MIREIA GINESTÍ ROSELL/
MIKEL L. FORCADA

LA TRADUCCIÓ AUTOMÀTICA EN LA PRÀCTICA: APLICACIONS, DIFICULTATS I ESTRATÈGIES DE DESENVOLUPAMENT

1. Introducció

La *traducció automàtica* (TA) és el procés de traducció, mitjançant un sistema informàtic (compost per ordinadors i programes), de textos informatitzats escrits en la llengua origen a textos informatitzats escrits en la llengua meta. Un *text informatitzat* és un fitxer d'ordinador que conté un text en algun format conegut.

La investigació sobre la traducció automàtica ha experimentat un gran auge en els últims anys. És un tema que interessa cada cop més en el nostre món multilingüe. 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 no solen ser adequades per a la majoria de propòsits comunicatius; com veurem més endavant, això és perquè la TA no ha aconseguit superar encara moltes de les dificultats del treball de traducció. Això no vol dir que no tingui utilitat en molts àmbits; només vol dir que cal ser conscient que té unes aplicacions concretes, que en força casos no coincideixen amb les aplicacions de la traducció humana, i que per tant no té sentit fer comparacions de

Caplletra 46 (Primavera 2009), pp. 43-60

qualitat entre un i altre mètode de traducció; el que cal és identificar els contextos en què es pot treure profit de la TA i què en podem esperar.

En aquest article volem presentar quines són les aplicacions actuals de la traducció automàtica, els principals problemes que planteja i les diferents estratègies utilitzades per a abordar-los. Presentarem un exemple concret de sistema de TA, la plataforma Apertium, i descriurem algunes de les tensions que es poden produir en el desenvolupament d'un sistema de traducció automàtica real i com es poden afrontar.

Finalment, presentarem els principals sistemes de TA existents per al català.

2. Aplicacions de la traducció automàtica

Ja ha quedat clar que l'objectiu de la TA no pot ser substituir la traducció professional. Els propòsits comunicatius de la TA es poden agrupar en dues grans categories:

1. Comprensió (també anomenada *assimilació*): traduir textos per tenir-ne una comprensió general, quan es desconeix la llengua origen i es vol tenir una idea aproximada del contingut del text. És el cas, per exemple, de la «navegació traduïda» de pàgines web, és a dir, la visita de pàgines d'Internet a través d'un traductor automàtic que les tradueix instantàniament a la llengua escollida. En aquesta aplicació, els errors en la traducció no són tan importants si s'aconsegueix transmetre el sentit general del text. En aquest context no té massa sentit usar els serveis d'un traductor professional; cal tenir en compte que una gran part dels textos traduïts no seran llegits ni guardats per a ser llegits més endavant. Un altre exemple de traducció automàtica per a l'assimilació és la traducció de converses en un *xat*, de manera que cada persona que hi participa pot usar la seva llengua i llegir les contribucions dels altres participants traduïdes també a la seva llengua.
2. Publicació (també anomenada *disseminació*): traduir textos com a pas intermedi en la producció d'un document en la llengua meta que serà publicat; el text traduït automàticament l'ha de revisar i corregir, o com se sol dir, *posteditar,* una persona especialitzada. Un sistema de TA útil en aquest sentit ens alliberarà de la part més mecànica (o «mecanitzable») de la tasca de traducció, però, per bo que sigui, no podem esperar que resolgui sempre les ambigüitats correctament i produeixi textos en una variant genuïna de la llengua meta. 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.

Perquè un sistema de TA sigui útil per al segon d'aquests propòsits (publicació), la proximitat lingüística entre la llengua origen i la llengua meta és determinant, almenys amb la tecnologia disponible avui dia. Per exemple, en el cas de la traducció entre l'espanyol i el català, l'ús d'algun dels sistemes de TA existents (que presentem a la secció 6) pot estalviar molta feina als traductors professionals; per a les combinacions anglès-català o anglès-castellà, en canvi, no hi ha cap sistema que pugui complir la mateixa funció. Per al primer grup de propòsits (comprensió), potser els sistemes de traducció entre llengües llunyanes són més útils, ja que la barrera lingüística que se supera, encara que sigui amb errors, és molt més alta.

3. Principals dificultats en la traducció automàtica

La ciència actual no és capaç d'expressar de manera formal, mitjançant regles, tots els mecanismes subjacents als llenguatges naturals ni els processos mentals involucrats en el seu ús. Aquesta impossibilitat per a obtenir una caracterització precisa, comuna a molts altres camps de la intel·ligència artificial, és un dels principals arguments que poden donar-se per a explicar la dificultat d'escriure un programa d'ordinador que tradueixi textos.

Tanmateix, els problemes de la TA no deriven només de la impossibilitat de caracteritzar formalment el funcionament del llenguatge humà. Hi ha característiques dels llenguatges naturals que fan que la traducció automàtica continuï essent un problema difícil de resoldre.

Un dels principals problemes de la TA és l'ambigüitat del llenguatge. Si tenim en compte l'anomenat *principi de composicionalitat semàntica*, és a dir, el principi segons el qual la interpretació d'una oració es construeix o compon a partir de les interpretacions dels mots seguint l'ordre dictat per l'estructura sintàctica, queda clar que una oració pot ser ambigua:

— perquè un o més dels seus mots tenen més d'una interpretació (*ambigüitat lèxica*),
— perquè l'oració té més d'un arbre d'anàlisi sintàctica (*ambigüitat estructural* o *sintàctica)*,
— o, en alguns casos, per ambdues coses alhora.

Així, si diem «*el gat està sota el cotxe*», pot ser que parlem d'un felí o d'una ferramenta (ambigüitat lèxica del mot *gat*); si diem «*hem rebut notícies de França*», no ens queda clar si és que les notícies són sobre França o que vénen d'aquest país (ambigüitat estructural: no sabem si el sintagma preposicional «*de França*» és complement

del nom *notícies* o del verb en el sintagma verbal «[*hem*] *rebut notícies*)». I si llegim el titular «*La direcció de l'empresa s'oposa a implantar mesures internes per evitar la crisi»,* no sabem amb seguretat si l'empresa s'oposa a les mesures perquè vol evitar la crisi, o si s'oposa a implantar mesures destinades a evitar la crisi (una altra ambigüitat estructural deguda a un sintagma preposicional); haurem de llegir la resta de la notícia per estar-ne segurs. Finalment, si un valencià ens diu «*les gallines han destrossat el sembrat però no les mates*», no sabem si està delimitant la magnitud del desastre o si ens està demanant que siguem clements amb les culpables (ambigüitat mixta: lèxica de *les*, pronom o article, i *mates*, nom o verb, que dóna lloc a una ambigüitat estructural, perquè la segona oració té dues sintaxis vàlides possibles i ben diferents).

En molts casos, l'elecció de la interpretació correcta d'una oració per part del sistema de traducció automàtica és necessària per a produir una traducció adequada. Si agafem els exemples anteriors, veurem que per traduir-los del català a l'espanyol només tindríem problemes en l'última frase («*no las mates*» vs. «*no las matas*»). En la resta de casos, es pot traduir sense resoldre l'ambigüitat. Per traduir del català a l'anglès, en canvi, són font d'errors l'ambigüitat de la primera, la segona i la quarta frase («*the cat*» vs. «*the jack*»; «*from France*» vs. «*about France*»; «*do not kill them*» vs. «*not the shrubs*»).

La resolució automàtica de l'ambigüitat és una tasca que està molt lluny de ser resolta. Mentre les persones poden usar el context i els seus coneixements, les seves expectatives i les seves creences sobre el món per a descartar amb seguretat moltes interpretacions (idealment, totes menys una), els sistemes de traducció automàtica han de prendre aquestes decisions usant únicament regles programables i factibles amb un consum de memòria i temps acceptable que processin la informació (sovint incompleta) que puguin extreure del text que envolta l'element ambigu.

L'ambigüitat no és l'únic obstacle que han de superar els programes de TA. Arnold (2003) classifica els problemes de la TA en quatre grups:

> 1. *La forma no determina completament el contingut*. Un text pot tenir diferents interpretacions. Fa referència, doncs, a l'ambigüitat del llenguatge, que acabem d'explicar.
>
> 2. *El contingut no determina completament la forma*. Existeixen multitud de formes d'expressar en una llengua un contingut donat i perquè un ordinador no hagi d'enfrontar-se a la complexitat que això implica, és necessari imposar estratègies que redueixin les possibilitats, encara que sigui a costa de perdre expressivitat. Per exemple, només una d'aquestes expressions és la manera adequada d'expressar el fet que volem saber en quin punt del dia ens trobem: «*Quina hora és?*», «*Quines hores són*» (com en portuguès «*que horas são?*»), «*Com és de tard?*» (com en alemany «*wie spät ist es?*»), etc.

3. *Diferents llengües usen estructures diferents per a expressar el mateix*. Considerem la frase «*m'agraden les pomes*»; la seva traducció a l'anglès és «*I like apples*», on pot veure's com el subjecte de la frase en espanyol (*les pomes*) s'ha transformat en un complement directe en el cas de l'anglès (*apples*). Encara que aquest és un exemple relativament senzill, en general, les estructures usades per llengües distintes poden ser tan divergents que facin que una simple traducció directa sigui inintel·ligible.

4. Finalment, com s'ha comentat a l'inici d'aquest apartat, és difícil caracteritzar amb la precisió necessària els principis involucrats en l'ús del llenguatge. De fet, construir un sistema de traducció automàtica comporta la gestió d'una gran quantitat de coneixement, que s'ha d'arreplegar, descriure, i representar en una forma útil (com veurem en la secció 5, aquest procés no està exempt de problemes). Aquest no és el cas en els sistemes de *traducció automàtica basada en corpus*, que utilitzen informació de naturalesa estadística obtinguda automàticament a partir del processament de grans corpus de text, i que s'expliquen breument en el següent apartat.

Tots aquests problemes es redueixen ostensiblement quan les llengües implicades en la traducció estan emparentades. En aquest cas les afinitats a nivell morfològic, sintàctic i semàntic simplifiquen el disseny d'aquests sistemes i permeten arribar fàcilment a traduccions en les quals menys d'un 10% del text s'hauria de posteditar. D'altra banda, existeixen tipus de textos que poden ser més fàcilment traduïts (cartes comercials, manuals d'instruccions, textos econòmics) i uns altres la traducció automàtica dels quals és a dia d'avui (i probablement serà per molt temps) totalment inviable (poesia, per exemple).

4 Tècniques de traducció automàtica

Al llarg de la història d'aquesta disciplina, la implicació de lingüistes en la investigació i el desenvolupament de la TA ha anat oscil·lant entre dos pols, entre sistemes amb poc o nul coneixement lingüístic (traducció rudimentària paraula per paraula, traducció estadística) i sistemes amb complexes codificacions lingüístiques (sistemes amb anàlisi sintàctica completa, sistemes basats en *interlingua*).

Actualment hi ha dos grans grups de tecnologies de traducció. Des de la dècada dels 80, l'aproximació dominant ha estat l'anomenada *traducció automàtica basada en regles*: equips amb informàtics i experts en traducció compilen diccionaris en forma electrònica, programen analitzadors morfològics i sintàctics, defineixen 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 milions de frases en una llengua s'han alineat amb les seves traduccions en l'altra llengua. Existeixen també aproximacions mixtes, per exemple, sistemes estadístics que incorporen algun tipus de coneixement lingüístic (com ara diccionaris bilingües o d'anàlisi morfològica) i sistemes de regles que incorporen algun tipus de coneixement no lingüístic (com ara memòries de traducció on es guarden les traduccions d'oracions completes provinents dels corpus de text o models estadístics de desambiguació).

Els sistemes basats en regles requereixen més temps de construcció (cal codificar explícitament les dades lingüístiques que utilitzarà el sistema) mentre que els sistemes basats en corpus són més ràpids de construir sempre que es disposi d'un gran volum de corpus de textos alineats; per tant, són difícils d'aplicar a la traducció d'una llengua minoritària amb poca disponibilitat de corpus bilingües digitalitzats.

D'altra banda, podríem dir que, en general, els sistemes basats en regles són els que millor qualitat de traducció proporcionen avui dia, especialment entre llengües properes, i els més utilitzats en els sistemes comercials, tot i que en els últims anys els sistemes estadístics han millorat considerablement i comencen a oferir bons resultats.[1]

Entre els sistemes basats en regles, els més habituals són els de traducció automàtica per transferència, que funcionen en tres fases ben diferenciades:

1. La fase d'*anàlisi* produeix, a partir de la frase en llengua origen, una representació intermèdia abstracta , en la qual s'estableixen classificacions i agrupaments lingüístics que permeten l'aplicació de regles generals de traducció. Per exemple, si l'anàlisi ens indica que el segment català «*el coixí còmode*» es compon d'un determinant, un substantiu i un adjectiu, sabrem que quan el traduïm a l'espanyol, el canvi de gènere del substantiu s'ha d'aplicar al determinant i a l'adjectiu perquè han de concordar-hi: «*la almohada cómoda*».

2. La fase de *transferència* converteix la representació intermèdia anterior en una nova representació intermèdia en la llengua meta, tot restablint-hi la concordança com en l'exemple citat, fent reordenaments o canvis lèxics com ara els que calen per a transformar «*el hombre cuya hija murió en Barcelona*» en «*l'home la filla del qual va morir a Barcelona*», etc.

3. La fase de *generació* produeix una frase concreta en la llengua meta a partir d'aquesta representació intermèdia abstracta.

1. El traductor estadístic de Google es pot provar en http://translate.google.com/translate_t?hl=es.

Les representacions intermèdies poden ser més o menys complexes. Alguns sistemes duen a terme una anàlisi sintàctica completa de cada oració, mentre que els anomenats sistemes *de transformació* (en anglès *transformer systems*: Arnold *et al.* (1994, secció 4.2)) substitueixen l'anàlisi sintàctica completa i les transformacions d'arbres per la detecció i transformació de *chunks* o segments sintàctics (patrons plans i no estructurats de categories com ara *article-nom-adjectiu*). Aquesta aproximació és molt adequada quan la llengua origen i la llengua meta són sintàcticament similars, ja que en aquest cas no cal fer grans transformacions sintàctiques ni cal, per tant, fer l'anàlisi sintàctica completa: de fet, només cal analitzar on hi ha divergència.[2]

El sistema de TA Apertium, que presentem a continuació, és un sistema de transferència que utilitza aquesta estratègia de transformació.

4.1 El sistema Apertium

El sistema Apertium[3] és una plataforma de codi obert (programari lliure)[4] que inclou els programes i les eines per a construir un sistema de traducció automàtica per transferència. És, per tant, un producte informàtic «acabat» al qual cal afegir les dades lingüístiques (diccionaris, regles de traducció, dades per a la desambiguació lèxica) necessàries per a la TA entre un parell de llengües determinat.

Apertium ha estat desenvolupat en el marc de diferents projectes de subvenció pública, en els quals han participat, fins al moment, diferents universitats (Universitat d'Alacant, Universidade de Vigo, Universitat Politècnica de Catalunya, Euskal Herriko Unibertsitatea i Universitat Pompeu Fabra) i empreses (Eleka Ingeniaritza Linguistikoa, Imaxin Software, Elhuyar Fundazioa i Prompsit Language Engineering), i es basa en l'experiència i els coneixements adquirits pel grup Transducens de la Universitat d'Alacant en el desenvolupament del sistema català–espanyol interNOSTRUM[5] (Canals-Marote *et al.* 2001) i del traductor portuguès–espanyol Traductor Universia[6] (Garrido-Alenda *et al.* 2004).

2 . Quan diem que "no cal", volem dir que no cal fer-ho per a aconseguir traduccions en brut on es pot aprofitar més del 90% del text tal com està.

3. http://www.apertium.org.

4. Se sol dir que un programa és lliure quan (a) qualsevol el pot usar lliurement per a qualsevol propòsit, (b) qualsevol el pot copiar i distribuir lliurement, (c) qui vulgui el pot modificar per a millorar-lo o per a adaptar-lo a un nou propòsit i (d) qui vulgui en pot distribuir les versions modificades. Com que per a les últimes dos condicions s'ha de tenir accés al codi (tal com l'ha escrit el programador), també se sol parlar de programes de codi obert. El programari lliure pot crear al seu voltant comunitats de desenrotlladors voluntaris que el millorin i l'amplïin; aquest comença a ser el cas d'Apertium.

5. http://www.internostrum.com.

6. http://traductor.universia.net/.

El fet que Apertium sigui de codi obert el diferencia de la majoria de programes de TA, que han estat tradicionalment sistemes tancats. En els últims anys, el canvi d'enfocament provocat pel programari lliure està arribant també a aquesta tecnologia. Un programa de TA de codi obert es pot adaptar lliurement a nous parells d'idiomes o dominis aprofitant dades, es pot integrar en altres aplicacions; també se'n poden usar els recursos en altres projectes d'investigació i desenvolupament.

Sobre la plataforma Apertium hi ha disponibles actualment, entre d'altres i en diferents graus de desenvolupament, traductors per als parells català–espanyol, espanyol–gallec, espanyol–portuguès, aranès–català, francès–català, anglès–català, anglès-espanyol, francès-espanyol, romanès-espanyol i català-esperanto.

Com hem dit abans, Apertium és un sistema de transferència que no realitza una anàlisi sintàctica completa de les oracions d'un text. El mòdul de transferència fa només una detecció de patrons de paraules que requereixen algun tipus de processament.

El sistema consta de diferents mòduls (vegeu la figura 1):

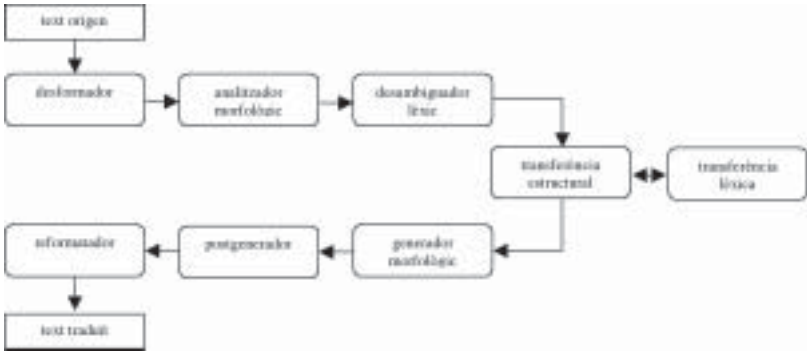

Figura 1: Els vuit mòduls del sistema Apertium

1. Un *desformatador* que separa el text que es vol traduir de la informació de format.

2. Un *analitzador morfològic* que separa el text en unitats lèxiques i proporciona, per a cada una d'elles, una anàlisi (o més d'una si la paraula és ambigua) que conté el lema, la categoria lèxica i la informació morfològica de flexió. Per exemple, l'anàlisi de la frase anglesa «the water is clean» seria així:

^the/the<det><def><sp>$ ^water/water<n><sg>$ ^is/be<vbser><pri><p3><sg>$ ^clean/clean<adj><sint>/clean<vblex><inf>/clean<vblex><pres>$

Veiem que l'última paraula (*clean*) pot ser adjectiu o una de dues formes d'un verb.

3. Un *desambiguador lèxic* que, quan es troba amb una unitat lèxica ambigua (que té més d'una anàlisi), en tria una d'acord amb un model estadístic. En la frase anterior, si el desambiguador funciona correctament triarà l'adjectiu com a categoria de *clean.* La sortida, doncs, seria:

^the<det><def><sp>$ ^water<n><sg>$ ^be<vbser><pri><p3><sg>$ ^clean<adj><sint>$

4. Un *mòdul de transferència lèxica* que llegeix cada forma analitzada i entrega la corresponent forma lèxica en llengua meta, consultat un diccionari bilingüe. En la traducció al català de l'exemple, les formes entregades serien:

the<det> → el<det>
water<n> → aigua<n><f>
be<vbser> → ser<vbser>
clean<adj> → net<adj>

5. Un *mòdul de transferència estructural* que detecta patrons de formes lèxiques que requereixen un processament addicional (reordenament dels mots, operacions de concordança, canvis de preposició, etc.) i els aplica les regles de transformació corresponents. Aquest mòdul pot ser únic (Apertium «nivell 1») o bé, si es vol realitzar transformacions estructurals més complexes, es pot dividir en tres submòduls: un primer que detecta els patrons de formes lèxiques, un segon que detecta patrons de patrons (és a dir, patrons de constituents, com ara el patró sintagma nominal–sintagma verbal), i un tercer que realitza operacions finals (Apertium «nivell 2»).

La traducció anglès–català funciona amb Apertium «nivell 2» i, per tant, amb tres submòduls de transferència. La sortida de cada un d'ells, per a la frase d'exemple, seria:

Submòdul 1:

^Det_nom<SN><DET><f><sg>{^el<det><def><3><4>$ ^aigua<n><3><4>$}$ ^be<Vcop><vbser><pri><p3><sg>{^ser<vbser><3><4><5>$}$ ^adj<SA><GD><ND>{^net<adj><2><3>$}$

Submòdul 2:

^Det_nom<SN><DET><f><sg>{^el<det><def><3><4>$ ^aigua<n><3><4>$}$ ^be<Vcop><vbser><pri><p3><sg>{^ser<vbser><3><4><5>$}$ ^adj<SA><f><sg>{^net<adj><2><3>$}$

Submòdul 3:

^El<det><def><f><sg>$ ^aigua<n><f><sg>$ ^ser<vbser><pri><p3><sg>$ ^net<adj><f><sg>$

Explicarem resumidament les operacions que fan aquests submòduls, perquè les etiquetes i els símbols que utilitza el programa poden resultar una mica confuses. Veiem que el primer submòdul agrupa dins de claus {...} les paraules en tres constituents o *chunks*: *SN* (sintagma nominal format per determinant i nom), *Vcop* (sintagma verbal format pel verb copulatiu) i *SA* (sintagma adjectival format per l'adjectiu). Fa també operacions de concordança dins dels patrons, en aquest cas, fa que el gènere i el nombre del determinant *el* concordin amb el nom *aigua* (femení singular).

En el segon submòdul, una regla detecta el patró *SN – Vcop – SA* i assigna el gènere i el nombre del *SN* al *SA* (femení singular).

El tercer submòdul desfà els chunks i propaga les etiquetes dels chunks als seus components. Així, l'adjectiu *net* acaba el procés amb les etiquetes *f* i *sg* (femení singular).

6. Un *generador morfològic* que proporciona, per a cada forma lèxica en llengua meta, una forma final flexionada. Per al nostre exemple, la sortida seria:

~La aigua és neta

7. Un *postgenerador* que fa operacions ortogràfiques com ara contraccions i apostrofacions. En el nostre exemple, s'apostrofarà l'article davant del nom. La sortida d'aquest mòdul seria:

L'aigua és neta

8. Un *reformatador*, que restaura la informació de format extreta pel desformatador.

5. TENSIONS EN EL DISSENY D'UN SISTEMA DE TA

Després de veure els diferents sistemes de TA i de presentar el sistema Apertium, volem explicar els dilemes a què s'enfronten els lingüistes a l'hora de desenvolupar un sistema de TA real. Sovint cal fer sacrificis, compromisos i aproximacions per passar del sistema *ideal* a un sistema *possible*.[7]

7. Aquesta discussió està basada en la de Forcada (2002).

No hi ha només un sistema possible. Cal tenir clar amb quin propòsit es desenvolupa un sistema de TA concret i de quins condicionants es parteix (com ara temps i recursos disponibles). No és el mateix dissenyar un sistema dirigit a un públic general per a la traducció de pàgines web, que un sistema per donar suport a la traducció d'informes tècnics en un ajuntament, o que un traductor de les converses de xat entre els estudiants d'una universitat a distància. Els fenòmens lingüístics que haurem d'abordar prioritàriament poden ser molt diferents, i aspectes tècnics com la velocitat de traducció també poden tenir molta més rellevància en un context que en un altre (per al xat, per exemple, la rapidesa és fonamental). Així com un traductor professional ha de tenir en compte els destinataris i el context de la traducció, el mateix han de fer els dissenyadors d'un traductor automàtic.

A continuació exposem els problemes pràctics a què s'han d'enfrontar els lingüistes que es dediquen a desenvolupar un sistema de TA, basant-nos en la nostra experiència en el desenvolupament de traductors basats en el sistema Apertium que hem descrit en l'apartat anterior.

5.1 Mecanitzabilitat: necessitat de coneixement explícit

Hem vist en l'apartat 3 que una de les dificultats de la TA (l'explicada en el quart punt) és la dificultat de caracteritzar amb precisió el funcionament del llenguatge. No sols és una tasca complicada, sinó que a aquesta dificultat s'hi afegeix la necessitat de fer una tria dels fenòmens que podem tractar en el nostre sistema real.

A l'hora de dissenyar les regles de traducció, doncs, ens caldrà escollir i codificar aquelles regles que abordin amb més eficàcia, generalitat i simplicitat els canvis estructurals que volem tractar. Només seran útils les regles que:

· puguin ser programades en un temps i amb un esforç raonables en el sistema de traducció automàtica (*requisits de desenvolupament*: necessitem un traductor útil en un temps raonable) i

· s'executin usant una quantitat de memòria i en un temps raonables *(requisits d'ús*: programa ràpid i compacte).

De vegades, la caracterització precisa i lingüísticament correcta d'un fenomen no és la millor estratègia per solucionar un problema concret de traducció. Ens cal distingir entre, per una banda, la descripció formal del llenguatge i, per altra banda, el tractament simplificat de problemes concrets de traducció amb mitjans senzills i eficaços que, si no en tots, solucionin el problema en la majoria de contextos.

Podem posar com a exemple d'aquesta simplificació el cas de la traducció del verb *haber* de l'espanyol al català en el nostre sistema. La seva traducció quan funciona

com a auxiliar és *haver*, però quan té el sentit de «existir» cal traduir-lo per *haver-hi*. En comptes d'intentar dissenyar regles per a tots els contextos en què pot traduir-se com a *haver-hi* (seguit de nom, seguit d'adverbi, seguit de preposició, seguit de determinant, seguit de coma, etc.), a Apertium hem decidit traduir-ho *sempre* com a *haver-hi* i crear una regla que fa que es tradueixi com a *haver* només quan va seguit de participi (tenint en compte que els diccionaris ja identifiquen expressions multiparaula com *haber de* i *haber que,* que tenen una traducció fixa sense *hi*). D'aquesta manera, amb una sola regla solucionem correctament quasi totes les situacions amb resultats molt satisfactoris.

5.2 Dues tensions de disseny

Les tensions durant el disseny d'un sistema de traducció automàtica es poden agrupar en dos tipus (no del tot desconnectats):

— Tensió *adequat vs. factible* o «*què hauríem de fer?*» *versus* «*què podem fer amb el disseny actual del sistema?*»

— Tensió *interessant vs. suficientment freqüent* o «*resoldre fenòmens lingüísticament interessants*» *versus* «*resoldre fenòmens que es presenten freqüentment en els textos d'entrada*»

5.2.1 Adequat vs. factible

Per a il·lustrar aquesta tensió, es donen alguns exemples de fenòmens que convindria tractar en la traducció espanyol–català i català–anglès, però que per ara Apertium no tracta perquè el disseny del sistema no ho permet:

— No podem eliminar la preposició *a* davant d'objecte directe en la traducció espanyol–català, ja que sense fer una anàlisi sintàctica i semàntica és molt difícil distingir els objectes directes dels indirectes o dels circumstancials.

— No podem resoldre la polisèmia perquè el sistema només considera un equivalent per a cada mot en llengua origen: *«vive en esta dirección» → «viu en aquesta adreça»,* però *«viene en esta dirección» → «ve en aquesta *adreça»*.

Els problemes de polisèmia, en alguns casos es resolen a través de col·locacions de longitud fixa incloses en els diccionaris (per exemple, traduïm *dirección* com a *adreça* però *dirección general* com a *direcció general*).

— Relacionat amb el punt anterior, hem d'escollir la traducció més adequada d'una paraula quan en la llengua meta té dues o més traduccions diferents. Per exemple, el nom *peix* pot traduir-se com a *pez* o *pescado*. La paraula *pez* semblaria la més general, però tenint en compte el propòsit del traductor

(dirigit sobretot a la traducció de textos periodístics i d'àmbit general) i observant la seva aparició en corpus, hem decidit traduir-la com a *pescado* perquè és més adient en la majoria de contextos observats.

— És molt difícil elegir quin és l'equivalent adequat d'alguns pronoms àtons en castellà, ja que en molts casos és necessari saber quin n'és l'antecedent: *se lo → li ho*, *li'l*, *els ho*, *els el*, *s'ho*, *se'l*?

— En la traducció a l'anglès dels possessius no podem determinar quasi mai el gènere del posseïdor: *«és el seu pare» →* (...) *«is his/her/its father»*.—El mateix passa amb els subjectes elidits que cal afegir: *és aquí → «he/she/it is here»*. En aquests casos, optem per traduir al masculí quan no tenim més informació.

— No podem determinar el gènere en català dels noms de professions en anglès: *teacher → mestre/mestra*. Aquí també optem per la forma masculina quan no tenim més informació, per ser la forma menys marcada.

Algunes d'aquestes decisions, com ara els dos últims punts, són doloroses: hem de renunciar a tractar correctament un fenomen i prendre una decisió general que en bastants casos no serà adequada.

5.2.2 Interessant vs. suficientment freqüent

En altres ocasions, un fenomen és *lingüísticament interessant*, sembla *tractable*, però no és *suficientment freqüent*: resoldre'l no tindria un impacte apreciable sobre la qualitat global de les traduccions i, en canvi, necessitaria un temps de desenvolupament que seria millor invertir en fenòmens més freqüents. La freqüència observada en *corpus representatius de text* ajuda els desenvolupadors del sistema de traducció automàtica a establir-hi *prioritats.*

Heus ací alguns exemples:

— *Construcció dels diccionaris*: encara que pugui ser interessant incloure l'equivalència entre el mot espanyol *vencetósigo* (nom de la planta herbàcia *Vincetoxicum officinale*) i el català *maseres* (mot que apareix només una vegada en els 52 milions de mots de la part de llengua no literària del *Corpus de Textos Informatitzats de la Llengua Catalana*, vegeu Rafel i Fontanals (1996)) sembla més interessant incloure mots més freqüents, fins i tot quan no són als diccionaris, com *blog → bloc*.

— *Reordenaments sintàctics llargs*: els lingüistes d'Apertium han escrit regles per a tractar les estructures espanyoles més freqüents amb *cuyo* (*a/os/as)*, com ara *cuyo* + substantiu, *cuyo* + substantiu + adjectiu, o *cuyo* + adjectiu +

substantiu, però no n'hi ha per a *cuyo* + substantiu + preposició + substantiu + adjectiu («cuyas señas de identidad básicas») perquè apareix menys d'una vegada per milió de mots de text.

— Seqüències difícils de desambiguar: hi ha seqüències de mots homògrafs especialment problemàtiques. El segment espanyol *así como* es pot traduir al català de quatre maneres, segons la interpretació de l'homògraf *así* (verb o adverbi) i *como* (conjunció o verb): *així com*, *vaig agafar com, així menjo* o *vaig agafar menjo*. Totes quatre són possibles en algun context; vegeu si no les frases següents: «*demandó medidas de apoyo al sector mediante subvenciones directas así como otras medidas* (...)» (*així com)*; « l*o así como me dijiste*» (*vaig agafar com*); «*me gusta comer con las manos porque así como más a gusto*» (*així menjo*); «*cada vez que pienso en el día en que la así como chocolate sin parar*» (*vaig agafar menjo*). No obstant això, en un corpus de textos periodístics de vora un milió de mots, *tots* els *así como* observats (523) s'han de traduir per *així com*. Per a evitar errors de resolució de l'ambigüitat i en vista de l'aclaparadora majoria d'una de les possibilitats, s'opta per no intentar resoldre-la i es codifica com una expressió fixa en els diccionaris, sacrificant les altres interpretacions. Un altre cas: en espanyol *sesenta y cinco* és quasi sempre el numeral català *seixanta-cinc*, però no sempre: «*murieron sesenta y cinco resultaron heridos*». Decidir en quin cas ens trobem és difícil, i propens a errors; tant, que és millor sacrificar la segona interpretació per improbable.

6. SISTEMES DE TRADUCCIÓ AUTOMÀTICA PER AL CATALÀ

Finalment, volem acabar aquest article amb una breu presentació dels sistemes de TA existents per al català. L'oferta és amplíssima, i, en alguns casos, no només connecta el català amb l'espanyol, sinó també amb altres llengües com l'anglès, el francès, l'alemany, l'occità o l'esperanto:

— SALT, el traductor de la Generalitat Valenciana, dissenyat per Josep Lacreu, va per la versió 4, i deu haver fet ja deu anys. Traduïx de l'espanyol al valencià i del valencià a l'espanyol, es pot descarregar gratuïtament de molts llocs, i funciona sobre els sistemes operatius Windows i Linux. És un programari excel·lent per la cobertura lèxica i està concebut perquè qui l'usa aprengui valencià mentre ho fa.

— Hi ha un programa tecnològicament molt similar a SALT, però de pagament, anomenat Ara (www.ara-autotrad.com), que genera textos en la variant de Catalunya.

— El traductor interNOSTRUM.com, finançat per Caja de Ahorros del Mediterráneo i la Universitat d'Alacant, està disponible en línia (www.internostrum.com). En la direcció espanyol-català permet triar la variant valenciana o la variant catalana, i quan tradueix a l'espanyol «entén» diverses variants del català. És probablement el programa de traducció més usat a Internet.

— Apertium.org és una plataforma de traducció automàtica de codi obert que ja hem descrit en aquest article (en la secció 4.1). Es pot usar en línia i es pot instal·lar en el sistema operatiu Linux. Sobre aquesta plataforma hi ha disponibles, entre d'altres, traductors entre el català i l'espanyol, el gallec, el francès, l'anglès, l'occità, el romanès i l'esperanto.

— Translendium (www.translendium.com) és una empresa que comercialitza, entre molts altres, sistemes de traducció automàtica entre el català i l'espanyol, l'anglès, l'alemany, el francès i l'occità aranès. Es pot provar gratuïtament en línia.

— AutomaticTrans (www.automatictrans.es) tradueix entre l'espanyol i el català; està relacionat amb el traductor que usa cada dia *El Periódico de Catalunya* per a produir la seva edició bilingüe.

— SisHiTra («sistema híbrid de traducció») és un traductor menys conegut, creat per la Universitat Politècnica de València amb finançament públic, que es pot usar directament a través de la web (prhltdemos.iti.es/%7Esishitra/). Els resultats són molt bons.

Tots aquests sistemes són, potser amb l'excepció de l'últim, sistemes basats quasi exclusivament en regles. Hi ha també un sistema experimental basat en corpus, entrenat sobre textos bilingües d'*El Periódico de Catalunya*, i desenvolupat també per la Universitat Politècnica de València, que està accessible per a fer proves (d'una frase) en http://dcomgp05.gnd.upv.es/WebTrans.debug/root, però que encara no és tan operatiu com els basats en regles.

7. Conclusions

Hem descrit breument una de les tecnologies lingüístiques més importants, la traducció automàtica, discutint les aplicacions possibles, les dificultats a què s'enfronta i les tècniques en què es basa. Com a exemple, hem descrit amb una miqueta més de detall el sistema Apertium coordinat des de la Universitat d'Alacant i hem exposat algunes de les dificultats i tensions que poden aparèixer en dissenyar un sistema de traducció real.

Hem vist que la traducció automàtica és una eina imperfecta que cal adequar a un propòsit determinat. Al llarg de l'article hem donat exemples concrets de problemes que poden sorgir en el desenvolupament de productes de TA dirigits als usuaris; els lingüistes que s'hi dediquen són conscients que treballen en un terreny pràctic i orientat als resultats, per la qual cosa sovint s'han d'allunyar de la lingüística teòrica per inventar solucions *ad hoc*.

Evidentment, els principis i les estratègies que regeixen aquesta feina són molt diferents dels que regeixen l'ampli camp actual d'investigació en traducció automàtica, en què els resultats immediats no són el primer objectiu. La investigació explora noves estratègies i obre noves possibilitats que en el futur poden arribar a aplicar-se a nous sistemes de traducció. Ambdós aspectes, la investigació i el desenvolupament, contribueixen a fer avançar aquesta disciplina.

Finalment, quant a la situació de la llengua catalana en aquest camp i vista l'àmplia oferta de sistemes i de combinacions lingüístiques disponibles, podem concloure que el català està en una posició excel·lent quant a l'existència de tecnologies de traducció automàtica, i que, probablement, aquest potencial no s'està realitzant encara en la seva totalitat.

Mireia Ginestí Rosell
Prompsit Language Engineering
Mikel L. Forcada
Universitat d'Alacant

REFERÈNCIES BIBLIOGRÀFIQUES

ARNOLD, D. (2003) «Why machine translation is difficult for computers» dins SOMERS, H. L. (coord.) *Computers and Translation: a translator's guide,* pp. 119-142, Amsterdam, John Benjamins.

ARNOLD, D., BALKAN, L., MEIJER, S., HUMPHREYS, R., i SADLER, L. (1994) *Machine translation: An introductory guide*, Oxford, NCC Blackwell (esgotat; disponible en http://clwww.essex.ac.uk/~doug/MTbook/).

ARMENTANO-OLLER. C., A. M., CORBÍ-BELLOT, M. L., FORCADA, M., GINESTÍ-ROSELL, M. A., MONTAVA, S., ORTIZ-ROJAS, J. A. ,PÉREZ-ORTIZ, G., RAMÍREZ-SÁNCHEZ i SÁNCHEZ-MARTÍNEZ, F. (2007) «Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática», dins *Proceedings of FLOSS (Free/Libre/Open Source Systems) International Conference*, pp. 5-20.

CANALS-MAROTE, R., ESTEVE-GUILLEN, A., GARRIDO-ALENDA, A., GUARDIOLA-SAVALL, M., ITURRASPE-BELLVER, A., MONTSERRAT-BUENDIA, S., ORTIZ-ROJAS, S., PASTOR-PINA, H., PEREZ-ANTÓN, P. M. i FORCADA, M. L. (2001) «The Spanish-Catalan machine translation system interNOSTRUM», dins *Proceedings of MT Summit VIII: Machine Translation in the Information Age,* Santiago de Compostel·la, pp. 18–22.

FORCADA ZUBIZARRETA, M. L. (2002) «Tensions entre l'adequació lingüística i la factibilitat informàtica durant el desenvolupament d'un sistema de traducció automàtica», dins *Interlingüística* 13, pp. 19-28.

GARRIDO-ALENDA, A., GILABERT-ZARCO, P., PÉREZ-ORTIZ, J. A., PERTUSA-IBÁÑEZ, A., RAMÍREZ-SÁNCHEZ, G., SÁNCHEZ-MARTÍNEZ, F., SCALCO, M. A. i FORCADA, M. L. «Shallow parsing for Portuguese - Spanish machine translation» dins BRANCO, A., MENDES, A., i RIBEIRO, R., (eds.) (2004), *Language technology for Portuguese: shallow processing tools and resources*, Lisboa, pp. 135–144.

RAFEL I FONTANALS, J. (dir.) (1996) *Diccionari de freqüències*, Barcelona, Institut d'Estudis Catalans.