



## El Periódico, una experiencia en traducció automàtica

Ricard Fité

Coordinador lingüístic de El Periódico de Catalunya

L'edició en llengua catalana d'*El Periódico de Catalunya* és present als quioscos des del 28 d'octubre del 1997 gràcies, en bona part, a l'ajuda d'un traductor automàtic. I diem en bona part perquè si bé és veritat que sense aquesta eina tecnològica el producte no es podria fer, no és menys cert que tampoc seria possible sense la participació diària d'un equip humà integrat per una quarantena llarga de lingüistes que deixen els textos resultants en condicions de ser publicats.

### AVANTATGES

Treballar amb un traductor automàtic té avantatges, però també inconvenients. Comencem pels avantatges. En un diari, és essencial perquè permet guanyar temps. El que nosaltres fem servir tradueix una pàgina de diari en tres segons, que és tota l'estona que s'ha d'esperar un lingüista per començar a treballar els textos en català un cop enllestits en castellà. Amb aquesta eina, una pàgina pot quedar perfectament traduïda en un temps que oscil·la entre els deu minuts i els tres quarts d'hora, segons la quantitat de text que contingui i la temàtica de què tracti. En conjunt, la mitjana de temps invertit ve a ser una mitja hora per pàgina i persona. Aquesta relació pàgina / temps és el que permet tancar l'edició puntualment, requisit imprescindible per poder arribar a l'hora a la rotativa i després als punts de venda. Perquè cal saber que l'edició diària en català s'ha de tancar com a màxim mitja hora més tard que l'edició en castellà. Així, si en castellà s'acaba a les 00.00 hores, que és el més habitual, en català no es pot tancar mai més tard de les 00.30 hores. Aquests horaris es compleixen pràcticament cada dia. No fer-ho genera uns costos econòmics (com ara haver de doblar les rutes de distribució dels exemplars) que s'han d'evitar tant com sigui possible. Però aquest factor temps no és decisiu només en l'edició diària: també ho és en les edicions setmanals dels suplementos, que tampoc es podrien tancar el dia corresponent sense el concurs del traductor.

A més de guanyar temps (perquè no s'ha de repicar el text i la traducció es converteix, per dir-ho d'alguna manera, en una correcció, però sense oblidar que som responsables d'una traducció), el traductor automàtic garanteix la correcció ortogràfica al cent per cent, cosa que no es pot dir de tothom que redacta o tradueix un text. És obvi que un traductor automàtic no s'oblidarà mai la *ela* geminada de *protocol·lari* quan hagi de traduir *protocolario*, ni afegirà una *ena* a *trastorn* ni se la descuidarà a *transvasament*. En ortografia hi guanyem, clarament. També ens simplifica i facilita la feina a mesura que anem coneixent com és i com treballa, homogeneïtzant els textos, garantint-ne la coherència formal, i permet establir i difondre un mateix estil. En serien exemples l'ús

sistemàtic del perfet perifràstic en lloc del simple (*dijeron* = van dir, i no pas *digueren*) i dels plurals en -os (*bosques* = boscos, i no pas *boscs*). Amb un traductor automàtic és més fàcil unificar l'estil periodístic en català de cent redactors d'El Periódico que escriuen en castellà, que no pas l'estil periodístic de cent redactors del diari *Avui* que escriuen en català.

En aquest mateix terreny, el traductor automàtic també facilita molt la fixació de grafies conflictives, com ara les dels topònims i antropònims procedents d'alfabet àrab (marroquins, palestins, iraquians, etcètera) o ciríl·lic (russos, ucraïnesos, búlgars, etc.). L'inconvenient és que en castellà aquests noms propis també fluctuen, i per tant s'ha d'estar al cas de quina és la grafia exacta que finalment adopten per poder-la introduir a la base de dades del traductor.

## INCONVENIENTS

El principal dels inconvenients del traductor automàtic és que no pensa la frase en català i que per tant tendeix a la literalitat (tot i que, diguem-ho de passada, quan pertoca tampoc és pas dolenta, la literalitat, ans al contrari!), de manera que és absolutament necessari proveir-se d'un antídoto en aquest sentit que actuï de manera permanent en l'inconscient del lingüista que controla el text (per exemple, preguntar-se constantment: "¿escrivint en català ho hauria redactat així?"). També hem de ser molt conscients que som davant d'una màquina de traduir cega que ignora la sintaxi (objectes directes i indirectes, pronoms febles, relatius, gerundis, règims verbals, ordre de la frase, etc.), i que en el camp del lèxic no pot resoldre la polisèmia. Però, per damunt de tot, hem de tenir molt present que la màquina ignora el coneixement del món, aquell coneixement que tenim les persones i que sense adonar-nos-en donem en general per sabut. Si no l'avises (i d'entrada no la pots avisar de tot aquest saber), et pots trobar de cop i volta que el diari *La Vanguardia* s'ha convertit en *L'Avantguarda*, o que el president Zapatero es diu Sabater, i la seva vicepresidenta De l'Horta.

## PRIMERA VERSIÓ DEL TRADUCTOR (1997)

Actualment treballem amb una versió evolucionada del traductor automàtic inicial, de fabricació pròpia. L'empresa va considerar en el seu moment que un equip d'informàtics de la casa sense formació lingüística específica, ajudat d'un equip de lingüistes sense coneixements d'informàtica (els mateixos lingüistes que després controlarien l'edició en català) podria crear l'eina més adequada per a les seves necessitats. I la jugada, d'entrada força agosarada, va sortir bé.

Des del punt de vista lingüístic, la versió primitiva era molt i molt simple. El motor de traducció posava en funcionament dues bases de dades, una de paraules i una de seqüències.

La base de dades de paraules contenia les paraules en castellà amb la catalana equivalent (*perro* = *gos*, *amarillo* = *groc*, *sorprender* = *sorprendre*, *ellos* = *ells*), i era tan simple com restrictiva, perquè només admetia equivalents únics, seguint l'estructura "paraula castellana = paraula catalana", i no distingia categories gramaticals ni relacionava aspectes morfològics. Així, per exemple, el fet que la base de dades contingués l'entrada *perro* = *gos* no permetia al traductor traduir el plural *perros*. I exactament el mateix passava amb els verbs: tenir introduït a la base l'infinitiu *sorprender* = *sorprendre* no comportava la traducció de cap altra forma de la conjugació verbal. És a dir, la base de dades de paraules només permetia la traducció dels mots de què disposava, considerats un per un i

aïlladament.

Com que el traductor no disposava de regles morfològiques ni sintàctiques que aportessin correcció gramatical al resultat de la traducció simple, la base de dades de seqüències intentava pal·liar aquesta mancança. Aquesta base funcionava exactament com la de paraules, només que amb seqüències. Però en la pràctica aquestes seqüències presentaven les mateixes restriccions morfològiques i sintàctiques que les paraules, i per tant s'havien d'introduir una per una, quan en vèiem la necessitat, i un cop introduïdes tenien el mateix rígid comportament que les paraules.

Posem-ne un exemple. Considerant només la base de dades de paraules, el sintagma castellà "su dimisión" era traduït per "el seu dimissió". ¿Per què? Perquè en aquesta base el possessiu "su" només pot ser traduït per "el seu". Per estrany que li pugui semblar a un expert en lingüística informàtica, a causa de la diguem-ne poc intel·ligent configuració d'aquest traductor s'havien de prendre decisions lingüístiques amb criteris estrictament numèrics, o fins i tot arbitraris, que com a mínim permetessin establir un cert ordre dins el caos, tant pel que fa al funcionament del traductor com a la tipologia dels errors que el lingüista hauria de corregir a posteriori. En el cas dels possessius de gènere invariable en castellà però variable en català (*mi, tu, su, mis, tus, sus*) vam prendre la decisió d'optar per la forma masculina. D'aquesta manera, asseguràvem el cent per cent d'èxit en la combinació "possessiu + substantiu masculí" però alhora el cent per cent de fracàs en la combinació "possessiu + substantiu femení". (L'única alternativa a això era l'opció contrària, és a dir decantar-se per la forma femenina, "la seva", i invertir l'èxit i el fracàs de les traduccions obtingudes amb la decisió anterior.)

Així doncs, per pal·liar, com dèiem, aquesta limitació de la base de dades de paraules, recorriem a la base de seqüències, en la qual, seguint amb el mateix exemple, introduïem el sintagma "su dimisión" i li dèiem que l'havia de traduir per "la seva dimissió". Per obtenir un èxit moderat en la majoria de seqüències d'aquestes característiques, no vam tenir més remei que anar introduint en aquesta base, una per una, totes les paraules femenines més habituals en els textos d'un diari precedides de la sèrie de possessius invariables en castellà: "mi dimisión = la meva dimissió", "tu dimisión = la teva dimissió", "su dimisión = la seva dimissió", amb els plurals corresponents. Deixant a part l'esforç humà permanent que representa anar dotant de precisió un traductor automàtic per aquest sistema, no cal dir que, seguint amb el mateix exemple, la mera intercalació d'un adjectiu entre el possessiu i el substantiu en el text original comportava un altre fracàs en la traducció: "su inesperada dimisión" era traduït per "el seu inesperada dimissió", de manera que si no volíem veure mai més aquesta errònia solució, havíem d'introduir a la base la seqüència concreta "su inesperada = la seva inesperada". I així successivament. A propòsit d'aquesta manera de treballar, sempre vam tenir la sensació d'estar fent una feina inacabable, infinita, semblant a la d'intentar buidar el mar Mediterrani amb un cubell d'aigua, amb l'agreujant que per l'estret de Gibraltar d'aigua n'hi va entrant contínuament.

La configuració d'aquesta primera versió del traductor automàtic es completava amb una única regla que resolvia força bé l'aspecte de l'apostrofació, inexistent en castellà. La resta consistia a anar omplint la base de dades de paraules i la de seqüències com millor ens convingués, d'acord amb l'experiència de la traducció diària.

## EVOLUCIÓ DE LES BASES DE DADES

Quan va sortir al carrer el primer número d'*El Periódico* en català, feia cinc mesos que treballàvem amb el traductor. Durant aquest període de proves, en què traduïem pàgines de diari, vam anar dotant-lo d'informació, a través d'aquestes dues bases, tot partint d'una base de paraules inicial de vint mil entrades i una base de seqüències amb zero entrades. Aquestes vint mil paraules van ser bolcades per un procés automatitzat des de la versió digitalitzada del *Diccionari català-castellà*, 2a edició, d'Enciclopèdia Catalana (heu llegit bé: *català-castellà*, no pas *castellà-català*, com hauria pertocat). La primera dificultat va consistir, doncs, a anar detectant els errors que es derivaven d'aquest bolcatge de dades, realitzat de manera automàtica i en sentit invers. En aquella època no existia encara la versió digitalitzada del *Diccionari castellà-català*, i des de la pragmàtica lògica informàtica (que no cal dir que no té res a veure amb la lògica dels lingüistes) es va considerar que les equivalències en la direcció de català a castellà no serien gaire diferents de les equivalències en la direcció inversa (més o menys, una cosa així com ara que "si *gos* és igual a *perro*, i *perro* és igual a *gos*" ja som al cap del carrer). Naturalment, la lògica informàtica no va tenir en compte que en multitud d'ocasions la simetria *gos* = *perro* / *perro* = *gos* no es compleix, ni de bon tros, i que sovint dóna lloc, en el millor dels casos, a traduccions còmiques (com ara *trabajador* = *feinater*, obtingut d'invertir *feinater* = *trabajador*; *cordel* = *ficel·la*, obtingut d'invertir *ficel·la* = *cordel*; *enchufar* = *fillolar*, obtingut d'invertir *fillolar* = *enchufar*), i en el pitjor dels casos a traduccions esperpèntiques o aberrants (com ara *montón* = *feinada*, obtingut d'invertir *feinada* = *montón m de trabajo*, perquè, per acabar-ho d'adobar, el procés automàtic de bolcatge només considerava la primera paraula de l'equivalència: ¡ves què li expliques a una màquina!).

Al cap d'un any, cap a l'octubre del 1998, les bases de dades ja tenien una mica de cara i ulls, i a més tancàvem el primer cicle anual, cosa que en certa manera garantia haver traduït tota mena de textos des del punt de vista temàtic almenys en alguna ocasió. En un any, en un diari ja has hagut d'afrontar tota la varietat temàtica que requereix un lèxic determinat: futbol, toros, bolets, neu, castells, Nadal, platges, moda, motor, etcètera, de manera que el lèxic de les bases de dades ja tenia força més a veure amb les nostres necessitats que al començament de l'experiència. Durant aquest primer any també vam carregar la base de dades de paraules amb les conjugacions verbals dels verbs considerats segurs des del punt de vista de la traducció, com ara *sorprender* = *sorprendre*. I el manteniment diari de les bases de dades, donant d'alta i de baixa les paraules i seqüències que ens convenien i ens deixaven de convenir, ens va portar al maig del 2005, data en què aquesta versió primitiva va ser rellevada per l'actual, a disposar d'una base de dades de més de 500.000 entrades, de les quals gairebé quatre cinquenes parts eren paraules i la resta seqüències.

Per al manteniment de les bases de dades vam seguir, i encara seguim, un criteri també molt simple. Com que cada entrada admet una traducció i només una, només donem traducció a les entrades que són segures. I en canvi, sempre que una paraula en castellà tingui més d'una traducció en català, no la donem mai d'alta a la base, o l'eliminem si hi era. D'aquesta manera, en el text traduït la paraula desconeguda pel traductor apareix entre angles (i amb un color diferent), sense traduir. En serien exemples totes les paraules polisèmiques en castellà (*fallo* = *error*, *decisión*; *alcanzar* = *aconseguir*, *arribar*).

## EL PROBLEMA DE LES MAJÚSCULES

Una altra de les dificultats que presentava aquesta primera versió del traductor era la no distinció inicial entre majúscules i minúscules. Això volia dir que d'entrada el traductor interpretava qualsevol paraula començada en majúscula com si fos una minúscula, i per tant posava en perill una gran quantitat de cognoms, noms d'empreses i tot allò que a cap traductor humà no se li acudiria de traduir. Per pal·liar aquest risc, es va crear una mena de regla per la qual el traductor ignorés els dos mots subsegüents a un nom propi de persona que apareguessin en majúscules (i aleshores va caldre introduir tot el santoral a la base de dades) i que al llarg del mateix text continués ignorant-los (és a dir, no traduint-los) també sense la presència del nom propi. Aquesta regla, doncs, permetia que l'economista Rodrigo Rato pogués continuar dient-se Rato tant si apareixia amb el nom davant com sense. Paral·lelament a aquesta regla diguem-ne de protecció, i com a part de la feina de manteniment de les bases de dades, teníem la possibilitat d'indicar quines paraules en majúscula no volíem que fossin traduïdes. Això sí, una per una.

## EINA EFICAÇ

Des del punt de vista de l'usuari, s'ha de reconèixer l'alta eficàcia de l'eina malgrat la descripció que n'hem fet. No deixa de ser paradoxal el contrast entre, d'una banda, les escasses expectatives d'èxit que genera un traductor automàtic concebut des del punt de vista lingüístic de manera tan simple, pedestre i rudimentària, i, de l'altra, l'altíssim rendiment que ens va acabar donant per traduir el diari. Deixant a part aspectes com la velocitat de traducció, imprescindible per al tipus de feina encomanada, el lingüista agraeix que la màquina no l'enganyi. Perquè com menys intel·ligent és el traductor automàtic, menys possibilitats té d'embolicar la troca. El lingüista sap, per tant, en quines situacions el traductor l'encerta sempre (traduccions segures del tipus *agua* = *aigua*), en quines pot fallar (aspectes morfològics i sintàctics diferents del castellà) i en quines no actua (polisèmies). És a dir, com menys intel·ligent és el traductor, més previsible és per a l'usuari que l'ha de controlar. Menys sorprenent, i per tant més fiable.

No cal dir que aquesta valoració d'un traductor automàtic feta per un lingüista (prioritzant la qualitat) és completament oposada a la que faria el cap de màrqueting d'una empresa fabricant de traductors, per al qual el més important seria recalcar que el seu producte tradueix el 99 per cent de les paraules que es troba, sense aclarir si ben traduïdes o mal traduïdes (prioritzant la quantitat). Un lingüista es fia més d'un traductor que li doni una mica més de feina a l'hora de revisar el text però que no l'enganyi, que no pas d'un altre que presumptament "sàpiga" donar equivalents per a tot però deixant al marge si realment l'està encertant o no.

## SEGONA VERSIÓ DEL TRADUCTOR (2005)

L'any 2004 *El Periódico de Catalunya* va decidir fer evolucionar aquest primitiu traductor cap a una segona versió, més intel·ligent. Els motius de la innovació eren de tipus tecnològic, no pas pràctics ni de millora de la producció, ni a requeriment dels lingüistes, que ja ens havíem acostumat a les peculiaritats de la màquina, l'havíem adaptat al màxim a les nostres necessitats, i més aviat contemplàvem amb una certa incertesa la perspectiva d'haver d'acostumar-nos a un altre traductor que vés a saber què ens faria, quan el que teníem ens anava,

com a lingüistes, la mar de bé. Però la innovació comportava també millores lingüístiques evidents, almenys en el pla teòric, de manera que per aquesta banda les expectatives eren bones.

Després d'uns quants mesos en període de proves, la nova versió del traductor es va incorporar a la producció del diari en català el mes de maig del 2005. Aquesta versió, més intel·ligent, va comptar amb la col·laboració inicial d'Enciclopèdia Catalana, que va aportar material lexicogràfic i també recursos humans des del punt de vista lingüístic i informàtic.

La base de dades del traductor nou és el resultat de sumar la informació dels diccionaris subministrats per Enciclopèdia Catalana i de les bases de dades del traductor vell. Si la traducció procedent del material lexicogràfic d'Enciclopèdia diferia de la solució que proporcionava el traductor vell, s'ha donat preferència a aquesta última. Les novetats que presenta el traductor nou són:

- 1) Les paraules estan etiquetades amb la categoria gramatical corresponent, i això ha permès formular regles que actuen sobre la traducció.
- 2) Les regles solucionen principalment aspectes de concordança (*su = el seu / la seva*, segons el gènere de la paraula següent) i algunes qüestions sintàctiques (distinció de *per / per a* només davant de substantiu, i sempre *per* davant d'infinitiu; *haver* davant d'un participi / *haver-hi* davant d'un no-verb). Com es pot deduir fàcilment, moltes d'aquestes regles serveixen per resoldre de manera general i sistemàtica problemes que en el traductor vell s'havien de solucionar introduint seqüències d'una en una, o que no es podien resoldre de cap manera.
- 3) Com a contrapartida, l'actuació d'aquestes regles obre la possibilitat a interaccions imprevistes (atesa la capacitat infinita de la llengua de generar frases) que poden produir traduccions agramaticals i diguem-ne estranyes, sense explicació aparent a primera vista. Però com sempre la producció diària determinarà quines ens convenen i quines no.
- 4) La base de dades distingeix, ara sí, entre majúscules i minúscules. Aquesta novetat millora la fiabilitat del traductor, però també fa augmentar la feina de manteniment. Sigui com sigui, podem dir al traductor que si es troba una determinada paraula en majúscula la pot traduir d'una manera, i si se la troba en minúscula d'una altra. (*Grosso modo*, perquè la casuística és molt i molt complexa: també podem distingir entre la paraula començada en majúscula i tota ella escrita en majúscules, i a més als diaris i als suplementos sovint moltes frases són escrites amb totes les lletres en majúscules.)

Aquestes regles de què parlem no són pròpiament regles sintàctiques tal com les entenem els lingüistes. En podríem dir *fórmules informàtico-morfosintàctiques*, sovint deduïdes a partir d'un cas concret que sembla generalitzable, expressades en un *llenguatge màquina* que no tindrà mai en compte contextos ni incisos imprevistos i imprevisibles. La possibilitat de dotar el traductor de regles permet fer volar la imaginació i pensar que es poden solucionar molts problemes. Això en part és veritat i en part és mentida, perquè una regla té un comportament sistemàtic, actua automàticament sempre que es compleixin les condicions formulades prèviament, i pot propiciar resultats estrambòtics aparentment inexplicables. Per exemple, com que són regles formulades amb condicions de contigüitat o juxtaposició, el traductor és capaç, malauradament, de processar l'element final d'un sintagma i el primer element del sintagma següent (no

separats per coma, no pas per error sinó perquè no cal que n'hi hagi) com si pertanyessin al mateix sintagma, amb resultats completament erronis.

Vegem-ho amb un altre exemple. Per aconseguir que el traductor distingeixi entre *haver* i *haver-hi* a partir d'*haber*, es va crear una regla que deia que "*haber* + NO VERBO = *hi* + *haver* + NO VERB". Aquesta regla permetia traduir bé "*había dos casos problemáticos*" per "*hi havia dos casos problemàtics*", però en canvi espatllava la traducció de "*había prácticamente terminado*", que convertia en "*hi havia pràcticament acabat*". Per tant, per evitar aquesta solució errònia es va haver de reformular la regla. En aquest cas, com en tots, la comparació del nombre d'encerts i errors de la regla ens permet valorar si ens convé mantenir-la, reformular-la o rebutjar-la, i així successivament amb cada regla.

## VALORACIÓ GLOBAL

Aquest exemple il·lustra la gran diferència entre la primera versió i la segona del traductor. Una gran diferència de la qual s'extreu una lliçó. Tenim un traductor més potent, més intel·ligent, més racional pel que fa a la seva arquitectura interna, que permet substituir quantitats infinites de seqüències per una sola regla (com és el cas de les concordances amb possessius esmentat més amunt), però tenim alhora un perill terrible que abans no teníem, i és que pot espatllar frases que en la versió anterior, sense regles, sortien ben traduïdes. Acostumats a una màquina simple que traduïa el que li dèiem mitjançant una equivalència tancada de paraula o seqüència i que no generava seqüeles, ens hem hagut de readaptar per veure a venir els danys col·laterals derivats de la potència de les regles, que també actuen més enllà dels casos específics per als quals han estat ideades (per haver estat mal formulades o perquè la llengua és més variada que no ens pensem).

Sigui com sigui (¡visca el progrés!), globalment hi hem sortit guanyant. La nova versió tradueix bé coses que l'antiga ens feia malament, i les que ens espatlla a causa de la potència de les regles les hem anat fent disminuir reformulant i fins i tot suprimint algunes d'aquestes mateixes regles. Perquè com a lingüistes al capdavant continuem preferint afrontar errors clars generats per la no intervenció del traductor automàtic, que no pas errors aparentment inexplicables causats per una sobreactuació dels processos automatitzats. Poder entendre fàcilment on i per què s'equivoca el traductor ens facilita molt la feina, tant de cara a la millora de l'eina com a l'actitud a prendre davant els resultats que proporciona.

Desembre 2006