducció, etc.). Són possibles organitzacions més centralitzades i jerarquitzades o més "horitzontals", depenent del projecte.
3. **Elicitació del coneixement lingüístic**. Aquest és un dels reptes més importants, especialment per a llengües per a les quals la perícia lingüística és escassa o fragmentària. Perquè siga útil per a codificar dades lingüístiques, el coneixement intuïtiu de la llengua per part dels parlants s'ha de fer explícit, és a dir, ha de ser *elicitat*. En la mesura que siga possible, el nivell de coneixements lingüístics necessari per a ser capaç de construir un nou sistema de traducció automàtica nou hauria de ser el mínim possible.
4. **Estandardització i documentació dels formats de dades lingüístiques.** S'ha de definir amb claredat i precisió un format sistemàtic per a cada font de dades lingüístiques utilitzada pel sistema. Una de les millors maneres de definir formats de dades lingüístiques és basar-se en el llenguatge extensible de marcatge XML:[5] els formats resultants són bastant autodescriptius, és possible comprovar automàticament si són vàlids per a l'aplicació abans d'usar-los i es facilita notablement l'intercanvi de les dades amb altres tecnologies i aplicacions lingüístiques.
5. **Modularitat.** Perquè el motor i les dades lingüístiques de traducció automàtica de codi obert siguen útils per a parells de llengües diferents o per a altres aplicacions de tecnologia lingüística, convé que siguen modulars. Per exemple, tenir un analitzador

4 Heus ací una altra explicació possible per algunes d'aquestes actituds tecnofòbiques: molts d'aquests professionals de llengua tendeixen a centrar-se normalment en fenòmens improbables que són propis de la idiosincràsia d'una llengua particular (les «joies» de la llengua), que els sistemes de traducció automàtica tendeixen a tractar incorrectament, en comptes de centrar-se en com aquests sistemes tracten estructures i paraules comunes que constitueixen el 95% dels textos de cada dia (els «maons» de la llengua).

5 http://www.w3c.org/XML/. XML són les sigles d'***ex**tensible **m**arkup **l**anguage*.

morfològic independent i el corresponent diccionari morfològic independent per una certa llengua permet que s'usen en un altre motor de traducció automàtica que té la mateixa *llengua origen* (o *llengua de partida*) i una *llengua meta* (o *llengua d'arribada*) diferent.

3. Apertium

Apertium[6] és una plataforma de traducció automàtica de codi obert, inicialment concebuda per a parells de llengües emparentades (en particular, llengües romàniques), però que ha estat recentment expandida per a poder tractar parells de llengües més divergents (com ara anglès–català). La plataforma proporciona

- un *motor* de traducció independent de les llengües (vegeu la secció 3.3);
- *eines* per a gestionar les dades lingüístiques necessàries per a construir un sistema de traducció automàtica per a un parell de llengües donat o per a adquirir automàticament («aprendre») regles de transferència estructural (Caseli et al. 2006; Sánchez-Martínez et al. 2008) i de desambiguació a partir de textos (Sánchez-Martínez et al. 2008);
- *dades lingüístiques* per a un nombre creixent de parells de llengües (vegeu les seccions 3.4 i 4).

3.1 Rerefons

El disseny inicial està basat en el de sistemes que ja havia desenvolupat pel grup Transducens de la Universitat d'Alacant, com ara interNOSTRUM[7] (espanyol–català), i Traductor Universia[8] (espanyol–portugués). Aquestes tecnologies, inicialment dissenyades per a parells de llengües relacionades, han estat esteses per a tractar parells de llengües que no estiguen tan relacionades.

3.2 La filosofia sobre la qual es fonamenta Apertium

3.2.1 Simplicitat de disseny i modularitat

Per a generar traduccions que siguen raonablement intel·ligibles i fàcils de corregir entre llengües relacionades com l'espanyol (es) i el català (ca) o el portugués (pt), etc., només cal millorar la traducció mot per mot amb: processament lèxic robust (incloent-hi unitats lèxiques multi-mot), desambiguació lèxica categorial (*part-of-speech tagging*) i processament estructural local basat en regles simples i ben formulades per a transformacions estructurals freqüents (reordenació, concordança).

Per a parells de llengües més difícils, no tan relacionats, hauria de ser possible estendre aquest model senzill i generalitzar-ne els conceptes de manera que la complexitat es mantinguera tan baixa com fóra possible, tal com s'ha discutit en 2.3.5.

Apertium té un disseny modular basat en conceptes lingüístics senzills, que es detalla en la secció 3.3.

3.2.2 Separació eficient de motor i dades

D'una banda, hauria de ser possible generar un sistema complet de traducció automàtica a partir de dades lingüístiques (diccionaris monolingües i bilingües, regles gramaticals), especificades de manera *declarativa*. Aquesta informació hauria d'estar en un format interoperable; per exemple, basat en XML (vegeu la secció 2.3.5).

D'altra banda, hauria de ser possible tenir un motor de traducció únic (independent de la llengua) que llegiria dades específiques per a cada parell de llengües («separació d'algorismes i dades»). Les dades lingüístiques del parell de llengües haurien de ser preprocessades de manera que el sistema siga ràpid (més de 10.000 mots per segon) i compacte; per exemple, les transformacions lèxiques es farien amb transductors d'estats finits (TEFs).

Apertium pot ser usat per a construir sistemes de traducció automàtica per a una gran varietat de parells de llengües; per a això, Apertium usa formats senzills basats en XML per a codificar les dades lingüístiques necessàries (fetes a mà o per conversió de dades existents) que es compilen, amb les eines que es proveeixen, en els formats de gran velocitat usats per un motor únic, independent del parell de llengües concret.

Aquests són els quatre tipus bàsics de dades d'Apertium:

- regles (independents de la llengua) per a tractar els diferents formats de text
- especificació del desambiguador lèxic categorial
- diccionaris morfològics i bilingües i diccionaris de regles de transformació ortogràfica

6 http://www.apertium.org
7 http://www.internostrum.com
8 http://traductor.universia.net

- regles de transferència estructural

3.2.3 Desenvolupament i distribució com a codi obert

Aquestes són les raons que van inspirar el desenvolupament d'Apertium en codi obert:

- Donar a tothom accés lliure i il·limitat a les millors tecnologies possibles de traducció automàtica.
- Establir una plataforma modular, documentada i oberta per a la traducció automàtica de transferència superficial i per a altres tasques de processament automàtic de la llengua.
- Afavorir l'intercanvi i la reutilització de les dades lingüístiques existents, tant per a crear nous sistemes de traducció automàtica com per a usar-los en altres tecnologies lingüístiques.
- Facilitar la integració amb altres tecnologies de codi obert.
- Beneficiar-se del desenvolupament col·laboratiu del motor de traducció i de les eines de dades per a parells de llengües existents o nous per part de la indústria, de les universitats o d'organitzacions de suport de llengües menors.
- Promoure el canvi de model de negoci en TA, del model basat en llicències (obsolescent) a un model basat en serveis.
- Garantir radicalment la reproducibilitat de la recerca en TA (vegeu la secció 3.7).
- Perquè no té sentit usar diners públics per a desenvolupar programari no lliure i de codi tancat.

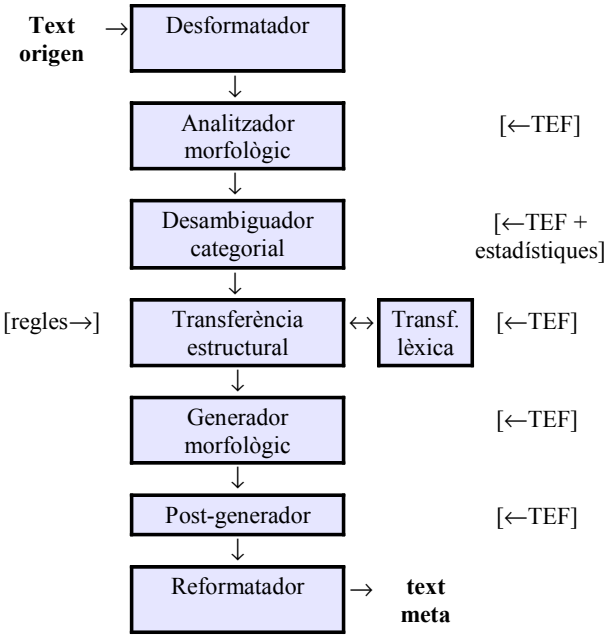
Apertium és, en el moment d'escriure aquest article, un dels pocs sistemes de TA de codi obert (basat en regles[9]) que poden ser utilitzats per a propòsits reals.[10]

3.3 Com funciona Apertium?

Apertium usa un motor de traducció de transferència superficial completament modular que processa el text d'entrada en etapes, com en una cadena de muntatge: desformatatge, anàlisi morfològica, desambiguació categorial, transferència estructural superficial, transferència lèxica, generació morfològica i reformatatge. La comunicació entre els mòduls que s'encarreguen de cada una d'aquestes etapes es fa en forma de text (usant les típiques canonades o *pipelines* d'Unix). Aquest esquema té avantatges clars: simplifica la diagnosi i la depuració d'errors, permet la modificació de dades entre dos mòduls, usant, per exemple, filtres, i facilita la inserció de mòduls alternatius (crucial per a la recerca i el desenvolupament, vegeu la secció 3.7).

Apertium és capaç de traduir textos en els formats de text més comuns (text pla, HTML, RTF, ODF, `.sxw` d'OpenOffice.org, etc.).

La següent figura resumeix el funcionament d'Apertium. Apertium usa transductors d'estats finits (en la figura, TEF) per a les operacions de processament lèxic (anàlisi i generació morfològica, transferència lèxica), models ocults de Markov (basats en estadístiques i tècniques d'estats finits) per a la desambiguació categorial i *chunking* (anàlisi sintàctica superficial) multietapa basat en patrons detectats mitjançant tècniques d'estats finits per a les regles de transferència superficial.



Segueix una breu descripció dels mòduls:

- El **desformatador** separa el text de la informació de format. Actualment hi ha desformatadors disponibles per a text pla, HTML, RTF, ODF, i `.sxw` d'OpenOffice.org. El funcionament està basat en tècniques d'estats finits. La majoria dels desformatadors es generen (usant un full d'estil XSLT) a partir d'un fitxer XML que especifica el seu funcionament per a cada format.

9 El sistema de TA de codi obert basada en corpus més usat és probablement Moses (http://www.statmt.org/moses/).

10 Com s'ha esmentat abans, hi ha també OpenLogos (http://ww.logos-os.dfki.de). Un altre sistema interessant és Matxin (http://matxin.sourceforge.net/), bastant relacionat amb Apertium.

- L'**analitzador morfològic** segmenta el text en llengua origen (LO) en *formes superficials* (FSs), assigna a cada FS una o més *formes lèxiques* (FLs), cada una amb lema, categoria lèxica o part de l'oració, i informació de flexió morfològica. És capaç de processar contraccions i unitats lèxiques multi-mot que poden ser invariables (es: *con cargo a*, *de suerte que*) o variables (es: *echaría de menos* → *echar de menos*). El mòdul lliga transductors d'estats finits *compilats* a partir d'un diccionari morfològic en XML.
- El **desambiguador lèxic categorial** tria una de les FLs corresponents a cada FS ambigua (al voltant del 30% en llengües romàniques) segons el context. Usa models de Markov ocults (preferències estadístiques) i restriccions escrites a mà. S'entrena usant corpus representatius per a la llengua origen (desambiguats manualment o no) o, més recentment, usant models estadístics de la llengua meta (Sánchez-Martínez et al. 2008, vegeu la secció 3.7). El seu comportament està controlat per un arxiu XML.
- El **mòdul de transferència estructural** reconeix *xuncs* o *chunks* (patrons de FLs de la LO) usant tècniques d'estats finits (d'esquerra a dreta i elegint el patró concordant més llarg), i executa les accions associades a cada patró en el fitxer de regles (de la forma *patró—acció*) per a generar el patró de FLs corresponent en la llengua meta. El fitxer de regles de transferència XML es preprocessa perquè siga interpretat més ràpidament. Per a parells de llengües "més difícils", hi ha disponible una transferència estructural en tres etapes:
  - Es detecten, processen i marquen patrons de FLs (*xuncs*)
  - Es detecten i processen patrons de *xuncs* : aquest processament *inter-xunc* permet transformacions sintàctiques d'abast més llarg
  - Els *xuncs* d'eixida son reprocessats si és necessari i les FLs que contenen s'envien a l'eixida.
- El mòdul de **transferència lèxica** llig cada FL de la LO i genera la FL corresponent en llengua meta (LM); usa transductors d'estats finits *compilats* a partir de diccionaris bilingües en XML, i és invocat quan és necessari pel mòdul de transferència estructural.
- El **generador morfològic** genera, flexionant adequadament cada FL en LM, la FS corresponent. Usa transductors d'estats finits *compilats* a partir de diccionaris morfològics en XML
- El **post-generador** realitza transformacions ortogràfiques com ara contraccions (ca: *de +els → dels* ; en: *can + not → cannot*), o inserció d'apòstrofs (ca: *de + amics → d'amics*), etc.; es basa en transductors d'estats finits *compilats* a partir de diccionaris de regles senzilles de post-generació.
- El **reformatador** reintegra la informació de format en el text traduït. Com el desformatador, es basa en tècniques d'estats finits i es genera a partir d'un fitxer d'especificació per a cada format. S'usa també per a modificar els URLs dels enllaços per a la modalitat *navegar i traduir*.

3.4 Dades lingüístiques (parells de llengües)

El projecte Apertium acull el desenvolupament col·laboratiu de dades per a un gran nombre de parells de llengües, amb un èmfasi especial sobre les llengües romàniques. Vegeu l'epígraf 4 per a més detalls.

3.5 Finançament

Des del 2004, Apertium ha estat finançat per nombroses institucions, sense les quals no hauria estat possible:

- Els ministeris d'Indústria, Turisme i Comerç, d'Educació i Ciència i de Ciència i Tecnologia d'Espanya
- La Secretaria de Telecomunicacions i Societat de la Informació (STSI) de la Generalitat de Catalunya
- El Ministeri d'Assumptes Exteriors de Romania
- La Universitat d'Alacant

Empreses: Prompsit Language Engineering, ABC Enciklopedioj, imaxin|software, Eleka Ingeniaritza Linguistikoa, Eolaistriu, etc.

3.6 La comunitat d'Apertium

Al voltant dels desenvolupadors originals (contractats amb el finançament descrit en la secció anterior), s'ha format una comunitat internacional de desenvolupadors (*instigada* fonamentalment per Francis Tyers). En

l'actualitat, hi ha 85 desenvolupadors inscrits en el projecte[11], molts de fora del grup original; el codi i les dades s'actualitzen molt freqüentment (centenars d'actualitzacions cada mes). Un *wiki* mantingut col·lectivament[12] documenta els components d'Apertium, mostra l'estat actual de desenvolupament i dóna consells per als desenvolupadors de dades lingüístiques o de programes. També s'han desenvolupat eines i codi externament: la interfície gràfica d'ús `apertium-tolk`, i l'eina de diagnòstic `apertium-view`; *plugins* per a OpenOffice.org, per al missatger Pidgin (abans Gaim), o per al sistema de gestió de continguts Wordpress; una versió dels diccionaris bilingües per a mòbils amb Java, i, recentment, per a PDA Palm (`TinyLex`); una aplicació para la traducció de subtítols de pel·lícules (`apertium-subtitles`), versions preliminars per al sistema operatiu Windows, etc. Molts dels desenvolupadors es troben en el canal de *xat* IRC `#apertium` (del servidor `irc.freenode.net`), per a discutir en línia assumptes d'Apertium de manera més o menys formal.

Des de fa dos anys els paquets estables estan disponibles com a part de la distribució Debian de GNU/Linux (i per tant, en la popular distribució Ubuntu Linux).

3.7 Apertium com a plataforma d'investigació

La plataforma de traducció automàtica (TA) de codi obert Apertium ha estat utilitzada com a plataforma d'investigació per a la implementació de nous mètodes que permeten el desenvolupament més ràpid i eficient d'alguns dels recursos necessaris per a la construcció de nous parells de llengües. De fet, recentment s'ha defensat una tesi doctoral en el marc del projecte (Sánchez-Martínez 2008).

Entre les recerques en què ha participat el grup Transducens de la Universitat d'Alacant, cal esmentar, a més de la tesi adés referida, els següents treballs:

- Caseli et al. (2006) proposen un mètode per a la inferència de recursos bilingües a partir de bitextos (textos en un idioma, juntament amb la seua traducció a un altre idioma). Els recursos obtinguts comprenen tant diccionaris bilingües com regles de transferència estructural superficial similars a les utilitzades en Apertium per a la TA entre llengües romàniques. El programari usat en aquest treball és també de codi obert[13] i s'ha usat per a iniciar el desenvolupament d'alguns diccionaris bilingües en Apertium.
- Sánchez-Martínez i Forcada (2009) fan ús de tècniques de TA estadística per a la inferència de regles de transferència estructural superficial a partir de bitextos; en aquest cas, no s'infereix cap diccionari bilingüe, sinó que se n'usa un d'existent. El mètode descrit per Sánchez-Martínez and Forcada (2009) ha estat implementat i alliberat com a codi obert dins d'Apertium de tal forma que s'integra fàcilment en el procés de desenvolupament de nous parells de llengües per a Apertium, ja que genera regles en el format XML utilitzat pel mòdul de transferència estructural.
- Sanchez-Martínez et al. (2008) han desenvolupat un nou mètode que permet l'entrenament dels desambiguadors lèxics categorials (*part-of-speech taggers*) basats en models ocults de Markov usats en Apertium de forma completament no supervisada mitjançant l'ús de textos tant en llengua origen com en llengua meta. Aquest mètode, que proporciona resultats clarament millors que els obtinguts pels mètodes d'entrenament no supervisats clàssics, ha estat alliberat com codi obert i s'integra plenament en el procés de desenvolupament de nous parells de llengües per a Apertium.

També hi ha recerques realitzades per investigadors externs:

- Homola i Kuboň (2008)descriuen un experiment realitzat amb Apertium sobre el parell portugués—espanyol, suggereixen una modificació de l'arquitectura del sistema que asseguren que millora la qualitat de traducció i discuteixen les implicacions de la millora de l'arquitectura per al disseny de recursos lingüístics per als sistemes de transferència sintàctica superficial com Apertium.
- Tyers i Donnelly (2009), com s'ha esmentat més amunt, descriuen un sistema obert de TA gal·lés-anglés basat en Apertium, pensat per a l'assimilació d'informació, n'avaluen els

11 En http://sourceforge.net/projects/apertium/

12 http://wiki.apertium.org

13 El programari forma part del projecte ReTraTos, i té l'adreça http://retratos.sourceforge.net/ .

resultats i discuteixen els avantatges del desenvolupament comunitari de sistemes basats en regles per a les llengües marginalitzades.

El fet que aquestes investigacions s'hagen fet sobre una plataforma oberta i disponible, facilita enormement la seua reproducibilitat a d'altres investigadors.

4. Apertium i les llengües romàniques

4.1 El grup de llengües millor representat

Entre els parells *estables*[14] disponibles a hores d'ara en la plataforma Apertium hi ha: **espanyol** ↔ **català**, **espanyol** ↔ **gallec**, **espanyol** ↔ **portugués**, **portugués** ↔ **català**, **portugués** ↔ **gallec**, anglés ↔ **català**, **francés** ↔ **català**, anglés ↔ **espanyol**, anglés ↔ **gallec francés** ↔ **espanyol**, **occità** ↔ **català**, **occità** ↔ **espanyol**, **romanés** → **espanyol**, **espanyol** → esperanto, **català** → esperanto, anglés → esperanto, basc → **espanyol** i gal·lés → anglés.[15] A més, hi ha un nombre creixent de parells de llengües en desenvolupament. Com es pot veure, la majoria dels parells estables inclouen una llengua romànica (en negretes). Això és perquè, de fet, la breu història d'Apertium (cinc anys) està molt lligada a les llengües romàniques, i la naturalesa col·laborativa del projecte ha atret desenvolupadors de procedències molt diverses, com veurem a la secció 4.2.

La taula "Parells de llengües d'Apertium..." dóna notícia de la data de l'última versió estable dels parells de llengües que inclouen una o dues llengües romàniques (a 15 de febrer de 2009). S'ha de tenir en compte que molts dels parells continuen en desenvolupament actiu encara que no se n'haja publicat cap versió estable recentment.

Parells de llengües d'Apertium que inclouen una llengua romànica

Parell de llengües	Última v. estable	Data de l'última versió estable
anglés↔**espanyol**	0.6	19 març 2008
anglés↔**català**	0.8.4	19 març 2008
anglés↔**gallec**	0.5.1	19 novembre 2008
basc→**espanyol**	0.3.0	11 novembre 2008
català→esperanto	0.9.0	20 febrer 2008
espanyol↔català	1.0	28 març 2006
espanyol↔gallec	1.0	7 octubre 2007
espanyol↔portugués	1.0.3	3 octubre 2007
espanyol→esperanto	0.9.0	20 febrer 2008
francés↔català	1.0	5 octubre 2007
francés↔espanyol	0.8.0	14 febrer 2008
occità↔català	1.0.5	12 juliol 2008
occità↔espanyol	1.0.5	12 juliol 2008
portugués↔català	0.8.0	18 juny 2008
portugués↔gallec	0.9.0	10 juny 2008
romanés→espanyol	0.7	8 octubre 2007

4.2 Breu història

Apertium naix, tal com s'esmenta a la secció 3.1, com una reescriptura en codi obert de les tecnologies de traducció existents en el grup Transducens de la Universitat d'Alacant. Aquestes tecnologies s'aplicaven aleshores a la traducció entre llengües romàniques: espanyol ↔ català i espanyol ↔ portugués. Aquesta reescriptura es va realitzar en el marc d'un projecte finançat pel Ministeri d'Indústria, Turisme i Comerç espanyol, en col·laboració amb universitats i empreses de tot Espanya. El resultat va ser un nou motor de traducció, completament redissenyat, i les dades per als parells espanyol ↔ català i espanyol ↔ gallec.

Més avant, amb suport de la Secretaria de Telecomunicacions i Societat de la Informació (STSI) de la Generalitat de Catalunya, es van llançar els parells francés ↔ català i català ↔ occità (inicialment, aranés), conjuntament amb l'anglés ↔ català. El cas de l'occità es descriu amb més detall en la secció següent.

Quasi paral·lelament, amb suport del Ministeri d'Assumptes Exteriors de Romania, i en un projecte dirigit per la Prof. Catalina Iliescu de la Universitat d'Alacant, es va començar a treballar en el parell romanés ↔ espanyol. Els problemes plantejats pel joc de caràcters del romanés van

14 L'ús de la denominació *estable* no fa referencia a la qualitat del traductor corresponent, sinó al fet que Apertium ha publicat paquets informàtics per a aquestes llengües, a punt per a poder-los instal·lar fàcilment.

15 Vegeu Tyers i Donnelly (2009)

motivar l'adaptació d'Apertium a Unicode (joc de caràcters universal, vàlid per a totes les llengües); això ha permés l'inici del desenvolupament de parells de llengües amb sistemes d'escriptura diferents (com el macedoni).

El parell espanyol ↔ portugués és també de la mateixa època. Aquest és, sens dubte, un dels parells de llengües romàniques més gran (darrere, potser, del parell espanyol ↔ francés). El grup Transducens va decidir muntar un paquet de dades (Armentano-Oller et al. 2006) a partir del coneixement que li havia permés desenvolupar el traductor Universia (http://traductor.universia.net/), ara comercial.

El 2006 es crea l'empresa Prompsit Language Engineering, amb programadors i lingüistes d'Apertium. Un dels primers parells que s'hi inicien, per encàrrec de l'empresa Eleka Ingeniaritza Linguistikoa, és l'espanyol ↔ francés, el qual continua en desenvolupament.

El 2007, la Universitat Pompeu Fabra i l'empresa ABC Enciklopedioj desenvolupen els sistemes espanyol → esperanto i català → esperanto. D'altra banda, Armentano-Oller i Forcada (2008) publiquen el primer prototip portugués ↔ català, construït a partir dels parells espanyol ↔ portugués i espanyol ↔ català.

El 2008, l'empresa imaxin|software publica el traductor portugués ↔ gallec, muntat a partir de les dades espanyol ↔ portugués i espanyol ↔ gallec.

També a finals de 2008, usant dades procedents del projecte Matxin,[16] la Universitat d'Alacant llança el primer prototip traductor basc → espanyol.

Actualment hi ha dos parells més de llengües en desenvolupament actiu en el projecte: espanyol—italià, finançat i desenvolupat per la Universitat d'Alacant, i bretó—francés cofinançat i desenvolupat per la Universitat d'Alacant i L'Ofis ar Brezhoneg (Oficina del Bretó).

4.3 Un exemple: Apertium i l'occità

El desenvolupament de TA per a l'occità per part de la Universitat d'Alacant i la Universitat Pompeu Fabra va començar en 2006 amb el parell aranés—català, finançat per la STSI de la Generalitat de Catalunya. Aquest parell connectava una llengua *mitjana* (el català, amb uns 6.000.000 parlants) i una variant estandarditzada *molt menuda* (l'aranés, amb uns 6.000 parlants) d'una llengua més gran, l'occità, amb potser 1.000.000 parlants. El desenvolupament (Armentano-Oller i Forcada 2006) es va iniciar partint de dades existents (espanyol—català), un exemple clar de reutilització de dades obertes.

Més avant, el 2007 les empreses alacantines Prompsit i Taller Digital guanyen un concurs públic i són contractades per la Generalitat de Catalunya per a construir els traductors oficials occità ↔ català i occità ↔ espanyol, tant per a l'aranés com per a l'occità general (*occitan larg*).

Un dels principals problemes d'aquest treball rau en l'estandardització de l'occità general, que avança molt lentament. Això convertia la iniciativa en autènticament pionera. Per a definir quin seria el model de llengua que produirà el sistema, es va crear una comissió d'experts lingüístics de *quasi* tot Occitània (2 experts per *regió*) amb participació d'una experta d'Apertium (Gema Ramírez). El model de llengua elegit (no sense llargues discussions) està basat en el dialecte llenguadocià.

En l'actualitat, amb un sistema bidireccional, completament operatiu, que es pot descarregar o usar en línia, i que té el 95% de cobertura i una taxa d'error del 10% per a la traducció aranés—català i del 25% d'error per a la traducció *occitan larg*—català (clarament millorable), es poden començar a produir els efectes següents:

- La quantitat de text en occità en la web, generat mitjançant traducció automàtica seguida de postedició, pot augmentar la visibilitat de la llengua.
- L'existència de traducció automàtica de qualitat pot promoure la difusió de les variants de l'occità elegides.
- La comunitat occitana general (la majoria a França) pot crear un traductor occità—francés a partir de les dades occità—català o occità—espanyol i francés—català o francès—espanyol ja existents en Apertium.
- Les dades públiques i obertes disponibles per a l'occità poden ser útils per a crear altres aplicacions de tecnologia lingüística per a aquesta llengua.

Els sistemes de traducció occità ↔ català i occità ↔ espanyol resultants, són, des del 5 de

16 http://matxin.sourceforge.net

novembre de 2008, els oficials de la Generalitat de Catalunya.[17]

5. Comentaris finals

El llançament, fa quatre anys, de la plataforma de traducció automàtica de codi obert Apertium (www.apertium.org) ha facilitat el desenvolupament col·laboratiu de sistemes de traducció automàtica oberts (i de tecnologia lingüística oberta, a punt per a ser transferida a d'altres aplicacions) per a moltes llengües, però molt especialment per a les llengües romàniques, per a les que va ser inicialment concebut. Això ha estat possible principalment gràcies al suport d'institucions públiques, però també d'empreses interessades a oferir serveis de traducció automàtica en el model de negoci emergent que possibilita el programari obert.

Crec que Apertium pot contribuir a una comunicació més fluida entre les comunitats de la Romània: d'una banda, ajudant en la producció de traduccions que es poden fer públiques amb poc esforç de correcció, i, d'altra, ajudant els internautes a llegir documents escrits en altres llengües romàniques per mitjà de traduccions aproximades instantànies.

En el cas particular de la llengua occitana, encara queda per avaluar quin serà l'impacte d'Apertium en l'estandardització pendent d'aquesta llengua.

Agraïments: Com ja he dit més amunt, Apertium ha estat finançat, des de 2004, pels governs espanyol, català i romanés, per la Universitat d'Alacant, i per nombroses empreses. Apertium (i aquest article) no serien possibles sense l'ajuda de molts investigadors i desenvolupadors, com Carme Armentano-Oller, Enrique Benimeli, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mireia Ginestí-Rosell, Juan Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Sergio Ortiz-Rojas, Míriam A. Scalco, Francis M. Tyers, i molts altres.

Referències

Armentano-Oller, C., Carrasco, R.C., Corbí-Bellot, A.M., Forcada, M.L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J.A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Scalco, M.A. (2005) "Open-source Portuguese-Spanish machine translation", in *Lecture Notes in Computer Science* **3960** (Computational Processing of the Portuguese Language, Proceedings of the 7th International Workshop on Computational Processing of Written and Spoken Portuguese, PROPOR 2006) 13-17 de maig de 2006, Itatiaia, Rio de Janeiro, Brasil., p. 50-59.

Armentano-Oller, C., Forcada, M.L. (2006) "Open-source machine translation between small languages: Catalan and Aranese Occitan", in Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages) (organitzat conjuntament amb l'LREC 2006 (22-28.05.2006)), p. 51-54.

Armentano-Oller, C., Forcada, M.L. (2008) "Reutilización de datos lingüísticos para la creación de un sistema de traducción automática para un nuevo par de lenguas", *Procesamiento del Lenguaje Natural* **41**, 243-250.

Caseli, H. M., M. G. V. Nunes, M. L. Forcada (2006). "Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation". *Machine Translation* **20**(4)227-245. Publicat el 2008.

González, J., Lagarda, A.L., Navarro, J.R., Eliodoro, L., Giménez A., Casacuberta, F., de Val, J.M., Fabregat, F. (2006) "SisHiTra: A Spanish-to-Catalan hybrid machine translation system". In *LREC-2006: Fifth International Conference on Language Resources and Evaluation. 5th SALTMIL Workshop on Minority Languages: "Strategies for developing machine translation for minority languages",* Gènova, Itàlia, 23 maig 2006; pp.69-73

Homola, P., Kuboň, V. (2008). "Improving Machine Translation Between Closely Related Romance Languages". In *Proceedings of the European Association of Machine Translation*, p. 72—77.

Sánchez-Martínez F. (2008). "Using unsupervised corpus-based methods to build rule-based machine translation systems". Tesi Doctoral, Departament de Llenguatges i Sistemes Infomàtics, Universitat d'Alacant.

Sánchez-Martínez, F., Pérez-Ortiz, J.A., Forcada, M.L. (2008). "Using target-language information to train part-of-speech taggers for machine translation". *Machine Translation*, **22**(1-2) 29-66.

Sánchez-Martínez, F., Forcada, M.L. (2009). "Inferring shallow-transfer machine translation rules from small parallel corpora". *Journal of Artificial Intelligence Research* (accepted).

Tyers, F. M. and Donnelly, K. (2009) "apertium-cy - a collaboratively-developed free RBMT system for Welsh to English". *The Prague Bulletin of Mathematical Linguistics* **91**: 57-66.

17 http://traductor.gencat.cat/