ELS TREBALLS DEL PROJECTE EUROTRA A L'ESTAT ESPANYOL

Antoni BADIA i CARDÚS

Eurotra (Barcelona)

1. INTRODUCCIÓ

Presentar els treballs de l'equip d'EUROTRA de l'Estat espanyol és presentar els treballs de tot EUROTRA en general, ja que es tracta d'un projecte unitari en el qual, tot i que el treball és descentralitzat, hi ha, evidentment, una perspectiva de conjunt. Aquesta presentació, doncs, es referirà a tot el projecte i hi esmentarem les possibles particularitats del nostre equip, si s'escau. En primer lloc, farem un breu repàs del desenvolupament del projecte, que ens servirà per delimitar-ne els objectius i fer palès el marc en què treballem. A continuació, una breu descripció del marc organitzatiu ens farà veure la complexitat d'un projecte d'aquest tipus. Finalment, presentarem els trets fonamentals del treball científic del projecte en allò que anomenem el seu marc científic.

2. DESENVOLUPAMENT

El projecte EUROTRA ha nascut de la constatació que la traducció de documents ocupa una part molt important del personal administratiu de la Comunitat Econòmica Europea. A mesura que aquesta ha anat creixent, ha anat augmentant el nombre d'aparellaments lingüístics de traducció unidireccional: de les quatre llengües inicials (francès, alemany, holandès i italià) amb 12 aparellaments s'ha passat a la situació actual de nou llengües (les abans esmentades, més l'anglès, el danès, el grec, el portuguès i el castellà) amb 72 aparellaments de traducció unidireccional. I això sense tenir en compte les llengües que, com el català, no tenen de mo-

ment l'estatut de llengües oficials de la Comunitat! Com que és una decisió política de la Comunitat que no s'afavoreixi una llengua per sobre de les altres, la situació ideal és que tots els documents siguin traduïts a totes les llengües. Això comporta que del 35 al 65 per cent de les despeses dels diversos serveis de la Comunitat siguin produïdes per la traducció; i, encara, sense aconseguir l'objectiu de traduir tots els documents a totes les llengües.

A aquestes dades, cal afegir-hi el fet que no es poden anar augmentant ilHimitadament els serveis de traducció (per una banda, la capacitat de traducció de la persona humana, del millor professional de la traducció, és limitada i, per l'altra, també és limitada la capacitat de contractació de nou personal). A la vegada, cal tenir en compte que la situació a l'interior de l'administració de la Comunitat és només una mostra de les necessitats de traducció de la societat (indústria, comerç, etc.) en el marc d'un mercat únic. Per tots aquests fets, la Comunitat va decidir patrocinar un projecte de recerca i desenvolupament (R + D) en el camp de la traducció automàtica. La decisió va ser presa formalment en la reunió del Consell de Ministres de la Comunitat que tingué lloc el 4 de novembre de 1982.

En la mateixa decisió del Consell de Ministres es marquen els dos objectius bàsics del projecte:

1. Construir el prototipus pre-industrial d'un sistema de traducció automàtica per als idiomes oficials de la Comunitat (sis en aquell moment; nou, ara).
2. Promoure la investigació i els estudis avançats en la traducció automàtica i camps afins.

Com es pot veure, les finalitats del projecte EUROTRA es mouen en dues direccions. En primer lloc, hi ha l'aspecte tècnic de creació d'un sistema de traducció automàtica. Aquest sistema no ha estat concebut per tractar tota mena de textos, sinó que és pensat per a textos no literaris d'unes característiques determinades (per exemple, textos tècnics o de l'administració). Cal tenir present que no és una feina pròpia del projecte elaborar el producte acabat que hauran d'usar els funcionaris de la Comunitat o que es podrà posar a la venda. El sistema que surti del projecte EUROTRA haurà de passar després per una fase d'industrialització que haurà de ser realitzada per la indústria informàtica. Per això, les característiques del prototipus elaborat han de ser tals que el sistema pugui ser fàcilment industrialitzat i conservat o ampliat posteriorment; és a dir, el seu manteniment i millora han de ser possibles. Això dóna una justificació addicional a una de les característiques pròpies del sistema, la seva modularitat, de la qual parlarem més endavant (vegeu l'apartat 4.2).

En segon lloc, i no menys important, el projecte pretenia, des del mo-

ment de la seva creació, estimular els estudis a Europa en els camps relacionats amb el processament del llenguatge natural (camps en els quals tant els Estats Units d'Amèrica com el Japó havien pres la davantera). Això només era possible si a la vegada s'estimulava la collaboració entre els diversos Estats i entre les diverses disciplines (lingüistes, informàtics, lògics, etc.). D'aquí naixerà un dels aspectes organitzatius cabdals del projecte: la descentralització . Malament es podia pensar a estimular el treball i la investigació en cada un dels Estats membres o a potenciar la collaboració entre aquests mateixos Estats amb una estructura centralitzada. Calia organitzar el projecte de manera descentralitzada i amb un cert grau d'autonomia per a cada equip de treball. Com que, per altra banda, el resultat final havia de ser un producte únic, calia organitzar bé les tasques de coordinació i control. D'aquí sorgirà la complexa estructura organitzativa que presentarem breument més endavant.

El projecte s'ha de desenvolupar en tres fases:

1. Primera fase, preparatòria (de 1983 a 1985):
   — Creació de la infrastructura (tant d'organització, com tècnica),
   — Creació dels grups lingüístics en cada Estat membre i
   — Definició de les especificacions lingüístiques i informàtiques bàsiques.

2. Segona fase (de 1985 a 1988):
   — Desenvolupament del *software* adequat per provar les eines lingüístiques i
   — Investigació lingüística bàsica:
     • Estudi dels models lingüístics per a l'anàlisi i la síntesi de cada una de les llengües i per a la transferència entre elles,
     • Preparació de la base de dades lèxica (amb unes 2.500 entrades per a cada llengua) i
     • Estudi de les estratègies lingüístiques adequades per a l'execució dels diversos processos per part de la màquina.

3. Tercera fase (de 1988 a 1990):
   — Ampliació del vocabulari (fins a unes 20.000 entrades de diccionari),
   — Millorament del *software,*
   — Revisió de les eines lingüístiques i ampliació dels fenòmens no tractats encara, i
   — Avaluació de les capacitats tècniques del sistema produït.

Pel que fa a l'estadi actual del projecte, cal dir que s'estan aconseguint els objectius marcats i que, per tant, s'està a l'inici de la tercera fase. A l'Estat espanyol (que, com Portugal, no va entrar operativament al pro-

jecte fins a mitjan 1987) estem pràcticament al nivell dels altres equips lingüístics. Val a dir que el grup lingüístic de l'Estat està dividit entre Barcelona i Madrid. A Barcelona hi ha l'equip central, amb la coordinació, i s'hi desenvolupa la sintaxi, la semàntica i els diccionaris bilingües. A Madrid tenen la responsabilitat de la morfologia i el lèxic.

3. El marc organizatiu

Tal com ja ho hem mencionat, en cada Estat membre de la CEE hi ha un grup lingüístic. En la majoria dels casos aquest grup lingüístic es dedica a la pròpia llengua oficial. De tota manera, com que d'Estats membres n'hi ha dotze i, en canvi, només hi ha nou llengües oficials, no tots els equips es dediquen a una única llengua oficial. Així, a Bèlgica hi ha dos subgrups lingüístics que treballen, l'un en col·laboració amb el grup francès sobre el francès, i l'altre sobre el flamenc en col·laboració amb el grup holandès. A Irlanda són responsables de la terminologia i a Luxemburg hi ha el centre de documentació. Als altres Estats, cada grup lingüístic s'ocupa de la pròpia llengua oficial. Són diversos els casos en què hi ha dos equips en un mateix Estat, de manera que l'aspiració de descentralització del projecte arriba fins i tot a l'estrutcura interna dels grups lingüístics dels Estats membres. Cada grup lingüístic és el responsable de l'anàlisi i la síntesi de la seva llengua oficial, així com de la transferència de les altres vuit llengües a la seva.

A més dels grups lingüístics hi ha, evidentment, unes estructures de coordinació i control que són les encarregades de donar cohesió al projecte. En primer lloc, el Grup d'Enllaç, en el qual són representats la Comissió Europea i cada un dels grups lingüístics, és el que porta la direcció tècnica i administrativa del projecte. La Comissió s'encarrega de les tasques de gestió i administració. Els Grups Especials d'Investigació són creats d'acord amb les necessitats i s'encarreguen de l'assistència tècnica als grups lingüístics. Finalment, hi ha el Comitè Consultiu, que dóna suport tècnic a les tasques de la Comissió, i el Comitè Supervisor, que controla la marxa del projecte; en tots dos són representats la Comissió i cada un dels Estats membres.

4. El marc científic

4.1. *Una opció per la transferència*

A l'inici de la traducció automàtica es pensava que era possible la traducció paraula per paraula o, almenys, amb un mínim d'anàlisi estruc-

tural. En adonar-se que això era impossible, l'eufòria que presidí els primers intents de traducció automàtica es va esvanir. De fet, fou l'informe ALPAC el que va constituir el veritable gerro d'aigua freda per a la traducció automàtica dels anys cinquanta i seixanta. Posteriorment, tant la informàtica com la lingüística teòrica (especialment la teoria de la sintaxi) han sofert un desenvolupament tan extraordinari que ja no és una utopia tornar a parlar de traducció automàtica. En aquesta represa els punts de vista han estat molt menys pretensiosos que els de la dècada dels cinquanta. És en aquest context que s'ha de comprendre la decisió d'EUROTRA de basar el seu procés de traducció en la transferència i no en una interlingua.

En un projecte de traducció automàtica, i més encara si afecta moltes llengües, la situació ideal seria tenir un nivell de representació de les frases de les llengües que fos prou abstracte perquè el procés d'anàlisi i de síntesi de cada llengua hi portés o en sortís sense dificultat. Aquest nivell de representació abstracte, comú a totes les llengües, és el que s'anomena *interlingua*. En la figura 1 veiem representat el que seria un procés de traducció automàtica basat en una interlingua.

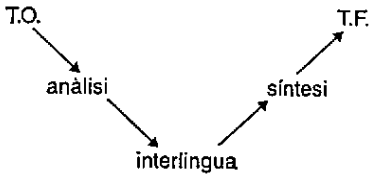


Figura 1

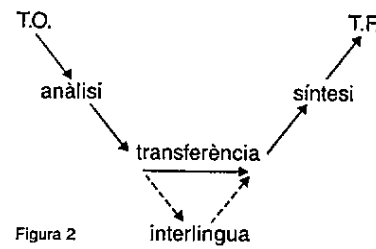


Figura 2

El text que s'ha de traduir (el text origen; T.O. en l'esquema) sofriria una sèrie d'anàlisis el resultat final de les quals seria una representació del text en la interlingua; a continuació, aquesta representació abstracta del text d'origen sofriria un procés de síntesi o generació fins que al final en sortiria el text final (T.F. en l'esquema). Com que el nivell de profunditat o d'abstracció de la representació en la intelingua seria el mateix per a totes les llengües a les quals s'apliqués el procediment, un projecte com EUROTRA (amb nou llengües involucrades en l'actualitat) hauria de construir únicament nou mòduls d'anàlisi i nou mòduls de síntesi. No caldria cap mòdul d'enllaç d'una llengua amb l'altra: el punt d'enllaç, el constituiria la representació comuna en la interlingua. El problema amb aquesta opció és que, de moment, no és gens clar com ha de ser representada aquesta interlingua. Ni la semàntica, a nivell lingüístic, ni la representació del coneixement, a nivell de la intel·ligència artificial, estan prou avançades

per permetre de pensar que s'és a punt d'aconseguir una representació d'interlingua.

Fonamentalment per aquesta raó, EUROTRA ha optat pel model de transferència. La característica fonamental d'aquest model és que el nivell de profunditat de la representació més abstracta no es troba al nivell de confluència entre totes les llengües involucrades (com en el cas de la interlingua), sinó que es troba en un punt intermedi en els processos d'anàlisi i de síntesi. La figura 2 representa esquemàticament aquest procés. Les conseqüències pràctiques són, per una banda, que el projecte és realitzable en un termini més o menys curt, ja que no té la barrera, de moment infranquejable, de la interlingua; i, per altra banda, que als mòduls d'anàlisi i de síntesi cal afegir els mòduls de transferència de la representació profunda d'una llengua a la de l'altra. En concret, a EUROTRA (amb, de moment, nou llengües) hi ha 72 mòduls de transferència (nou llengües que han de ser traduïdes a unes altres vuit: $9 \times 8 = 72$).

Un altre aspecte a tenir en compte és que un model de traducció automàtica basat en la transferència ha de decidir a quin nivell de profunditat colloca la transferència. Això, certament, depèn en gran mesura de les llengües involucrades. Com més distants siguin aquestes entre elles, a més profunditat hauran d'intervenir els mòduls de transferència. En el cas d'EUROTRA, la similaritat entre les llengües involucrades (totes són de la gran família indoeuropea: quatre de romàniques, quatre de germàniques i la grega) ha permès una llibertat bastant gran a l'hora d'escollir el nivell de profunditat en el qual s'ha de realitzar la transferència. A la pràctica aquesta es fa a un nivell bastant simple de representació semàntica.

4.2. *Modularitat*

El procés es realitza per mòduls independents els uns dels altres. Per exemple, els mòduls de l'anàlisi de l'anglès són independents de les característiques dels mòduls de síntesi de l'alemany, de manera que, en primer lloc, l'organització del treball del grup anglès no ha d'estar pendent de l'organització del grup alemany, i, en segon lloc, l'únic punt de contacte entre l'anglès i l'alemany (en el cas d'una traducció del primer al segon) és el mòdul de transferència de l'anglès a l'alemany).

Aquesta característica, justificada àmpliament per tota la pràctica informàtica i les tècniques de programació, respon també a una necessitat pràctica en funció de la mateixa concepció del projecte. Seria impossible el treball descentralitzat i amb molts quilòmetres de distància entre els diversos centres d'investigació si no fos per la independència entre els mòduls en què es descompon el sistema.

Un altre aspecte a tenir en compte és el manteniment i la millora del

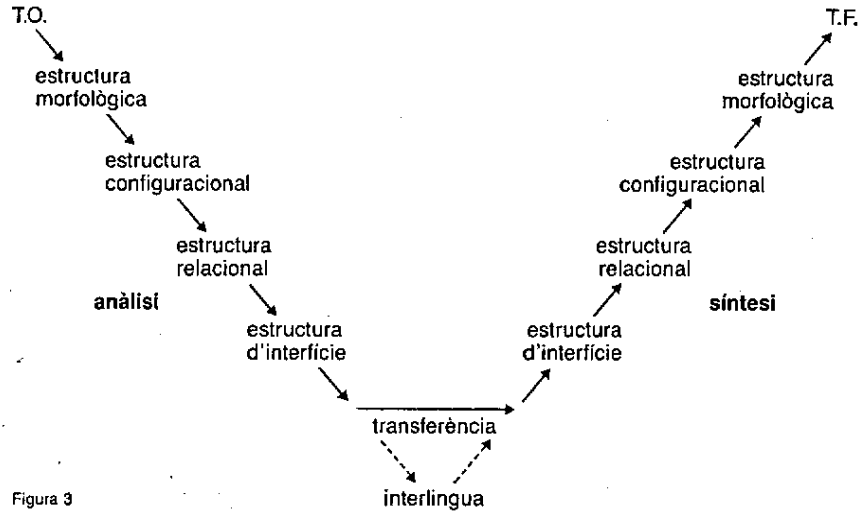


Figura 3

sistema, que també es faciliten amb una estructuració per mòduls. Un retoc en un d'ells no afecta tots els altres. Això facilita extraordinàriament el treball per tal com es pot anar provant el funcionament del sistema a mesura que es va construint sense que esdevingui una pèrdua de temps o d'energies.

El procés d'anàlisi i el de síntesi són descompostos en quatre mòduls diferents, que donen una representació del text a un nivell cada vegada més profund. En la figura 3 tenim representada esquemàticament l'estructura global del procés. Així que el text d'origen (T.O.) ha sofert les adaptacions necessàries per poder ser processat (tractament de caràcters especials, uniformització de lletra, etc.), el primer mòdul d'anàlisi ofereix com a resultat una *estructura morfològica.* Aquesta és processada pel segon mòdul, que dóna com a resultat una *representació configuracional* o de sintaxi de constituents. El pas següent és realitzat pel mòdul de l'estructura relacional, que produeix una *representació relacional.* Aquesta, al seu torn, és processada per l'últim mòdul d'anàlisi per tal de produir la *representació d'interfície,* que és la que ha de ser l'entrada dels mòduls de transferència. El procés de síntesi és similar al d'anàlisi amb la diferència que actua en sentit invers, de la representació més profunda a la cadena del text.

Tècnicament, cada un d'aquests nivells de representació està constituït per un constructor, és a dir, una gramàtica que crea les representacions pròpies del seu nivell. A més, cada nivell està enllaçat amb el següent mitjançant un traductor que recull les representacions del nivell de sortida

i les adequa per tal de ser representacions del nivell d'entrada, que seran *consolidades* o acceptades pel seu constructor.

El que fins ara hem estat anomenant mòduls de transferència no són més que traductors, la característica diferenciadora dels quals és que estan formats sobretot per regles de traducció lèxica.

4.3. *Composicionalitat*

Si s'aplica el concepte de la composicionalitat a la traducció, resulta que la traducció d'una expressió és composicional quan està relacionada amb la traducció de les seves parts d'una manera sistemàtica. Dit d'una altra manera, la traducció d'una expressió és una funció de la traducció de les seves parts i de la manera com aquestes parts s'hi combinen.

Aquest concepte de composicionalitat s'aplica, a EUROTRA, a tots els traductors del procés; tant si formen part del procés d'anàlisi o del de síntesi, com si es tracta dels mòduls de transferència. Actualment hi ha en el sistema d'EUROTRA algun mecanisme no composicional, però s'intenta que tingui el menor pes possible.

4.4. *Unificació*

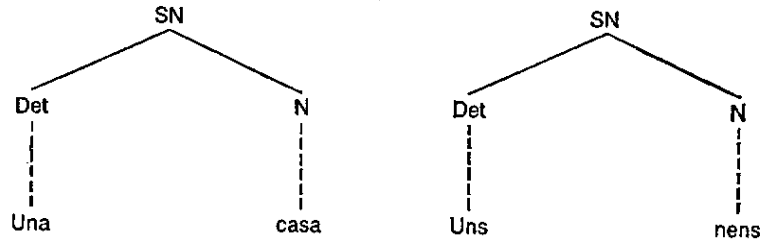
La unificació és un dels mecanismes més potents i, a la vegada, més usats en les gramàtiques actuals. És un concepte que originàriament neix en la teoria informàtica de la resolució de problemes i que ha estat adoptat amb èxit per la gran majoria de les teories o formalismes gramaticals que s'usen en les aplicacions computacionals de la lingüística.

Els objectes construïts per una gramàtica són arbres. Els nusos terminals dels arbres són les paraules (o morfemes, si es tracta de la representació morfològica), mentre que els seus nusos no terminals són els que defineixen l'estructura de l'expressió en qüestió; són els anomenats *sintagma nominal, sintagma verbal,* etc. Ara bé, aquests nusos no són tan sols uns punts en una estructura arbòria, sinó que contenen informació, que tant pot ser morfològica, com sintàctica o semàntica. Aquesta informació és continguda o bé en conjunts o bé en estructures de trets. Cada tret és un parell ordenat d'un atribut i un valor.

Per exemple, a un nivell sintàctic configuracional, podríem tenir les representacions arbòries *a'*) i *b'*) corresponents als sintagmes següents *a*) i *b*):

a) Una casa
b) Uns nens

a')



Alguns dels trets que podrien formar part de la informació relacionada amb el nus terminal de 'una' són:

c) < gènere = femení, nombre = singular>,

i els trets següents podrien formar part de la informació del nus terminal de 'uns':

d) < gènere = masculí, nombre = plural >.

En tots dos casos tenim un conjunt de dos trets, el primer dels quals té com a atribut el *gènere* i el segon, el *nombre*; en cada cas el valor de l'atribut és diferent.

Doncs bé, la unificació és una operació que controla el flux de la informació en una estructura arbòria. La podem definir com aquella operació que combina la informació de dos conjunts (o estructures) de trets per obtenir un altre conjunt (o estructura) de trets que contingui tota la informació dels dos, sempre que aquesta informació no sigui incompatible. Els conjunts de trets *c*) i *d*) no es poden unificar perquè tota la informació que contenen és incompatible. En canvi, els conjunts de trets *e*) i *f*) sí que es poden unificar; el conjunt resultant de la seva unificació és *g*):

e) < gènere = masculí, nombre = plural >
f) < persona = tercera >
g) < gènere = masculí, nombre = plural, persona = tercera >.

Podem fer servir aquest mecanisme en les regles del constructor per tal de controlar la bona formació dels sintagmes. Si tenim una regla de formació del sintagma nominal com *h*) i uns nusos terminals com *i*) i *j*), els resultat d'aplicar aquesta regla *h*) a aquests nusos terminals és, per unificació, la representació *k*):

h) SN<gènere=G, nombre=N, persona=P, determinació=D, nucli=M>

→

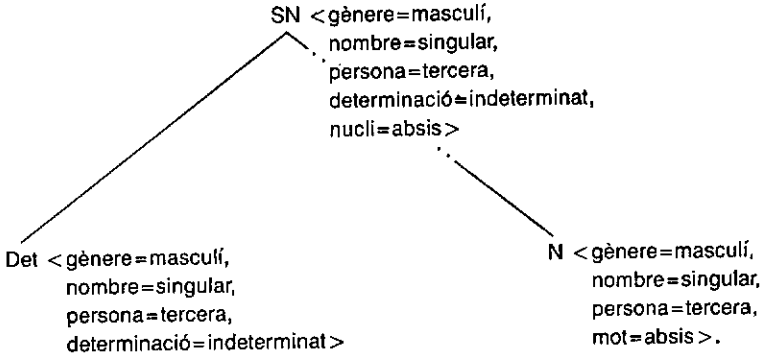
Det<gènere=G, nombre=N, persona=P, determinació=D>.

N<gènere=G, nombre=N, persona=P, mot=M.

i) un<gènere=masculí, nombre=singular, determinació=indeterminat>.

j) absis<gènere=masculí, persona=tercera, mot=absis>.

k)



Doncs bé, les gramàtiques d'EUROTRA, com gairebé tots els formalismes gramaticals actuals, usen aquest mecanisme d'una manera molt semblant a com l'hem descrita suara.

5. Conclusió

El projecte EUROTRA ha estat concebut, des del seu origen, com un projecte d'investigació per desenrotllar un sistema de traducció automàtica entre les llengües oficials de la CEE i per promoure la recerca en el camp de la lingüística computacional, així com la collaboració entre especialistes en diverses branques i de diversos Estats.

Per les seves característiques tècniques, el projecte s'inclou dintre del gran corrent de treball de la lingüística computacional actual, és a dir, el de les gramàtiques o formalismes basats en la unificació.

A més, les seves característiques organitzatives li donen un caràcter únic en el món de la traducció automàtica actual.