Informàtica Aplicada a la Traducció: notes de classe

Mikel L. Forcada i Juan Antonio Pérez Ortiz
{mlf,japerez}@dlsi.ua.es
http://www.dlsi.ua.es/~mlf
Departament de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant

Curs 2001–2002, versió del 29.10.2001

Índex

1 Introducció **1**

2 Ordinadors i programes **3**
2.1 Maquinari . . . 3
2.2 Programari . . . 5
2.3 Tipus d'ordinadors . . . 9
2.4 Memòria . . . 10
2.5 Configuració d'un ordinador personal . . . 11
2.6 Sobre algorismes . . . 11
2.7 Un petit glossari . . . 13
2.8 Qüestions i exercicis . . . 15
2.9 Solucions . . . 18

3 Algunes nocions bàsiques sobre Internet **19**
3.1 Què és Internet? . . . 19
3.2 Números IP . . . 19
3.3 Noms . . . 20
3.4 Adreces de correu electrònic . . . 21
3.5 News . . . 21
3.6 Els localitzadors . . . 21
3.7 Navegadors . . . 22
3.8 Accés domèstic a Internet . . . 23
3.9 Exercicis i qüestions . . . 23
3.10 Solucions . . . 24

4 L'entrada i el processament de textos **25**
4.1 Formats de text . . . 25
4.1.1 L'ASCII original de 7 bits . . . 25
4.1.2 Els ASCII estesos a 8 bits . . . 26
4.1.3 Unicode . . . 26
4.1.4 Limitacions . . . 28
4.1.5 Formats més avançats . . . 28
4.1.6 SGML i XML . . . 29

4.1.7 HTML . . . 31
4.1.8 RTF . . . 32
4.2 Processadors de textos . . . 34
4.3 Reconeixement automàtic de la parla . . . 37
4.4 Reconeixement automàtic de textos escrits . . . 38
4.5 Qüestions i exercicis . . . 39
4.6 Solucions . . . 39

5 Bases de dades **41**
5.1 Bases de dades lèxiques o terminològiques . . . 43

6 Traducció i traducció automàtica **45**
6.1 Què és la traducció? . . . 45
6.2 Traducció automàtica . . . 47
6.2.1 Definició . . . 47
6.2.2 Sobre el nom en altres llengües . . . 48
6.3 Història de la traducció automàtica . . . 48
6.3.1 Els pioners, –1954 . . . 48
6.3.2 El decenni de l'optimisme, 1954–1966 . . . 48
6.3.3 Després de l'informe ALPAC (1966) als vuitanta . . . 50
6.3.4 Els primers vuitanta . . . 50
6.3.5 Els primers noranta . . . 51
6.3.6 Del darrers noranta a l'actualitat . . . 52

7 Usos de la traducció automàtica **53**
7.1 Utilitat de la traducció automàtica . . . 53
7.2 Assimilació i disseminació . . . 54
7.2.1 Assimilació . . . 54
7.2.2 Disseminació . . . 56
7.3 CAT, HAMT i MAHT . . . 56
7.4 Preedició i postedició . . . 57
7.5 Llenguatges controlats . . . 57
7.6 Qüestions, exercicis i problemes . . . 59
7.7 Solucions . . . 60

8 Ambigüitat **63**
8.1 Ambigüitat deguda a l'ambigüitat lèxica . . . 64
8.2 Ambigüitat estructural pura . . . 67
8.3 Ambigüitats mixtes . . . 72
8.4 Ambigüitat deguda a l'abast dels quantificadors . . . 74
8.5 Estratègies de resolució de l'ambigüitat . . . 74
8.5.1 Resolució de l'ambigüitat lèxica categorial . . . 75
8.5.2 Resolució de la polisèmia . . . 75
8.5.3 Resolució de l'anàfora . . . 76

8.5.4 Resolució de l'ambigüitat estructural 76
8.6 Qüestions, exercicis i problemes 78
8.7 Solucions 80

9 Tècniques de traducció automàtica 83
9.1 Traducció directa i traducció indirecta 83
9.2 Traducció *directa*? 84
9.3 Traducció indirecta per transferència 84
9.3.1 Sistemes de transferència morfològica 85
9.3.2 Anàlisi i generació morfològiques 87
9.3.3 Sistemes de transferència sintàctica 92
9.3.4 Anàlisi sintàctica 96
9.3.5 Sistemes de transferència semàntica 99
9.4 Sistemes sense transferència: els sistemes d'*interlingua* 100
9.5 Qüestions, exercicis i problemes 102
9.6 Solucions 105

10 Memòries de traducció 107

11 Avaluació de la traducció automàtica 111
11.1 Qüestions bàsiques 111
11.2 Tipus d'avaluació 111
11.3 Traducció automàtica i traducció humana 114
11.4 Avaluació predictiva 115

12 Problemàtica de la TA castellà–català 117
12.1 Introducció 117
12.2 Segmentació del text origen 118
12.3 Homografia 118
12.4 Divergències de traducció 121
12.5 Qüestions i exercicis 122
12.6 Solucions 123

13 Experiències de TA castellà–català 125
13.1 SALT, de la Generalitat Valenciana 125
13.2 Ara, d'Autotrad 126
13.3 Es-Ca, de Incyta 126
13.4 El traductor d'*El Periódico de Catalunya* 126
13.5 interNOSTRUM 126
13.5.1 Característiques informàtiques 127
13.5.2 Característiques lingüístiques 127
13.5.3 Eines de suport a interNOSTRUM 131

Capítol 1

Introducció

Aquestes pàgines contenen una bona part dels continguts[1] de l'assignatura *Informàtica Aplicada a la Traducció* que cursarà l'alumnat de tercer curs de la llicenciatura en Traducció i Interpretació de la Universitat d'Alacant. La lectura d'aquestes notes —que poden fins i tot contenir algun error no detectat— no pot mai substituir l'estudi de bons llibres sobre la matèria, molts dels quals se citen en el text i es llisten en la bibliografia.

Veureu que els continguts d'aquestes notes es poden dividir en tres parts: la primera presenta alguns conceptes bàsics de la informàtica (capítol 2), i, més concretament, d'Internet (capítol 3), sobre l'entrada i el processament de textos (capítol 4), i sobre les bases de dades (capítol 5); la segona és una introducció a alguns aspectes generals de la traducció automàtica (capítols 6 a 11), i la tercera, un recull d'aspectes més concretament referits a la traducció castellà–català (capítols 12 i 13), els quals poden servir com a il·lustració de l'estudiat en un cas concret. Els continguts d'aquesta tercera part són, per tant, complementaris.

De fet, aquestes notes es poden millorar molt. No seria gens estrany que s'editara una nova versió de les notes cap al final del curs. Heus ací una llista d'algunes de les coses que sabem que ens queden per fer tan prompte com siga possible:

- S'ha de millorar molt i actualitzar el capítol que tracta sobre conceptes bàsics de la informàtica. Hi falten referències bibliogràfiques.
- La descripció dels conceptes de fitxer i directori és molt curta i cal ampliar-la, amb gràfics si és possible.
- Cal explicar com funcionen els diversos tipus d'impressora.
- Falta una secció que descriga recursos per a traductors que es poden

[1] Hi ha continguts —diguem-ne millor habilitats— que s'aprenen com a part de les sessions de laboratori i que no figuren en aquest document.

trobar en Internet: diccionaris, traductors, lliçons introductòries (*tutorials*), etc.

- Cal explicar altres modalitats d'accés domèstic a Internet (ADSL, cable).
- Hi ha capítols sobre temes de molt interés que encara són massa curts (per exemple, els capítols 5 i 10); en tots dos falten esquemes i gràfics que expliquen millor els conceptes i els processos.
- S'ha de millorar la descripció de les tècniques de resolució de l'ambigüitat lèxica de transferència (polisèmia), tot donant un exemple de xarxa semàntica.
- Cal millorar l'apartat sobre resolució de l'ambigüitat estructural pura.
- Cal ampliar l'apartat sobre llenguatges controlats.
- S'ha d'ampliar la descripció dels sistemes de traducció automàtica basats en transferència semàntica i en interlingua.
- Cal actualitzar les descripcions de sistemes de traducció automàtica castellà–català que hi ha al final del llibre.
- Cal ampliar el nombre de qüestions i exercicis d'alguns capítols (alguns, de fet, no en tenen), recuperant els d'exàmens, etc.

A més, és possible que hi haja errades que s'hagen de corregir. El text està obert, per descomptat, a suggeriments i a correccions que el facen més útil, tant a l'alumnat de l'assignatura com a altres persones que volen saber sobre el tema.

Capítol 2

Ordinadors i programes

Tots els sistemes informàtics[1] es poden dividir en dues parts: *maquinari* i *programari*.

Maquinari (o *hardware*): l'equipament físic que es pot veure i tocar. Per exemple, la pantalla, el processador central, el teclat, els xips[2] de memòria i les impressores.

Programari (o *software*): un o més *programes* (i les dades associades) que fan alguna funció útil. Per exemple, un processador de textos com Microsoft Word o WordPerfect pot estar compost per més d'un *programa*. Un *programa* és un conjunt ordenat (llista) d'instruccions que són seguides pel maquinari, de tal manera que realitzen alguna tasca determinada. Normalment, els ordinadors estan organitzats al voltant d'un *processador central* (vegeu més endavant) que és capaç de comprendre i executar instruccions preses d'un conjunt determinat (el *conjunt d'instruccions* del processador). Els programes poden estar escrits a mà en un paper, guardats en un disc o carregats en la memòria de l'ordinador mentre són executats pel processador.

Ara es consideren el programari i el maquinari amb més detall.

2.1 Maquinari

Tots els sistemes informàtics tenen maquinari de les classes següents:

Entrada: la funció primària dels dispositius d'entrada és que l'usuari puga interactuar amb la màquina i amb els programes que executa amb la

[1] És a dir, totes les instal·lacions basades en ordinadors

[2] El xip és l'element bàsic de la microelectrònica i de la microinformàtica; es tracta d'un o més circuits integrats en una placa de silici de dimensions molt reduïdes, que normalment es col·loca en una capsa hermètica amb contactes metàl·lics.

finalitat d'*introduir-hi* dades o informació. El dispositiu d'entrada més comú és el teclat. N'hi ha d'altres, com el ratolí o la maneta de jocs, o l'escànner, un dispositiu que llegeix una imatge impresa i la converteix en un fitxer (vegeu la pàg. pg:fitxer) que conté la imatge digitalitzada[3].

Emmagatzematge: els dispositius d'emmagatzematge es poden dividir en dos grups:

Memòria primària: memòria ràpida de curt termini, volàtil (s'esborra quan s'apaga l'ordinador), que serveix per a guardar-hi programes i dades mentre l'ordinador està funcionant. Normalment consisteix en xips RAM (*random-access memory*[4]) de silici.

Memòria secundària: memòria de llarg termini, permanent. Exemples: disquets, discos fixos o durs, unitats de cinta, i diverses formes de ROM (*read-only memory*, memòria de lectura només), com els xips ROM o el CD-ROM. Els disquets, els discos fixos i les cintes són dispositius d'emmagatzematge magnètic, poc més o menys com ho són les cassets. La memòria ROM sol estar feta de xips de silici. Els CD-ROM —idèntics en aparença i molt similars en molts aspectes als CD de música— emmagatzemen la informació òpticament.

Fitxers i directoris: És comú que les dades emmagatzemades en memòria secundària estiguen organitzades en *fitxers*[5] (conjunts de dades amb un nom i que es manipulen —s'obrin, es tanquen, es copien, s'esborren— com un tot) i que els fitxers estiguen organitzats en *directoris*[6] (fitxers especials que agrupen els noms i les característiques d'altres fitxers), de manera jeràrquica o arbòria (els directoris poden contenir fitxers o també altres directoris, i així successivament). Normalment, cada disc té un *directori principal* o *directori arrel* (el més elevat en la jerarquia de directoris). Dos fitxers —també dos directoris— només poden tenir el mateix nom si es troben en directoris diferents. Per raons històriques, els noms de fitxers solen tenir dues parts: el *nom* pròpiament dit i l'*extensió*, separades per un punt (per exemple,

[3] Quan la imatge és la d'un text imprés, un programa de *reconeixement òptic de caràcters* (OCR, *optical character recognition*) la pot convertir en una representació del text adequada per a ser manipulada amb un processador de textos (vegeu la secció 4.2), generalment amb alguns errors tipogràfics menors.

[4] La raó d'aquest nom és l'oposició entre *accés aleatori* (*random*, a voluntat) i *accés seqüencial*; per exemple, a les dades enregistrades en una cinta magnètica —com, per exemple, la d'una casset— només es pot accedir seqüencialment: per a arribar a la dada número 1000 hem de passar per les 999 anteriors.

[5] També anomenats *documents* o *arxius*.

[6] També anomenats *carpetes*.

`alacant.txt`. El nom sol ser normalment lliure, però l'extensió sol ser curta (entre una i quatre lletres) i sol identificar el programa que s'ha d'usar per a processar-lo o el format en què es troben les dades que conté (per exemple, l'extensió `.txt` identifica normalment un text senzill en format ASCII o ANSI, vegeu l'apartat 4.1).

Processament: Els dispositius de processament són els que fan realment el treball. La majoria dels sistemes contenen una CPU (*central processing unit*, unitat central [de processament]), o senzillament, un *processador* que és responsable d'executar totes les instruccions de programa, de processar dades, i de controlar el funcionament d'altres components del maquinari. En els ordinadors personals, la unitat central és un únic xip de silici.

Eixida: Aquesta és la família dels dispositius que l'ordinador usa per a comunicar dades o informació a l'usuari. El monitor (la pantalla) n'és el més comú. Altres dispositius d'eixida són les impressores, les traçadores gràfiques, etc.

2.2 Programari

Com ja s'ha dit més amunt, un programari és un conjunt de programes, cada un dels quals consisteix en una llista d'instruccions vàlides (executables per l'ordinador) que s'executen en l'ordre indicat, de la primera a l'última, excepte quan s'hi presenta alguna instrucció de *salt* que indica quina és la següent instrucció que s'ha de executar.

Per exemple, un programa que suma tots els nombres enters del 1 al 10 podria ser el següent, el qual usa dues posicions de memòria RAM per a guardar valors necessaris per al càlcul. Cada una de les ordres es correspon amb una instrucció bàsica de les que pot entendre qualsevol processador.

1. Fes que l'acumulador (un registre de la memòria interna del processador) valga 1.
2. Guarda el valor de l'acumulador en una posició de memòria que anomenarem *índex*.
3. Fes que l'acumulador valga 0.
4. Guarda el valor de l'acumulador en una posició de memòria que anomenarem *suma*, la qual contindrà la suma total.
5. Carrega el valor de *suma* en l'acumulador.
6. Suma el valor d'*índex* a l'acumulador.

7. Guarda el valor de l'acumulador en *suma*.
8. Carrega el valor d'*índex* en l'acumulador.
9. Compara el valor de l'acumulador amb 10.
10. Si és igual, salta a la instrucció 14
11. Incrementa en 1 el valor de l'acumulador.
12. Guarda el valor de l'acumulador en *índex*.
13. Salta a la instrucció 5.
14. Para.

Moltes voltes s'usen noms curts (en anglés *mnemonics*) per a les instruccions del processador i també noms elegits pel programador per a referir-se a posicions del programa (aquesta notació se sol anomenar *llenguatge assemblador*). El programa de dalt tindria l'aparença següent:

```
        mov #1,A
        mov A,index
        mov #0,A
        mov A,suma
altre:  mov suma,A
        add A,index
        mov A,suma
        mov index,A
        cmp A,#10
        jeq final
        inc A
        mov A,index
        jmp altre
final:  hlt
```

Programes i algorismes: Moltes voltes, un programa és la realització (entre informàtics se'n diu *implementació*) d'un *algorisme*. Un algorisme (també *algoritme*, per interferència amb *aritmètica*) és una seqüència finita d'operacions executables i no ambigües[7] que defineix un procediment que sempre es deté. Normalment els algorismes especifiquen un mètode general per a obtenir la resposta (correcta o incorrecta) a qualsevol cas particular d'un problema o pregunta, com ara "quin és el quocient de la divisió de dos nombres enters"?

Els algorismes es poden convertir fàcilment en programes d'ordinador si es canvien les instruccions bàsiques de l'algorisme per instruccions que

[7] però no necessàriament comprensibles per a un ordinador

entenga l'ordinador corresponent. Al final d'aquest document hi ha una secció dedicada als algorismes i a la seua relació amb els programes.

Tipus de programari: Hi ha tres classes bàsiques de programari:

Sistemes operatius i *firmware*: són els programes que permeten el funcionament bàsic de l'ordinador. Per exemple, és el sistema operatiu qui controla la visualització en pantalla. S'anomena *firmware* el programari del sistema que s'usa tan freqüentment que s'emmagatzema permanentment en xips ROM.

El sistema operatiu és el primer programa que comença a executar-se quan connectem l'ordinador. Permet que l'usuari hi execute programes i gestione els fitxers de dades. Quant a l'aparença i la forma d'interaccionar amb l'usuari, els sistemes operatius es poden dividir en *gràfics* (Windows, Macintosh OS, OS/2) i de *línia d'ordres* (Unix primigeni, VAX/VMS, MS-DOS). Els sistemes poden ser també *monousuari* (MS-DOS, Windows 3.11) o *multiusuari* (Unix, VAX/VMS), si poden o no donar accés i suport a més d'un usuari alhora, o *monotasca* (MS-DOS) i *multitasca* (quasi tots els altres), si poden o no executar més d'un programa alhora. Hi havia també sistemes operatius especialitzats per a connectar ordinadors formant una xarxa, com Novell Netware, però són poc usats perquè els sistemes operatius actuals ja estan preparats per a les xarxes més bàsiques[8]

Algunes de les operacions bàsiques que fan els sistemes operatius són:

- Copiar, moure i esborrar fitxers de dades.
- Crear, moure i esborrar directoris de fitxers.
- Establir connexions entre ordinadors.
- Executar programes i controlar-ne l'execució.

De fet, els altres programes solen estar escrits per a ser executats *sobre un sistema operatiu*[9], és a dir, assumeixen que el sistema operatiu farà totes aquestes operacions senzilles i no contenen instruccions de programa per a fer-les, sinó només instruccions per a invocar els programes corresponents del sistema operatiu. Això simplifica enormement l'escriptura de programes d'ordinador.

Processadors de llenguatges de programació: les instruccions que executa el processador central d'un ordinador són massa senzilles perquè

[8] L'organització dels ordinadors en una xarxa permet la comunicació d'informació entre ells i la compartició de recursos, com ara una impressora. Internet (vegeu el capítol 3) no és més que una gran xarxa global que interconnecta moltes xarxes més locals.

[9] I, per tant, quan comprem un programa, hem d'especificar per a quin sistema operatiu el volem.

un programador humà en faça programes útils; seria llarg i enutjós, com hem vist en l'exemple de programa que sumava els enters de l'1 al 10. Els programadors normalment escriuen els seus programes en *llenguatges de programació* basats en instruccions més potents (com ara BASIC, FORTRAN, C, Pascal) i usen programes especials —els processadors de llenguatges— per a traduir-los a les instruccions senzilles que entén la màquina.[10] Quasi tots els programes que s'executen en un ordinador han estat escrits en algun llenguatge de programació. El programa que suma els nombres de l'1 al 10 quedaría així en Pascal:

```
program SUMA;
var
   index, suma: integer;
begin
   suma:=0;
   for index:=1 to 10
      suma:=suma+index;
end.
```

Programes d'aplicació: (de vegades s'anomenen simplement *aplicacions*) Programari dissenyat específicament per a satisfer les necessitats dels usuaris. Se'n podrien fer dos grups:

Programari d'ús específic: Programari dissenyat per a un usuari molt concret amb unes necessitats molt concretes: per exemple, el programa que gestiona els préstecs, les quotes i les adquisicions d'un videoclub.

Programari d'ús general: Programari dissenyat per a fer tasques més genèriques, interessants per a moltes classes d'usuaris. Ací en teniu alguns exemples:

Processadors de text per a preparar, modificar, emmagatzemar i imprimir documents de text (vegeu la secció 4.2).

Fulls de càlcul, que permeten automatitzar càlculs que es repeteixen sobre un conjunt més o menys gran de dades (per exemple, per a calcular la nota mitjana de cada estudiant d'una classe sencera a partir de les notes parcials), i presentarne els resultats de diverses maneres, per exemple, en gràfics de molts tipus.

Gestors de bases de dades que serveixen per a emmagatzemar, organitzar i gestionar de diverses maneres la informació continguda en *bases* o bancs de dades (vegeu el capítol 5).

[10] Hi ha dos famílies bàsiques de processadors de llenguatges: els *compiladors*, que tradueixen tot el programa al llenguatge de la màquina abans d'executar-lo, i els *intèrprets*, que lligen el programa línia a línia i executen petits programes ja escrits en el llenguatge de la màquina i que corresponen a les sentències del llenguatge de programació.

Programes de comunicacions, que permeten connectar el nostre ordinador a altres ordinadors, transferir fitxers, etc. Actualment, en molts casos, els programes de comunicació formen part dels sistemes operatius i l'usuari no s'adona que estiguen actius.

Navegadors d'Internet: programes que permeten accedir de manera senzilla als documents d'Internet en màquines connectades a aquesta xarxa[11]. Exemples: *Microsoft Internet Explorer*, *Netscape*, *Mosaic*. Vegeu el capítol 3.

Programes de gràfics, que ens ajuden a presentar de manera més útil les dades de què disposem (aquests programes apareixen moltes voltes associats a fulls de càlcul).

Programes d'autopublicació o autoedició, que integren textos, imatges, etc. fins a produir un document imprés amb característiques de publicació. Quan el disseny de la publicació no és massa complex, és possible usar senzillament un processador de textos.

Els programes d'aplicació els activa l'usuari per mitjà del sistema operatiu, i utilitzen el sistema operatiu per a accedir als recursos (maquinari i altres programes) del sistema.

2.3 Tipus d'ordinadors

Una classificació que pràcticament ha passat de moda és la següent:

"Mainframes": Ordinadors que normalment contenen més d'un processador central i són capaços de donar servei a més d'un usuari alhora mitjançant estratègies de compartició del temps. La noció de "mainframe" està associada a la de "centre de processament de dades" (d'una universitat, d'un ministeri, d'una gran empresa, etc.). N'hi ha alguns que es diuen "supercomputadors" i s'usen en enginyeria i en ciència.

Miniordinadors: Versió reduïda dels "mainframes" (normalment basats en un processador únic) que s'usen molt comunament en ambients de recerca i de manufacturació. També poden ser usats per més d'un usuari i executar més d'una tasca alhora.

Microordinadors: Ara se'ls anomena més comunament *ordinadors personals*, i estan dissenyats per a ser usats per un únic usuari. Els PC i els Macintosh en són exemples.

[11] El nom *navegador* s'usa per l'analogia —dèbil— existent entre els mecanismes d'accés als documents de la Internet i la navegació mitjançant un mapa en una zona desconeguda. Altres noms: *browser* (fullejador), *explorador* i *visor*.

La classificació no pot ser molt rígida. En concret, els PC són cada volta més potents i, amb un sistema operatiu adequat —multitasca i multiusuari— poden fer les funcions que fa uns cinc anys només feien els miniordinadors i en fa uns deu només feien els "mainframes".

2.4 Memòria

Tota la informació —instruccions de programa o dades— que s'emmagatzema en la memòria d'un ordinador s'hi guarda en forma binària; és a dir, cada dada és una cadena de dígits binaris o *bits*. Un bit pot tenir dos valors: 0 (apagat, inactiu) o 1 (encés, actiu); això és perquè el dispositiu electrònic corresponent pot estar en dos estats. Si necessitem guardar objectes o unitats d'informació que tenen més de dos valors possibles, no tindrem prou amb 1 bit; haurem de combinar més d'un bit. Per exemple, si tenim una unitat d'informació que pot presentar-se en 778 formes diferents[12], necessitarem 10 bits, perquè amb 9 bits només podem fer

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^9 = 512$$

combinacions, però amb 10, ja en podem fer suficients, perquè $2^{10} = 1.024$ (en quedarien $1024 - 778 = 246$ combinacions sense usar).

Els bits s'agrupen normalment en grups de vuit, anomenats *octets* o *bytes*. Un octet pot estar, per tant, en $2^8 = 256$ estats diferents. Per exemple, els caràcters i símbols més comunament usats en textos se solen guardar cada un en un octet, usant el codi ASCII (*American Standard Code for Information Interchange*), on el codi de la "`A`" és "`01000001`" o el de la "`z`" és "`01111010`" (vegeu l'epígraf 4.1). El codi ASCII és el codi bàsic per a emmagatzemar textos; quan els textos són més rics i contenen informació sobre tipus i grandàries de lletra, diagramació, notes a peu de pàgina, etc., s'usen formats més avançats que s'expliquen en l'epígraf 4.1. Un octet pot contenir, per tant, molt poca informació (un caràcter, una instrucció senzilla del processador central, un nombre de 0 ("`00000000`") a 255 ("`11111111`"), etc.). Per exemple, un document de text com aquest té desenes de milers de caràcters, i una enciclopèdia, centenars de milions. En les imatges en blanc i negre, cada punt és un bit; una pantalla d'ordinador en conté més o menys un milió. Si són de colors, cal més d'un bit per a cada punt. Les instruccions dels programes, com el de l'exemple, que sumava tots els enters del 1 al 10, també s'emmagatzemen en octets: per exemple, la instrucció `inc A`, que incrementa l'acumulador en 1, podria ser l'octet `11010110`.

Com que un octet pot contenir poca informació, normalment es parla de

- *kilooctets* o *kilobytes* (kb), o milers d'octets. De fet, per fidelitat al sistema binari, un kilooctet no té 1.000, sinó 1.024 octets (2^{10} és 1.024), és a dir 1.024×8=8.192 bits.

[12] Com, per exemple, els signes d'alguna escriptura ideogràfica oriental

- *megaoctets* o *megabytes* (Mb), o milions d'octets. De fet, com en el cas dels kilooctets, no exactament:

$$1 \text{ Mb} = 1.024 \times 1.024 \text{ octets} = 1.048.576 \text{ octets}.$$

- *gigaoctets* o *gigabytes* (Gb), o milers de milions —una mica més— d'octets:

$$1 \text{ Gb} = 1.024 \text{ Mb} = 1.048.576 \text{ kb} = 1.073.741.824 \text{ octets}.$$

Per exemple, actualment és comú que un PC tinga un processador Pentium (vegeu el glossari, apartat 2.7), 32 o 64 Mb de memòria RAM i uns 10 Gb de disc dur. Els disquets d'ús més comú en ordinadors personals, els disquets de densitat alta (marcats "HD") de 3,5 polzades, emmagatzemen 1,44 Mb de dades. Un CD-ROM conté uns 650 Mb.

2.5 Configuració d'un ordinador personal

La configuració clàssica d'un ordinador personal model 2001 sol ser més o menys com segueix:

- La unitat base (la "caixa" o la "torre") conté el processador (típicament un Pentium III o millor), la memòria RAM (per exemple, de 128 MB), una bona targeta gràfica, un disc fix de 30 MB, la unitat de disquet, la unitat lectora de CD-ROM[13] o de DVD (una classe de disc òptic més veloç i amb més capacitat) i la placa de so amb altaveus i micròfon. Pot haver-hi també plaques de comunicacions (mòdems o targetes de xarxa, vegeu el glossari, apartat 2.7).
- Hi ha un monitor o pantalla, de colors, de 15 polzades o més.
- Hi ha un teclat separat i un ratolí de dos o tres botons.
- Pot haver-hi una impressora (d'injecció o de raig de tinta —la més típica—, o làser[14]).

2.6 Sobre algorismes

El nom *algorisme* ve del nom d'un matemàtic que treballava a Bagdad vora l'any 825 de la nostra era, Muhammad ibn Musa al Khwarizmi, anomenat

[13] En les unitats de CD-ROM és important la velocitat màxima de transferència de dades, que es dóna com a múltiple de l'estàndard (la d'un CD de música, de l'ordre d'uns 150 kilooctets per segon): quàdrupla (4×), sèxtupla (6×), etc. Actualment no és estrany que una unitat de CD-ROM siga 40× o més. Les velocitats *mitjanes* solen ser més baixes.

[14] Impressores matricials o d'agulles ja no es fan.

així perquè sembla que era d'una ciutat al sud del mar d'Aral anomenada Khwarizm (ara Kheva). Aquest matemàtic va introduir el sistema decimal hindú en el món àrab. El seu llibre, *Kitab al-jabr wa al-muqabalah*, o *Llibre de la integració i l'equació*, va donar nom també a l'*àlgebra* quan es va traduir al llatí en el segle XII. El llibre és bàsicament una compilació de procediments algebraics i geomètrics, és a dir, d'*algorismes*.

Un exemple d'algorisme/programa és el següent, que diu si dues seqüències de caràcters (també anomenades *cadenes* de caràcters) són iguals o no. Aquest algorisme està relacionat amb la recerca d'un mot en un diccionari, per exemple. Els algorismes s'enuncien *en imperatiu*, com si fossen ordres que es donen a algú perquè les execute seqüencialment, excepte quan es trenca l'ordre amb una instrucció "Vés a".

ALGORISME COMPARA
Entrada: Dues seqüències de caràcters A i B.
Eixida: *Sí*, si són iguals; *no* si no ho són.

1. Fes que p valga 1 (p indica la posició actual dins de les seqüències que estem comparant: $p = 1$, per al primer caràcter, $p = 2$, per al segon, etc.)
2. Si no existeix la posició p de A, vés a 7.
3. Si no existeix la posició p de B, vés a 8.
4. Si el caràcter en la posició p de A no és igual al caràcter en la posició p de B, vés a 8.
5. Suma 1 a p.
6. Vés al pas 2.
7. Si no existeix la posició p de B, digues *sí* i para.
8. Digues *no* i para.

Si es pot suposar que a les cadenes de caràcters s'afegeix un símbol especial de final de cadena (per exemple, '$'), l'algorisme se simplifica:

ALGORISME COMPARA2
Entrada: Dues seqüències de caràcters A i B, acabades en '$'.
Eixida: *Sí*, si són iguals; *no* si no ho són.

1. Fes que p valga 1
2. Si el caràcter en la posició p de A no és igual al caràcter en la posició p de B, digues *no* i para.

3. Si el caràcter en la posició p de A és '$', digues *sí* i para.

4. Vés al pas 2.

Aquest algorisme es pot convertir a un programa d'ordinador. Si existira un llenguatge de programació de l'estil de Pascal en valencià, potser tindria una forma com aquesta:

```
func compara(A, B : sequencia de car);
 var p :  entera ;
 etiquetes :  2 ;
 inici
    p:=1;
 2: si A[p]<>B[p] llavors retorna 'no';
    si A[p]='$' llavors retorna 'si'
    ves_a 2;
 fi;
```

on `A[p]` representa el caràcter que hi ha en la posició `p` de la seqüència `A`.

2.7 Un petit glossari

Aquest glossari arreplega alguns termes d'ús comú en la descripció d'ordinadors i programes que no han estat definits més amunt.

386: Un model de processador central de la família 86, dissenyat per Intel i comercialitzat també per altres companyies, que es troba en ordinadors personals del tipus PC. És capaç de processar 32 bits per cicle. Deriva del processador de 16 bits 286 (80286), que al seu torn deriva del 8086, un dels primers processadors instal·lat en els PC.

486: Un model de processador central de la família 86, dissenyat per Intel i també per altres companyies, que es troba en ordinadors personals del tipus PC. És el successor del 386 —té un conjunt d'instruccions més complet—, i processa, com aquest, 32 bits per cicle; tots els programes escrits per a ser executats pels processadors 386 poden ser executats pels processadors 486.

adaptador de vídeo (també anomenada targeta gràfica o controlador de vídeo): targeta que permet connectar un monitor a l'ordinador. Hi ha molts tipus d'adaptadors de vídeo. Se n'ha de considerar la *resolució*, és a dir, el nombre de punts, elements d'imatge (*píxels*) que caben en una imatge, per exemple 1024×768 (horitzontal $\times$ vertical), i altres paràmetres com la *freqüència de refrescament* (que es mesura en hertzs o cicles per segon; vegeu "megahertz"). Els adaptadors de vídeo

bàsics més comuns són els adaptadors VGA o SuperVGA, encara que actualment n'hi ha de molt més avançats.

cache: Memòria RAM intermèdia, d'accés més ràpid per part del processador, on es copia de quan en quan un bloc (també "pàgina") complet de posicions consecutives de la memòria RAM general per a simplificar accessos repetits a posicions en la mateixa zona. Per exemple, en un ordinador amb 512 kilooctets (524.288 octets) de *cache*, és molt probable que després d'accedir a la posició 2.000.000 és molt probable que el processador vulga accedir a la posició 2.000.003. Si quan s'ha demanat la 2.000.000 es copien en el *cache* les 524.288 posicions que van de la 1.572.864 a la 2.097.151, l'accés a la posició 2.000.003 serà més ràpida.

densitat alta: Quan els disquets de 3 polzades i mitja (uns 90 mm) són de densitat alta solen estar marcats amb les lletres HD (*high density*) i poden emmagatzemar 1,44 Mb de dades. Els anomenats de *densitat doble* (DD) —molt poc usats actualment— poden contenir 720 kb, és a dir, la meitat. Es distingeixen perquè els primers tenen un petit foradet quadrat addicional (a la part inferior dreta si posem la finestra dalt).

Linux: un sistema operatiu multitasca i multiusuari gratuït, de l'estil de l'Unix que es podia trobar en els miniordinadors, desenvolupat de manera no lucrativa per milers de voluntaris arreu del món i que es pot copiar lliurement complint certes condicions. Es pot instal·lar Linux en un PC amb processador 386 o superior i en molts altres tipus d'ordinador.

Macintosh o *Mac*: nom genèric (i comercial) d'una família d'ordinadors construïts per Apple Computer i que són bàsicament equivalents als PC, encara que no compatibles. Algunes diferències entre els PC i els *Mac* són: el processador dels *Mac* no és de la família 86 (386, 486, Pentium) d'Intel, sinó d'una altra (de Motorola); els PC i els accessoris corresponents són fabricats i comercialitzats per nombroses firmes, mentre que els *Mac* i els seus accessoris només són fets i venuts per Apple i per alguna altra firma autoritzada.

megahertz: Un megahertz (MHz) és un milió de hertzs (Hz), és a dir, un milió de cicles per segon. La velocitat de les unitats centrals dels ordinadors es mesura en MHz, és a dir, en milions de cicles bàsics de processament d'informació —corresponents als *tics* o impulsos del rellotge que sincronitza tots els circuits de l'ordinador— per segon. L'execució d'una instrucció per part del processador sol consumir un nombre menut de cicles, quasi sempre més d'un. Una velocitat típica

(any 2001) és 1000 MHz. Una velocitat més gran implica una velocitat d'execució més gran, sempre que no hi haja altres circumstàncies limitants (per exemple, una falta de memòria).

MHz: vegeu megahertz.

mòdem: abreviatura de modulador-desmodulador. Es tracta d'un dispositiu (normalment una placa interna, encara que també pot ser extern) que permet usar la línia telefònica (senyals analògics) per a comunicacions informàtiques (digitals) entre dos ordinadors. Un dels paràmetres més interessants d'un mòdem és la *velocitat* de transmissió de dades, que es mesura en bps (bits per segon). Una velocitat clàssica en mòdems domèstics és 33.600 bps (més recentment, 57.600 bps; les línies telefòniques actuals poden admetre potser velocitats al voltant dels 100.000 bps). Això permet enviar una carta d'una pàgina en unes dècimes de segon. Actualment, un dels usos més populars dels mòdems és la connexió del PC domèstic a la Internet (vegeu l'apartat 3.8).

Pentium: nom genèric de diverses famílies de processadors d'Intel, successors del 486, amb un conjunt d'instruccions més complet i amb moltes millores internes. El Pentium original ha anat evolucionant durant els últims anys: Pentium Pro, Pentium II, Pentium III, Pentium IV etc.

placa de so: Una placa (o targeta) de so permet usar l'ordinador per a processar, enregistrar, reproduir, i manipular sons digitalitzats.

SoundBlaster: una de les marques més comunes de plaques (targetes) especialitzades de so per a PC. Vegeu "placa de so".

targeta de xarxa: per a connectar ordinadors i formar una xarxa (normalment local) per a compartir recursos, cada ordinador ha de tenir una placa o targeta de xarxa. Hi ha diversos estàndards de connexió en xarxa; els més anomenats són Ethernet i Token Ring.

2.8 Qüestions i exercicis

1. Quants *kilobytes* (kilooctets) hi ha en un *gigabyte* (gigaoctet)?

   (a) 1.024

   (b) 1.073.741.824

   (c) 1.048.576

2. Si un ordinador té un disc dur de 6 *gigabytes* (gigaoctets) i una pàgina de text típica té 50 línies de 60 caràcters (contant els blancs), quantes pàgines caben aproximadament en la memòria?

(a) 200

(b) 20000

(c) 2000000

3. Una persona connectada a Internet per telèfon observa que les velocitats de transferència que li indica el seu navegador (vegeu el capítol 3) varien al voltant dels 3 kilooctets (*kilobytes*) per segon. Una d'aquestes tres *no* pot ser la velocitat del seu mòdem:

   (a) 9600 bits per segon

   (b) 57600 bits per segon

   (c) 38400 bits per segon

4. Quina d'aquestes condicions impedeix que un procediment o mètode siga un *algorisme*?

   (a) Que la resposta siga incorrecta.

   (b) Que s'execute indefinidament.

   (c) Que continga instruccions innecesàries però que no destorben el seu funcionament.

5. Quina d'aquestes afirmacions és incorrecta?

   (a) Els mòdems converteixen informació digital en senyals analògics però no al revés.

   (b) Actualment no es pot transferir un megabit per segon per una línia telefònica domèstica convencional.

   (c) Els mòdems serveixen per a connectar-nos a Internet usant la línia telèfonica.

6. Es podria enregistrar (guardar) en un CD-ROM tota la informació continguda en un instant determinat en la memòria RAM d'un ordinador típic de l'any 2001?

   (a) Sí.

   (b) No, perquè no hi cap.

   (c) No, perquè un suport és electrònic i l'altre òptic.

7. Quina d'aquestes afirmacions es certa?

   (a) En qualsevol disc (disquet, dur, CD-ROM) sempre hi ha un directori especial que es diu arrel.

   (b) Un disquet no pot contenir més de dos nivells jeràrquics de carpetes.

(c) Una carpeta no pot contenir només una altra carpeta.

8. Pot haver-hi dos carpetes amb el mateix nom una dins de l'altra?

   (a) No.
   (b) Sí, si tenen data i hora diferents.
   (c) Sí.

9. Quants valors possibles pot prendre un *byte*?

   (a) 2
   (b) 256
   (c) 8

10. Quin dels tres mitjans d'emmagatzemament següents no és magnètic:

    (a) Un CD-ROM
    (b) Un disquet flexible
    (c) Un disc fix

11. Quina capacitat (aproximada) té un disquet de 3,5 polzades de densitat alta?

    (a) 720 Gb
    (b) 1,44 kb
    (c) 1440 kb

12. On resideix un programa d'ordinador mentre l'estem executant?

    (a) En el disc dur o en un disquet.
    (b) En la memòria RAM (almenys parcialment).
    (c) En el CD-ROM.

13. Quina d'aquestes definicions de fitxer és més correcta?

    (a) Un conjunt de dades que es manipula com un tot, resideix en algun mitjà d'emmagatzemament i té un nom.
    (b) Una estructura que conté els noms d'altres fitxers.
    (c) Una estructura de dades que representa el text generat per un processador de textos i que té un nom associat.

14. Quines són les característiques de la memòria RAM?

    (a) És lenta, volàtil i d'accés aleatori.
    (b) És ràpida, volàtil i d'accés aleatori.

(c) És ràpida, permanent i d'accés seqüencial.

15. Es pot fer que diversos ordinadors compartisquen un recurs connectat a un d'ells com, per exemple, una impressora?

    (a) Sí, si estan connectats formant una xarxa.
    (b) Només si la impressora és compatible amb Internet.
    (c) Sí, usant un mòdem en la impressora.

2.9 Solucions

1. (c). Un gigaoctet té 1024 megaoctets, i un megaoctet, 1024 kilooctets: $1024 \times 1024 = 1048576$.
2. (c). Un disc de 6 gigaoctets conté aproximadament 6.000.000.000 octets. Un carácter ocupa un octet; per tant, la pàgina de 50×60 ocupa 3.000 octets. Caben 6.000.000.000/3.000 = 2.000.000 pàgines en un disc.
3. (a). Una velocitat de 3 kilooctets per segon equival a uns $3000 \times 8 = 24000$ bits per segon.
4. (b). Un algorisme sempre s'ha de detenir.
5. (a). Els mòdems modulen (converteixen senyals digitals a analògics) i desmodulen (converteixen senyals analògics en digitals) per a enviar i rebre dades per la línia telefònica. Les línies telefòniques domèstiques actuals admenten uns 100.000 bits per segon.
6. (a). La RAM típica en l'any 2001 és de 128 megaoctets. Un CD-ROM en pot emmagatzemar 650.
7. (a)
8. (c)
9. (b). $2^8 = 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 256$.
10. (a). Els CD-ROM usen un sistema òptic.
11. (c)
12. (b). Almenys la porció del programa que s'està executant ha de residir en la RAM.
13. (a)
14. (b)
15. (a)

Capítol 3

Algunes nocions bàsiques sobre Internet

Una de les eines informàtiques bàsiques que es troben a l'abast del professional de la traducció és Internet. Per exemple, Internet permet que els professionals puguen usar recursos remots, o que puguen rebre originals dels seus clients i remetre'ls-en les traduccions.

3.1 Què és Internet?

S'anomena *Internet* un conjunt d'ordinadors, distribuïts arreu del món i interconnectats mitjançant un protocol estàndard (el protocol d'Internet o IP) de manera que els recursos presents en uns ordinadors (normalment, informació) estan disponibles per a ser usats pels usuaris d'altres ordinadors. Es diu que els ordinadors d'Internet formen una *xarxa*, en la qual els nodes o nusos són els ordinadors i els fils, les connexions; les connexions poden ser de naturalesa molt diversa (línies telefòniques, fibra òptica, enllaços de ràdio terrestres o per satèl·lit, etc.), però el protocol d'Internet està dissenyat de manera que la naturalesa de la connexió no siga rellevant per a l'usuari. Altres noms que s'usen (més recentment) en comptes d'*Internet* són *World Wide Web* o *WWW* ("teranyina d'abast mundial") o simplement *Web* ("teranyina").

3.2 Números IP

Cada node (cada ordinador) de la xarxa Internet té un *número IP* únic, el qual es compon de 4 octets (4 enters del 0 al 255) separats per punts, com ara `192.168.5.5`. Els enters inicials s'usen per a designar grans subxarxes, mentre que els finals s'usen per a designar xarxes més menudes, i dins d'aquestes, ordinadors concrets (en això recorden els números de telèfon: dos abonats pròxims normalment comparteixen les xifres inicials).

INDICATIU	PAÍS
`.es`	Espanya
`.fr`	França
`.pt`	Portugal
`.it`	Itàlia
`.uk`	Regne Unit
`.ru`	Rússia
`.za`	Sud-àfrica
`.ie`	Irlanda

Taula 3.1: Indicatius Internet d'alguns països

INDICATIU	TIPUS
`.gov`	governamental
`.mil`	militar
`.com`	comercial
`.org`	organització no lucrativa
`.edu`	institució educativa

Taula 3.2: Alguns indicatius Internet usats originalment als Estats Units d'Amèrica i més recentment arreu del món (recentment s'estan llançant nous indicatius com ara `.biz` (*business*), `.info`, etc.)

3.3 Noms

Com que recordar números IP no és fàcil, normalment s'usen *noms* o *adreces* per a referir-se a les màquines; alguns dels ordinadors de la xarxa (anomenats *servidors de noms*) s'encarreguen de traduir els noms a números IP. Per exemple, un nom podria ser `altea.dlsi.ua.es`, on `altea` es refereix a una màquina concreta del Departament de Llenguatges i Sistemes Informàtics (`dlsi`) de la Universitat d'Alacant (`ua`), que es troba a Espanya (`es`) (en això els noms s'assemblen a les adreces postals: primer es dóna el més concret i al final el país; aquest ordre és l'invers al dels números IP).

La taula 3.1 dóna alguns exemples d'indicatius de països. De vegades, l'últim component d'un nom no es correspon amb l'indicatiu d'un país, sinó que indica la naturalesa del lloc (es tracta d'una relíquia del sistema de noms anterior); fins fa poc calia sobreentendre que es tractava d'un ordinador situat físicament als Estats Units d'Amèrica, però ara això ja no és necessàriament així. Aquests indicatius apareixen en la taula 3.2. En altres països (`.uk`, `.nz`, `.za`) s'usen indicatius similars (`.co`(mercial), `.ac`(adèmic), etc.) davant de l'indicatiu de país (per exemple, `www.shef.ac.uk` és la Universitat de Sheffield).

3.4 Adreces de correu electrònic

Un dels serveis més usats d'Internet és el correu electrònic (en anglés *electronic mail* o *e-mail*), que ens permet enviar missatges (textos informatitzats que poden contenir, a més, fitxers adjunts o *attachments*) a usuaris d'altres ordinadors. Les adreces de correu electrònic tenen dues parts, separades pel caràcter "`@`", que se sol pronunciar *at* (en anglés, per part dels informàtics més vells), *arrova*, *rova* o, per què no, *ensaïmada*. La primera part és l'identificador de l'usuari i la segona, el nom d'un ordinador (o d'un grup d'ordinadors que comparteixen un nom). Per exemple, una adreça electrònica vàlida podria ser

`Monica.Lewinsky@whitehouse.gov`

De vegades, una adreça electrònica identifica una llista d'usuaris (*àlies*) o una llista de distribució (la qual envia una còpia de cada missatge que rep a tots els inscrits en la llista). En els dos casos, si hi enviem un missatge, el reben tots els inscrits, de manera que es pot usar per a establir, per exemple, fòrums de discussió.[1] Els missatges poden contenir, a més del text mateix del missatge, fitxers *annexos* (en anglés *attachments*) com ara imatges o documents de text.

3.5 News

Aquest és un servei similar a les llistes de distribució i, per tant, també serveix per a constituir grups de discussió o de notícies (*newsgroups*) sobre un tema; els missatges que s'hi envien hi queden guardats i qui vulga saber si hi ha novetats en un grup de discussió determinat només s'ha de connectar al *servidor de notícies* més pròxim, seleccionar el grup i llegir els missatges nous. Els grups de notícies tenen noms compostos per diversos mots separats per punts, com ara `comp.ai.neural-nets`.

3.6 Els localitzadors

Els serveis i els documents concrets presents en un ordinador que els fa disponibles (un servidor d'Internet) es poden designar mitjançant el seu *localitzador uniforme de recursos* o, més comunament, *URL* (de l'anglés *uniform resource locator*). L'URL és, per tant, una *adreça* que identifica o localitza uniformement un servei o document (un recurs) de qualsevol dels que s'ofereixen en Internet.

[1] Per exemple, la llista de distribució sobre traducció automàtica MT-List, mantiguda per l'EAMT (*European Association for Machine Translation*, associació europea per a la traducció automàtica), té l'adreça `mt-list@eamt.org`; per a formar part de la llista cal subscriure's en l'URL `http://www.pairlist.net/mailman/listinfo/mt-list`. Si s'envia un missatge a `mt-list@eamt.org` el reben tots els subscrits.

Un URL té generalment tres parts, encara que s'hi donen algunes variacions:

- el protocol, que indica la classe de document i com l'ha d'usar l'ordinador sol·licitant (o *client*)
- el nom de l'ordinador servidor o el seu número IP
- (opcionalment) informació sobre la localització del servei o document dins de l'ordinador servidor

Per exemple, l'URL

`http://www.mtn.co.za/regulars/sms/index.html`

es refereix a un document d'hipertext[2] compatible amb el protocol `http` (*hypertext transfer protocol* o protocol de transferència d'hipertextos) situat en l'ordinador `www.mtn.co.za` (de l'empresa comercial Mobile Telephone Networks de Sud-àfrica), i, dins d'aquesta, en el directori `regulars`, subdirectori `sms`. El fitxer que conté l'hipertext s'anomena `index.html` (on les sigles HTML corresponen a *hypertext markup language*, nom del llenguatge o sistema de marques més usat per a donar format als hipertextos, descrit en l'epígraf 4.1).

En canvi, l'URL `mailto:anton@dlsi.ua.es` serveix per a enviar correu electrònic (`mailto`) a l'usuari que té l'adreça de correu electrònic `anton@dlsi.ua.es`.

Altres protocols són: `ftp://`, *file transfer protocol*, usat per a descarregar (transferir) fitxers per a guardar-los en el nostre ordinador; i `telnet:`, per a connectar-nos a l'ordinador remot i usar-lo com si ens trobàrem allà mateix, o `news:` per a identificar grups de notícies.

3.7 Navegadors

Els programes navegadors es coneixen també per altres noms: *browsers* (fullejadors), *exploradors* i *visors*. Són programes que permeten accedir de manera senzilla als documents o serveis d'Internet en ordinadors connectats a aquesta xarxa; entre altres coses, els navegadors interpreten els hipertextos escrits en HTML i els presenten a la persona usuària en el format que indiquen les marques, de manera que els enllaços a altres hipertextos queden clarament destacats i siguen *actius*, és a dir, que responguen a un clic del ratolí *saltant* a l'hipertext enllaçat. Els navegadors més usats són *Microsoft Internet Explorer* (empaquetat, no sense polèmiques legals, amb el sistema operatiu Windows 98) i *Netscape*. Els navegadors normalment porten un programa de correu electrònic incorporat.

[2] Un hipertext és un document de text que conté enllaços que permeten accedir directament a altres hipertextos relacionats.

3.8 Accés domèstic a Internet

La majoria dels usuaris particulars d'Internet s'hi connecten mitjançant una línia telefònica com les que s'usen normalment per a parlar (potser la mateixa que usen per a parlar). Per a poder usar Internet des de casa, cal:

- un mòdem
- un programa per a l'accés telefònic a xarxes (normalment present en el sistema operatiu)
- haver-se donat d'alta amb algun proveïdor d'Internet (ISP, *Internet service provider*)

El proveïdor d'Internet sol estar accessible mitjançant un número de telèfon local tot usant un nom d'usuari i una contrasenya. Fins fa poc, els proveïdors d'Internet solien oferir normalment una tarifa fixa (una certa quantitat al mes, independentment del temps de connexió), però el cost de la telefonada local necessària es pagava sempre per minuts. En l'actualitat, d'una banda, la major part dels ISP ofereixen el servei d'internet gratuïtament de manera que només s'han de pagar les telefonades, però, d'altra banda, molts proveïdors d'Internet estan associats a companyies telefòniques i ofereixen descomptes[3] si l'accés es fa a través d'una companyia concreta.

Durant el temps de la connexió, el proveïdor d'Internet assigna un número IP temporal al nostre ordinador domèstic, de manera que forme part d'Internet i puga accedir com a client a tots els serveis i documents disponibles en qualsevol màquina de la xarxa, però normalment no com a servidor.

3.9 Exercicis i qüestions

1. La primera part d'un URL (localitzador uniforme de recursos) especifica
   (a) el protocol d'accés.
   (b) el nom del servidor.
   (c) el directori on es troba el servei.
2. Després del protocol d'accés, un URL (localitzador uniforme de recursos) especifica
   (a) la velocitat de transferència.

[3] Hi ha diverses modalitats de descompte: preus per minut reduïts, bons de preu fix per un nombre d'hores al mes, quotes mensuals fixes per a totes les connexions a certes hores, o fins i tot les anomenades *tarifes planes*: quotes mensuals fixes que permeten connectar-s'hi a qualsevol hora sense limitacions de temps.

   (b) el nom del servidor.
   (c) el directori on es troba el servei.

3. Què es "http://www.tharaka.org.ke/nkoru"?
   (a) Un URL.
   (b) Una adreça de correu electrònic.
   (c) El nom d'un fitxer local del nostre ordinador.

4. Els números IP es componen de 4 números del 0 al 255 separats per punts. Quants bits són necessaris per a emmagatzemar-los?
   (a) 16
   (b) 32
   (c) 4

3.10 Solucions

1. (a)
2. (b)
3. (a)
4. (b). Cada número del 0 al 255 es pot emmagatzemar en 8 bits ($2^8 = 256$) i n'hi ha quatre: $8 \times 4 = 32$.

Capítol 4

L'entrada i el processament de textos

El tipus de fitxer bàsic de treball del professional de la traducció sol ser un fitxer amb text, és a dir, un text informatitzat. Aquest fitxer pot contenir, a més del text mateix, informació sobre el format dels paràgrafs i de les pàgines, sobre els tipus i les grandàries de lletra que s'usen amb cada mot, etc. Un text informatitzat pot tenir orígens diversos:

- Pot haver estat generat per un altre programa d'ordinador, per exemple a partir de les dades contingudes en alguna base de dades (vegeu la pàg. 11).
- El podem haver rebut annex a un missatge electrònic (vegeu la pàg. pg:annex).
- El podem haver descarregat (copiat) d'algun servidor d'Internet (vegeu la pàg. pg:ftp).
- El podem haver generat, potser a partir d'un altre text, usant un *processador de textos* (vegeu l'apartat 4.2).
- El pot haver generat un *sistema de reconeixement de la parla* (vegeu la secció 4.3) a partir de la veu de la persona que l'ha dictat.
- El pot haver generat un *sistema de reconeixement de textos escrits* a partir d'un text tipografiat o manuscrit (vegeu la secció 4.4).

4.1 Formats de text

4.1.1 L'ASCII original de 7 bits

Com ja s'ha comentat en la pàgina pg:ASCII, per a emmagatzemar textos s'usa el codi ASCII com a codi bàsic. Aquest codi assigna un número de 7 bits a cada caràcter, de manera que permet emmagatzemar un caràcter per

octet i encara en sobra un bit. La taula 4.1 mostra alguns exemples de codis ASCII. Els codis ASCII del 0 al 31 no corresponen a caràcters imprimibles sinó a *caràcters de control* que tenen noms especials i s'usen per a un control rudimentari del format i de la transmissió dels textos.

4.1.2 Els ASCII estesos a 8 bits

ASCII no té codis per als caràcters especials que usen algunes llengües europees com ç, ä, ú, etc.; amb l'arribada dels microordinadors es va decidir ampliar el codi ASCII, de 7 bits i per tant amb $2^7 = 128$ possibilitats, a un codi de 8 bits amb $2^8 = 256$ possibilitats; el vuité bit o "bit 7"[1]—el primer per l'esquerra— és 1 per als caràcters especials (numerats del 128 al 255) i zero per als caràcters estàndards d'ASCII. El fet que hi haja més d'una manera estàndard d'usar els nous codis fa que de vegades els textos amb caràcters especials no queden bé quan passem d'un processador de textos (o un editor de textos) a un altre. En la nostra àrea geogràfica s'usa normalment la codificació ANSI Latin-1, també coneguda com a ISO-8859-1; aquesta codificació serveix per a les llengües següents: *afrikaans*, alemany, anglés, basc, català, danés, escocés, espanyol, feroés, finés, francés, gallec, irlandés, islandés, italià, neerlandés, noruec, portugués i suec.[2]

4.1.3 Unicode

Els codis de 8 bits com ANSI són adequats per a la major part de les llengües europees, les quals es basen en l'alfabet llatí amb algunes modificacions, però hi ha llengües al món que tenen sistemes d'escriptura molt complexos amb milers de símbols diferents, com ara el xinés o el japonés. Dues-centes cinquanta-sis combinacions no són suficients per a aquestes llengües i s'hi han proposat diverses solucions. *Unicode* (ISO 10646) és un nou estàndard per a fitxers de text que permet codificar pràcticament totes les llengües del món i fins i tot mesclar diversos alfabets en un mateix fitxer.[3] La versió més comunament usada d'Unicode (BMP, *Basic Multilingual Plane*) té 65534 caràcters; això suposaria l'ús de 2 octets (16 bits) en comptes d'un ($2^{16} = 65536$); això faria que un text Unicode senzill fóra el doble de gran que el text ASCII corresponent, però hi ha mètodes de codificació d'Unicode, com l'UTF-8, que en el cas de les llengües europees amb alfabet llatí estalvia espai perquè usa un únic octet per als codis ASCII (del 0 al 127, els més freqüents), i més d'un octet per als codis següents (així, a més, és compatible amb l'ASCII).

[1] Recordeu que en informàtica és comú comptar començant pel zero.

[2] Recentment s'ha fet una modificació anomenada ISO-8859-15, per a incloure, entre altres, el símbol de l'euro i resoldre alguns problemes referents al francés i al finés.

[3] Unicode té 31 bits; és a dir, permet $2^{31} = 2.147.483.648$ caràcters diferents

Codi binari	codi decimal	caràcter
0000000	0	NUL (caràcter nul)
...	...	...
0001001	9	TAB (tabulador)
0001010	10	NL (nova línia)
...	...	...
0001101	13	CR (retorn del carro)
...	...	...
0100000	32	(un espai en blanc)
0100001	33	!
0100010	34	"
...	...	...
0110000	48	0
0110001	49	1
...	...	...
0111000	56	8
0111001	57	9
...	...	...
1000000	64	@
1000001	65	A
1000010	66	B
...	...	...
1011010	90	Z
1011011	91	[
...	...	...
1100000	96	‘
1100001	97	a
1100010	98	b
...	...	...
1111010	122	z
1111011	123	{
...	...	...
1111110	126	~

Taula 4.1: Alguns exemples del codi ASCII. Els codis del 0 al 31 no corresponen a caràcters imprimibles sinó a caràcters de control.

4.1.4 Limitacions

Tot i que ampliem l'ASCII a ANSI o Unicode, encara és molt limitat. Per exemple, si volem que un text tinga un cert format, només podrem usar caràcters de control con l'espai en blanc, el tabulador, el salt de línia, etc. Per exemple, no podrem canviar fàcilment de tipus o de grandària de lletra. De qualsevol manera, ASCII i ANSI s'usen en aplicacions com ara el correu electrònic, o quan volem que un text —el contingut del qual és molt més important que l'aparença— puga ser llegit per qualsevol usuari sense importar el processador de textos que use; els textos d'aquesta mena s'anomenen de vegades *textos plans* i s'emmagatzemen en fitxers amb l'extensió `.txt`. Aquests textos es poden produir i llegir amb qualsevol *editor de textos* (vegeu l'epígraf 4.2).

4.1.5 Formats més avançats

Però els documents de text són en general més rics que simples seqüències de caràcters: contenen informació sobre tipus i grandàries de lletra, formats, notes a peu de pàgina, etc. Per a guardar aquesta informació, s'usen:

- D'una banda, codificacions o formats basats en l'ASCII com ara SGML (*standardized generalized markup language*), la seua versió simplificada XML (*extensible markup language*), el tipus SGML anomenat HTML (*hypertext markup language*), i el format proposat per Microsoft anomenat RTF (*rich text format*, no relacionat amb SGML). Tots aquests formats usen combinacions especials de caràcters ASCII per a indicar aquestes característiques de presentació[4].

- D'altra banda, hi ha els formats particulars dels processadors de text com ara WordPerfect o Microsoft Word, que usen esquemes diferents, basats en codis binaris no relacionats amb l'ASCII.

Però l'ús de formats de text més avançats no només serveix per a determinarne la presentació en la pantalla o quan són impresos; com veurem més avall, en el cas de SGML i XML, el format serveix per a *estructurar* el document de text en unitats directament relacionades amb el contingut del document, com ara seccions, títols de secció, llistes, paràgrafs, etc.; aquesta estructuració interna del document pot ser usada després per a fer recerques d'informació amb l'ajuda de l'estructura definida, com ara buscar un mot concret només en títols de secció, o també per a produir-ne una presentació concreta del document. De fet, recentment, amb l'aparició de XML (vegeu més avall), s'observa una tendència cap a l'adopció de formats de document estructurats, és a dir, no relacionats únicament amb la presentació, sinó també amb

[4] Aquests caràcters són normalment invisibles per a la persona usuària mentre redacta el document, excepte si demana explícitament que els vol veure.

l'estructura pròpia del document, formats normalment concebuts de manera que la presentació desitjada es puga produir a partir de l'estructura usant fitxers (anomenats *fulls d'estil*) amb regles d'estil ben definides.

4.1.6 SGML i XML

SGML

SGML, el llenguatge estàndar generalitzat de marques, havia tingut un èxit relatiu fins a mitjans dels noranta; però l'aparició d'una versió restringida i simplificada de SGML anomenada XML ha impulsat enormement l'adopció dels formats d'estructuració de documents, de tal manera que en l'actualitat s'usa XML moltíssim més que el SGML original;[5] per això, ens centrarem en aquest últim format.

XML

Un document XML és un document ASCII on, a més de text, podem trobar *etiquetes* o *marques* (en anglés *tags*) que donen informació sobre la naturalesa i l'organització de cada un dels continguts del document; com ja s'ha dit, un document XML és un document *estructurat*. Per exemple, un document XML corresponent a un FAX podria tenir l'aparença que es mostra en la figura 4.1. La primera línia declara que el document és un document XML de la versió 1.0 i que el joc de caràcters que usa és l'ISO-8859-1 (o 'Latin 1'). Com s'hi pot veure, les etiquetes que apareixen entre parèntesis angulars indiquen les diverses parts del document. Típicament, s'obrin amb `<`*nom*`>` i es tanquen amb `</`*nom*`>`. En l'exemple, es pot veure que un fax (`<FAX>`...`</FAX>`) té un destinatari, un remitent, una data i un text. Tant el destinatari com el remitent tenen nom i número, i el text es compon de paràgrafs (`<P>`...`</P>`).

DTD

L'estructura de les etiquetes d'un tipus determinat de document XML es pot especificar usant una DTD (*document type definition* o definició del tipus de document). La segona línia del fax de la figura 4.1 especifica el tipus del document tot indicant d'una banda l'etiqueta arrel o principal del document (`FAX`) i l'URL (`SYSTEM`) on es troba la DTD. Aquesta DTD es veu en la figura 4.2; examinem la DTD línia a línia per a comprendre com s'usen les DTD per a definir famílies (tipus) de documents en XML:

1. La primera línia declara que la DTD és una DTD de la versió 1.0 i que el joc de caràcters que s'usa és l'ISO-8859-1, també conegut com a Latin-1.

[5] Per exemple, XML és especialment popular com a eina per a definir l'estructura dels documents presents en les biblioteques digitals.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
```

2. La segona línia és un comentari. Els comentaris comencen amb `<!--` i acaben amb `-->` i es poden situar en qualsevol part d'una DTD.

   ```
   <!-- Aquest és l'exemple de DTD de FAX -->
   ```

3. Les línies següents defineixen l'estructura del document definint els seus *elements*. La línia

   ```
   <!ELEMENT FAX (DESTINATARI, REMITENT?, DATA, TEXT)>
   ```

   defineix l'element arrel o principal, `FAX`, i especifica que es compon (en l'ordre especificat) d'un `DESTINATARI`, un `REMITENT` opcional (indicat amb `?`), d'una `DATA` i d'un `TEXT`.

4. El `DESTINATARI` del fax té dues parts: el `NOM` (opcional) i el número (`NUM`).

   ```
   <!ELEMENT DESTINATARI (NOM?, NUM)>
   ```

5. El remitent es defineix igual:

   ```
   <!ELEMENT REMITENT (NOM?, NUM)>
   ```

6. El `NOM`, el número `NUM` i la `DATA` contenen text sense marques (indicat amb `#PCDATA`).

   ```
   <!ELEMENT NOM (#PCDATA)>
   <!ELEMENT NUM (#PCDATA)>
   <!ELEMENT DATA (#PCDATA)>
   ```

7. El `TEXT` es compon d'un o més (`+`) paràgrafs (`P`).

   ```
   <!ELEMENT TEXT (P)+>
   ```

8. Finalment, els paràgrafs contenen text.

   ```
   <!ELEMENT P (#PCDATA)>
   ```

Una de les aplicacions més importants de les DTD és que serveixen per a la validació automàtica dels documents: un programa *validador* llegeix la DTD i el document XML i decideix si és vàlid, és a dir, si segueix l'especificació donada en la DTD, però les DTD poden servir també per a construir programes que processen o transformen els documents XML, com ara quan es vol produir una presentació visual del document (o de determinades estructures del document). De fet, el *significat* de les etiquetes (és a dir, quines

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE FAX SYSTEM "http://www.dlsi.ua.es/%7Emlf/iat/fax.dtd">
<FAX>
<DESTINATARI>
       <NOM>Mikel L. Forcada</NOM>
       <NUM>+34-96-590-9326</NUM>
</DESTINATARI>
<REMITENT>
       <NOM>Letícia Ibaizábal</NOM>
       <NUM>+34-96-999-9999</NUM>
</REMITENT>
<DATA>23 de setembre de 2000</DATA>
<TEXT>
<P>Mikel, m'agradaria que m'enviares el text de l'última
pràctica d'IAT.</P>
<P>Ah, i també; l'examen de febrer. Gràcies.</P>
</TEXT>
</FAX>
```

Figura 4.1: Un FAX en XML

conseqüències tindran quan es processe el document XML) l'ha d'establir el programa o els programes que processaran els documents. Com ja s'ha dit, el significat de les etiquetes pot estar associat, per exemple, a la manera de presentar el document quan s'imprimeix (per exemple, el destinatari del fax pot anar en negretes), però també podria servir per a facilitar el processament de la informació (per exemple, buscar tots els faxos que tenen un determinat destinatari, o, en llibres codificats en XML, decidir quines parts han de ser traduïdes automàticament del castellà a l'anglés i quines no perquè són cites literàries).[6] Fins i tot fitxers que normalment no consideraríem documents, com ara les memòries de traducció (vegeu el capítol 10) s'estructuren de manera estàndard usant un format basat en XML anomenat TMX.

4.1.7 HTML

El format HTML (*hypertext markup language* o *llenguatge de marques per a hipertextos*) és un dels tipus de document que es poden definir amb

[6] De fet, existeix una DTD SGML (i la DTD XML simplificada corresponent) per a codificar documents de text de l'estil dels llibres, els articles, etc., que s'anomena TEI, de l'anglés *text encoding initiative*, 'iniciativa de codificació de textos".

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- Aquest és l'exemple de DTD de FAX -->
<!ELEMENT FAX (DESTINATARI, REMITENT?, DATA, TEXT)>
<!ELEMENT DESTINATARI (NOM?, NUM)>
<!ELEMENT REMITENT (NOM?, NUM)>
<!ELEMENT NOM (#PCDATA)>
<!ELEMENT NUM (#PCDATA)>
<!ELEMENT DATA (#PCDATA)>
<!ELEMENT TEXT (P)+>
<!ELEMENT P (#PCDATA)>
```

Figura 4.2: La DTD que defineix els faxos com el de la fig. 4.1.

SGML (però no exactament amb XML; la versió XML de HTML s'anomena XHTML); HTML és molt important perquè és el llenguatge en el que estan escrits els hipertextos d'Internet (vegeu el capítol 3) i el que interpreten els navegadors (vegeu l'apartat 3.7). En HTML, les marques tenen un significat determinat. Originalment, les marques estaven pensades per a expressar l'estructura del document, però amb el pas del temps el significat de les marques ha canviat i actualment sol estar més aviat associat a la presentació del document durant la navegació (tot i que, darrerament, sembla que a poc a poc es va tornant al plantejament original amb iniciatives com ara XHTML, una versió compatible amb XML del llenguatge HTML). Per exemple, HTML indica el començament d'un segment de text en negretes amb la marca "`<B>`" (3 caràcters ASCII) i el final amb la marca "`</B>`" (4 caràcters). Els enllaços (hiperreferències) a altres documents comencen amb "`<A HREF="` *URL* `">`" —on *URL* és el localitzador del document enllaçat— i acaben amb "`</A>`", etc. Els documents HTML comencen idealment amb la marca "`<HTML>`" i acaben amb la marca "`</HTML>`", i tenen, entre altres elements, un títol ("`<TITLE>...</TITLE>`") i un cos ("`<BODY>...</BODY>`"). Per exemple, el document HTML que es mostra en la figura 4.3 es mostraria en un navegador aproximadament com en la figura 4.4.

Quan estem mirant un document HTML amb un navegador, podem veure les etiquetes HTML que el formaten si seleccionem l'opció "veure font HTML" ("view HTML source") o similar que hi ha normalment en el menú "veure" ("view").

4.1.8 RTF

RTF (*rich text format*, és a dir, *format de text ric*) és un format impulsat per l'empresa Microsoft per a facilitar l'intercanvi de documents entre pro-

```
<HTML>
<HEAD>
<TITLE>T&iacute;tol del document</TITLE>
</HEAD>
<BODY>
<H1>Encap&ccedil;alament de nivell 1</H1>
<H2>Encap&ccedil;alament de nivell 2</H2>

<p>Aquest &eacute;s el <I>primer</I> par&agrave;graf
d'aquest document. El navegador decideix com dividir-lo
en l&iacute;nies per a presentar-lo. Idealment, hauria
d'acabar amb una marca de final de par&agrave;graf. </p>

<H2>Un altre encap&ccedil;alament de nivell 2</H2>

<p>Aquest &eacute;s l'<I>&uacute;ltim</I> par&agrave;graf
d'aquest document HTML. Els documents HTML poden contenir
<A HREF="http://www.internostrum.com">enlla&ccedil;os</A>
a altres documents HTML, locals o remots. </p>
</BODY>
</HTML>
```

Figura 4.3: Un document HTML, tal com el presentaria un editor normal de textos o usant l'opció "view HTML source" (veure font HTML) del navegador. En aquest document HTML no s'ha especificat el joc de caràcters que s'usarà i per això els caràcters especials com ara les lletres accentuades apareixen representades per entitats que comencen amb "`&`" i acaben amb "`;`".

Títol del document

Encapçalament de nivell 1

Encapçalament de nivell 2

Aquest és el *primer* paràgraf d'aquest document. El navegador decideix com dividir-lo en línies per a presentar-lo. Idealment, hauria d'acabar amb una marca de final de paràgraf.

Un altre encapçalament de nivell 2

Aquest és l'*últim* paràgraf d'aquest document HTML. Els documents HTML poden contenir enllaços a altres documents HTML, locals o remots.

Figura 4.4: El document HTML de la figura 4.3, vist a través d'un navegador determinat.

```
{\rtf1\ansi\ansicpg1252
                                    ...
\par
{\b T\‘edtol en negretes}\par
Text del par\‘a0graf en lletra normal amb alguns incisos
{\i en cursives} i una marca de final de par\‘a0graf al
final.\par
Els car\‘a0cters que no pertanyen a l’ASCII est\‘a0ndard
s’indiquen amb codis especials (en aquest cas s’ha usat
ANSI, amb {\i codepage} 1252, com es veu al principi del
document), com per exemple en el mot
{\i ling\’fc\’edstica}.\par
                                    ...
```

Figura 4.5: Part d'un document de text en format RTF

cessadors de textos mantenint-ne el format. RTF també té etiquetes, les quals comencen normalment per una barra invertida (\); però els àmbits d'acció de les etiquetes estan delimitats per claus ("`{...}`") en comptes de per parelles d'etiquetes; per exemple, un segment en negretes s'indica amb "`{\b...}`", mentre que en HTML s'usa "`<B>...</B>`". La figura 4.5 mostra part d'un document RTF, en la qual es veuen algunes comandes de l'encapçalament (començant amb "`{\rtf1`...") i on també s'observa la manera especial com es codifiquen els caràcters internacionals.

4.2 Processadors de textos

Un *processador de textos* és un programa que permet crear i modificar documents de text informatitzats. També s'hi poden usar *editors*: la diferència entre un processador de textos i un *editor* és que aquest últim programa és un processador de textos ASCII o ANSI senzills (sense informació de format, etc.) que normalment s'usa per a preparar textos en algun llenguatge artificial (per exemple, programes escrits en algun llenguatge de programació) que serviran de entrada per a un altre programa, o textos molt senzills on el format no és crucial, com un missatge electrònic senzill.

El processament de textos també s'anomena *tractament de textos* (paral·lelament al francés, *traitement textes*). En anglés, l'èmfasi és sobre les paraules: *word processing*.

Per descomptat, aquesta secció no pretén instruir en l'ús de cap pro-

cessador de textos concret, sinó que vol descriure breument algunes característiques comunes als processadors de textos que s'usen en l'actualitat. De fet, l'ús dels processadors de text s'aprén molt millor en el laboratori; a més, en vista del fet que els processadors de textos canvien constantment, potser és millor no aprendre a usar un processador concret sinó a buscar en cada processador les eines que necessitem. Això és possible perquè la major part dels processadors van fornits de manuals o de sistemes d'ajuda en línia; alguns tenen fins i tot "assistents" que observen el que fa la persona usuària i li suggereixen —amb més o menys fortuna— possibles accions en cada moment.

Quant a l'*aparença* del programa, la major part dels processadors de text es manifesten bàsicament com una o diverses finestres, cada una de les quals mostra una secció d'algun dels documents de text informatitzats que estem creant i modificant (els documents que tenim *oberts*). La tendència actual afavoreix que el text es mostre tan paregut com siga possible a la versió impresa que se'n produirà, quant a format, tipus de lletra, etc. (en anglés, aquest concepte de fidelitat visual es resumeix amb el mot *wysiwyg*, fet amb les sigles de "what you see is what you get", és a dir, "el que veieu és el que obtindreu").

Quant a l'*operació*, els processadors de text assumeixen que la major part dels caràcters que teclegem s'han d'inserir darrere del caràcter que actualment es troba destacat amb una marca anomenada *cursor* de text (pot ser diferent del cursor o apuntador que indica la posició virtual del ratolí en la pantalla), o bé l'han de sobrescriure. No obstant això, es reserven determinades tecles (algunes senzilles, i altres en combinació amb les tecles especials "Alt" o "Control") per a fer operacions, algunes molt bàsiques com ara moure el cursor de text o esborrar caràcters i altres més complexes, com ara apegar-hi un bloc de text que havíem esborrat prèviament o enregistrar el text complet en el disc[7]. Però moltes d'aquestes operacions, conjuntament amb d'altres que no s'usen tan sovint, també estan accessibles mitjançant *menús*; els noms d'aquests menús solen estar situats típicament en la part de dalt de la finestra: si s'hi fa un clic del ratolí, es despleguen i ens mostren les opcions que contenen, que podem elegir amb el ratolí.

A més de l'accés als menús, hi ha operacions que normalment es fan amb el ratolí: una de les més importants és *marcar* una porció de text per a alguna operació posterior (per exemple, copiar-la o modificar-ne el tipus de lletra); típicament, es fa prement el botó principal del ratolí en un extrem del text que volem marcar i, anant, sense soltar-lo, a l'altre extrem.

Heus ací algunes de les operacions bàsiques que es poden fer amb un processador de textos, organitzades de manera similar als menús que trobarem en un processador de text:

[7] Aquestes tecles i combinacions de tecles que permeten un accés ràpid a operacions rutinàries se solen anomenar en anglés *hotkeys*.

- Operacions amb fitxers:
  - *crear* un nou document de text;
  - *obrir* un fitxer de document existent en el disc per a treballar-hi;
  - *guardar* el document en curs en un fitxer amb el mateix nom i en el mateix format que tenia quan el vam obrir;
  - guardar el document en curs amb un altre nom o en un altre format, per exemple el d'un altre processador de textos (*guardar com*);
  - *imprimir* el document actual;
  - fer una *presentació preliminar* en pantalla del document tal com quedarà imprés (quan el processador no és completament *wysiwyg*)
  - *eixir* del processador de textos.
- Operacions d'edició. A més de les operacions que només es fan des del teclat o amb el ratolí, com ara inserir un caràcter, esborrar-lo, moure'ns pel text, o marcar-hi un passatge, hi ha operacions de modificació del text molt importants que estan accessibles en el menú d'*edició*:
  - esborrar (*retallar*) la part marcada;
  - *copiar* la part marcada a un *portapapers* (una memòria intermèdia) per a usar-la posteriorment;
  - inserir (*enganxar* o *apegar*) el contingut del portapapers en un punt del document actual;
  - *desfer* o invertir l'última operació (hi ha processadors de text que recorden un nombre considerable d'operacions bàsiques i permeten desfer-ne més d'una, en ordre invers, per descomptat; altres només en recorden l'última operació).
- Operacions de cerca i substitució:
  - *buscar* una determinada paraula o seqüència de caràcters en el text;[8]
  - repetir l'última recerca;
  - buscar una determinada paraula o seqüència de caràcters en el text i *substituir-la* per una altra (interactivament o automàticament).

[8] Alguns programes permeten buscar usant les anomenades *expressions regulars*, clàssiques en el sistema operatiu Unix, les quals permeten, mitjançant caràcters especials anomenats *jòquers* (anglés *wildcards*), buscar tots els mots que segueixen un patró determinat. Per exemple, una recerca amb l'expressió regular `pres*a` trobaria els mots *prea*, *presa*, *pressa*, *presssa*, etc.

- Operacions amb tipus de lletra:
  - seleccionar la grandària de la lletra (en *punts*, 1 polzada =2,54 cm = 72 punts)
  - seleccionar la família tipogràfica (Times, Courier, Helvetica...)
  - seleccionar el tipus: redona, cursiva, negreta, subíndex, superíndex, etc.
- Operacions de formatatge (normalment els paràgrafs es formaten sols, sense intervenció de la persona usuària en el procés, segons que el va teclejant):
  - Seleccionar l'alineació o la justificació de les línies del text (alineades a la dreta, a l'esquerra, centrades, o justificades[9]);
  - seleccionar el format de la pàgina (grandària del paper, marges inferior, superior, dret i esquerre), etc.
- Altres eines
  - demanar una *correcció ortogràfica* del text;
  - accedir a un *diccionari de sinònims*, etc.

4.3 Reconeixement automàtic de la parla

El *reconeixement automàtic de la parla* (RAP) es pot definir com la producció de textos informatitzats —en *temps real*, és a dir, tan instantàniament com siga possible— a partir de la veu humana (vegeu Samuelson-Brown (1996)). El RAP de propòsit general està encara molt lluny de ser perfecte, i, de fet, és encara un camp de recerca actiu. En canvi, el RAP per a un propòsit específic (per exemple, la consulta telefònica d'horaris de trens) està molt més avançat. La major part de la inversió de la comunitat internacional en RAP és, per raons òbvies, sobre l'anglés.

El RAP genera text a partir de la veu recollida a través d'un micròfon utilitzant una targeta de so per a digitalitzar-la i després un sistema de reconeixement automàtic de la veu (*automatic speech recognition*) per a detectar fonemes, síl·labes o paraules completes (depén del sistema concret) i traduir-les posteriorment a un text informatitzat. Hi ha sistemes de reconeixement *independents del parlant* i sistemes *dependents del parlant* (els quals normalment han de ser *entrenats* per la persona abans de l'ús). El RAP és especialment difícil per la gran variabilitat acústica que presenten els fonemes:

[9] No s'ha de dir *justificat a la dreta*: la justificació és sempre als dos marges alhora; el que s'ha de dir en aquest cas és *alineat a la dreta*.

- segons el context articulatori (per exemple, no és igual el so del fonema palatal representat pel dígraf *ig* en "passeig curt" —sord— que en "passeig allargat" —sonor—);
- segons el parlant (cada persona té uns organs fonadors de forma diferent —acústicament diferents— i processos de producció de la parla diferents —per exemple, n'hi ha qui parla més a poc a poc i qui parla molt de pressa—);
- segons el dialecte del parlant (per exemple, els valencians fem africades les *j* que en català central són palatals fricatives sonores).
- segons l'estat emocional del parlant, etc.

És un fet ben establert que, per a superar aquestes dificultats, els humans fem un ús molt intensiu dels coneixements lingüístics que tenim sobre l'idioma que estem escoltant i del context comunicatiu: així, si sentim dir "*percà nom* passes l'*antre xoc* de *craus*" a un amic quan veiem que no pot obrir el cotxe, entenem perfectament què ens vol dir; o si sentim dir en veu alta "me hu han dim moltis baltes" és molt probable que entenguem clarament "me ho han dit moltes voltes" a pesar dels canvis fonètics, ja que inconscientment busquem la interpretació correcta més propera al que hem sentit (en el context concret en què es diu la frase)[10]. Els resultats de la RAP són especialment dependents de les particularitats lingüístiques de la llengua involucrada i l'èxit depén de l'existència d'un bon *model de llengua* —ràpid i concís, és a dir, computacionalment eficient— que simule la part no contextual de la comprensió humana i permeta obtenir el text més probable en un idioma determinat a partir del text en brut produït pel sistema de RAP. La major part dels sistemes usen vocabularis grans i models estadístics.

4.4 Reconeixement automàtic de textos escrits

El *reconeixement automàtic de textos escrits* (RATE) es pot definir com la producció de textos informatitzats a partir de textos manuscrits o tipografiats. En el cas de textos tipografiats la tasca és molt més senzilla; en el cas de manuscrits, la complexitat és comparable a la del reconeixement de la parla.

El RATE genera un text informatitzat a partir d'un document imprés, usant un escàner (o *scanner*) i un programa de reconeixement òptic (també se'n diu *automàtic*) de caràcters (OCR, *optical character recognition*). Primerament, el document imprés és llegit (escandit o escanejat) usant l'escàner, i se'n genera un fitxer que en conté la imatge digital (per exemple, una graella molt fina de quadrats blancs i negres). Després, el programa

[10] Considereu aquest doblet anglés clàssic sobre el tema: *people can easily recognize speech* no és molt diferent de *people can easily wreck a nice beach.*

d'OCR llig la pàgina, descobreix on són els paràgrafs, les línies i, finalment, els caràcters concrets, i els transforma en un text informatitzat (normalment bastant imperfecte, especialment si és manuscrit). Com en el cas del reconeixement de la parla, és crucial l'ús d'informació sobre l'idioma concret (diccionaris, estadística sobre les seqüencies de lletres) per a corregir els errors de l'OCR. Per exemple, si un programa de lectura automàtica de textos produeix per error el text "4ixò 6s uua mcrda" no cal dir què hi llegim sense massa problemes, malgrat els errors en tots els mots; això és gràcies als nostres coneixements sobre les seqüències de lletres comunes en català.

4.5 Qüestions i exercicis

1. Com es diuen les combinacions especials de caràcters ASCII de l'estil de "`<B>`" o "`</B>`" que s'usen per a indicar tipus de lletra, grandàries de lletra, formats, etc., en algunes codificacions?

   (a) Marques

   (b) Protocols

   (c) Carpetes

4.6 Solucions

1. (a)

Capítol 5

Bases de dades

Com ja s'ha dit en la pàg. pg:fitxer, anomenem *fitxers* els conjunts de dades que es guarden en un mitjà d'emmagatzematge secundari, que es manipulen com un tot i que s'identifiquen per un nom. Molts dels fitxers que usen les persones que es dediquen a la traducció són fitxers (o *documents*) de text de diversos formats, com els descrits en l'epígraf 4.1, però també n'hi ha que es corresponen amb el significat del mot *fitxer* fora de la informàtica: contenen *fitxes*, totes amb un format més o menys constant; per exemple, totes les fitxes d'un fitxer bibliogràfic contenen informació sobre els autors, el títol, l'any de publicació, etc. En informàtica, els fitxers d'aquesta mena se solen anomenar normalment *bases de dades*; les fitxes s'anomenen *registres* i cada element d'informació de la fitxa s'anomena *camp*; així, els registres d'una base de dades bibliogràfica contenen informació en camps: un per als autors, altre per al títol, etc.

Quan volem buscar una determinada informació en un fitxer de fitxes de cartolina (per exemple, quins autors han usat mot *arbre* en el títol de les seues obres), i aquest fitxer no està ordenat segons cap criteri convenient, ens veurem obligats a mirar totes les fitxes una per una. Però és comú que els fitxers estiguen ordenats segons un dels seus camps: per exemple, un fitxer bibliogràfic pot estar ordenat pel cognom del primer autor, o per la matèria. Si la consulta o la recerca que volem fer es refereix al camp pel qual s'ha establit l'ordenació, és molt més senzilla que si es refereix a un altre camp, i es pot completar sense mirar totes les fitxes; per exemple, fent-hi una *recerca dicotòmica*: mirem la fitxa que hi ha enmig del fitxer; si ens hem passat, repetim l'operació amb la primera meitat del fitxer, i si ens hem quedat curts, ho fem amb la segona meitat. Es pot demostrar que la recerca dicotòmica mira com a molt n fitxes si el fitxer té entre 2^{n-1} i 2^n fitxes; per exemple, si el fitxer té 1234 fitxes, hi ha prou amb $n = 11$ recerques perque $2^{10} = 1024$ i $2^{11} = 2048$.

Però un fitxer de fitxes de cartolina només es pot ordenar seguint un únic criteri. Si volem facilitar les consultes associades a més d'un camp (per

exemple, autors i matèries) ens veurem obligats a mantenir dues còpies del fitxer sencer, cada còpia ordenada per un criteri.

L'organització de la informació en forma de base de dades simplifica enormement les consultes, ja que els ordinadors són molt més ràpids i segurs a l'hora de, per exemple, comparar el contingut de un determinat camp de totes les fitxes amb un cert valor o patró (per exemple, els autors que comencen per *Per*) i llistar el contingut d'un altre camp (per exemple, el títol) per a cada fitxa coincident. El programa que permet fer, entre altres, operacions de consulta d'una base de dades, és un programa *gestor de bases de dades* o bé l'inclou o l'invoca.

Si les consultes més freqüents es refereixen a un camp[1] que pren valors que es poden ordenar, els registres es poden ordenar per aquest camp, igual que el fitxer de fitxes de cartolina. Però un dels avantatges més clars de les bases de dades és que permeten que els registres estiguen ordenats per més d'un camp, sense haver de duplicar la base de dades. Això s'aconsegueix mitjançant un procediment anomenat *indexació*: bàsicament, s'assigna un número a cada fitxa i es construeix una taula o *índex* ordenat (una altra base de dades) que conté registres amb dos camps: un, el camp pel qual es vol ordenar, i l'altre, la posició en la base de dades del registre que conté aquest valor del camp (en cert sentit, aquest índex no és massa diferent de l'índex alfabètic que hi ha al final d'alguns llibres: busquem el mot alfabèticament i ens diu en quina o quines pàgines se'n parla).

Es pot construir un índex per a cada un dels camps associats a les consultes més freqüents i així s'evita recórrer tota la base de dades cada vegada que es fa una consulta: es busca el valor del camp en el registre corresponent i, quan es troba, s'usa la posició del registre per a accedir-hi directament. Una base de dades amb aquestes propietats està *indexada*. A més de fer-hi *consultes*, hi ha altres operacions que es poden fer amb una base de dades; les més importants són les *altes* o addicions de registres, les *baixes* o eliminacions de registres, i les *modificacions* de la informació continguda en un o més registres. Els índexs s'han de refer parcialment quan es fan aquestes operacions per tal que les consultes continuen sent eficients. Quan creem una nova base de dades i definim l'estructura de camps que tindrà cada un dels seus registres, podem designar quins dels camps corresponen als índexs.

Normalment, els usuaris reals no executen un programa gestor de bases de dades universal, sinó que usen programes o *aplicacions* que simplifiquen la creació, el manteniment i l'ús de la base de dades per a un perfil d'usuari concret.

[1] O a una combinació de camps, com ara el dia, el mes i l'any que formen una data.

5.1 Bases de dades lèxiques o terminològiques

Un dels programes més comunament usats pels professionals de la traducció són els gestors de bases de dades lèxiques (normalment anomenats gestors de bases de dades *terminològiques*, encara que es poden usar per a moltes altres aplicacions a més de les estrictament terminològiques).

Els registres d'una base de dades lèxica multilingüe poden contenir molts tipus de camps:

- El terme en cada una de les llengües (camps que normalment s'usen d'índex per a fer les recerques més eficients).
- El sentit (entre els possibles sentits del terme) al qual es refereix la fitxa o registre actual.
- L'autor de la fitxa (quan més d'una persona gestiona la base de dades).
- La definició del terme en una o més llengües.
- El camp temàtic de la fitxa.
- Altres termes relacionats.
- Informació sobre la morfologia o la flexió del terme en cada una de les llengües.

Una base de dades d'aquesta mena la pot consultar una persona mentre està fent una traducció manualment o pot estar inclosa dins d'un programa de traducció automàtica o assistida per ordinador. Per exemple, moltes memòries de traducció (vegeu el capítol 10) inclouen bases de dades terminològiques d'aquesta mena i permeten que la persona usuària les mantinga i les consulte, bé usant un programa independent, o bé des del processador de textos que preferisca.

Capítol 6

Traducció i traducció automàtica

Aquest curs tracta sobre la *traducció automàtica* (TA). Abans de considerar l'automatització de la traducció fóra bo que ens paràrem un poc per a discutir què vol dir exactament el mot *traducció*. Sobre la relació entre traducció humana i traducció automàtica, vegeu la secció 11.3.

6.1 Què és la traducció?

Per començar, s'ha de tenir en compte que el mot *traducció* és ambigu perquè es pot referir al *procés* de traduir o al *producte* (resultat) d'aquest procés.

Sager (1993)[1] comença la seua definició dient que, com a procés, es pot anomenar traducció "un rang d'activitats humanes deliberades, que es fan com a resultat d'instruccions rebudes d'un tercer, i que consisteixen en la producció de textos en una llengua meta (LM), basada, entre altres coses, en la modificació d'un text en una llengua origen (LO) per a fer-lo adequat a un propòsit nou", però encara no descriu la naturalesa de la modificació.

Com a producte, una *traducció* es pot identificar com a tal perquè és un document (en LM) derivat d'un altre document en un altre idioma (LO), i que manté una certa similitud de contingut amb aquest.

Es poden dir encara més coses sobre la traducció:

- les traduccions solen estar escrites en un subllenguatge particular (registre, especialitat...) de la comunitat lingüística de la LM, basat en un subllenguatge paral·lel de la LO;
- els documents i les traduccions corresponents es poden classificar en tipus[2] i aquesta tipologia afecta la traducció;

[1] Els conceptes d'aquesta secció estan presos quasi íntegrament d'aquesta obra.

[2] Per exemple, *carta comercial*, *edicte municipal*, *comentari editorial*, *manual tècnic informàtic* o *recull de poemes*.

- la traducció es veu afectada per elements extralingüístics perquè, normalment, els documents són entitats que uneixen l'expressió lingüística amb l'expressió no lingüística;

- les traduccions tenen un *receptor* o *lector*; una traducció, com a acte comunicatiu ha de considerar, a més de la intenció de la traducció, les expectatives dels lectors, que resulten del seu rerefons cultural i de les seues necessitats comunicatives, i que influeixen en la recepció del text traduït;

- la traducció sempre té una motivació: la superació de barreres comunicatives; per això se n'ha creat una professió.

Es pot aprofundir un poc més en la definició de traducció que hem considerat més amunt, revisant definicions existents (algunes preses de Sager (1993)):

- Nida (1966, p. 19): "La traducció consisteix a produir en la llengua meta l'equivalent natural més proper del missatge en la llengua origen, primerament quant al significat i segonament quant a l'estil." Sager diu que, més que *natural* (en sentit absolut) caldria dir *adequat* (a la tasca concreta). Aquesta definició introdueix dues de les tres dimensions bàsiques d'un document escrit (original o traduït): el *contingut* (significat) i la *forma* (estil), però oblida el *propòsit*.

- Flamand (1983): traduir és representar amb precisió (fidelitat a l'autor) un missatge en LO en una forma autèntica i correcta de la LM, adaptada al contingut i al receptor (fidelitat al lector)". El problema d'aquesta definició és la indefinició del concepte de *fidelitat*.

- Jakobson (1966): "Traducció és la interpretació de signes verbals per mitjà d'una altra llengua". Aquesta definició evita el concepte d'*equivalència* i introdueix el d'*interpretació* com a conjunt de processos cognitius que tenen lloc en la ment del traductor.

- En el *Diccionari de la Llengua Catalana*[3] es defineix *traducció* com la "reproducció del contingut d'un text o d'un enunciat oral, formulat en una llengua, en formes pròpies d'una altra llengua" (i *traduir* com "escriure o dir en una llengua allò que ha estat escrit o dit en una altra"). La definició inclou, per tant, el tractament i la producció de missatges no textuals (orals).

- Alcaraz Varó i Martinez Linares (1997) defineixen traducció com "expresión de un enunciado en la lengua de llegada [lengua meta] que sea

[3] Editorial Enciclopedia Catalana, 7a ed., 1987-

> equivalente al de la lengua de partida [lengua origen]"; queda per definir la noció d'*equivalència*, que els mateixos autors defineixen així: "la posesión del mismo valor por parte de los enunciados de la lengua de partida y de la de llegada"; l'equivalència pot ser *semàntica*, *estilística* i *textual*.

Per a acabar aquest apartat, convé esmentar alguns processos que no s'anomenaran *traducció* en aquest curs:

- l'adaptació de textos antics a la forma moderna d'un idioma;
- la traducció de mots i frases quan s'ensenya un nou idioma;
- la interpretació (de missatges parlats);
- la codificació (en Morse, etc.).

6.2 Traducció automàtica

6.2.1 Definició

La traducció automàtica (TA) es pot definir com el procés (o el producte) de traduir un text informatitzat[4] en una llengua origen a un text informatitzat en una llengua meta mitjançant l'ús d'un programa d'ordinador. Normalment es reserva la denominació *traducció automàtica* per a la completament automàtica; quan s'hi produeix intervenció humana es parla de *traducció assistida per l'ordinador* o de *traducció semi-automàtica*. El capítol 7 està dedicat a analitzar les diverses modalitats d'interacció entre persones i màquines en traducció.

Un aclariment és necessari sobre el tractament dels textos informatitzats. Quan els programes de traducció automàtica i semiautomàtica han de tractar documents estructurats (com els discutits en els epígrafs 4.1.6 i 4.1.8) han de ser capaços d'identificar les parts dels documents que corresponen als textos que s'han de traduir, destriant-les de les etiquetes. Normalment, els programes tenen un mòdul inicial que podríem anomenar *desformatador* i un mòdul final que podríem anomenar *reformatador* i que restitueix les etiquetes de manera que el format i l'estructura del document es conserven tant com siga possible. En general, aquestes operacions es poden considerar bàsicament independents del procés de traducció mateix —com farem en aquest llibre—, però hi ha programes més avançats que són fins i tot capaços d'usar la informació de les etiquetes com a context per a elegir una traducció on hi ha més d'una alternativa.

[4] Anomenarem *text informatitzat* un fitxer d'ordinador que conté un text codificat en un format conegut (vegeu el capítol 4)

Les referències que s'han fet en l'epígraf 6.1 al propòsit o motivació de la traducció i a la tipologia dels documents que han de ser traduïts són també molt importants a l'hora d'analitzar la traducció automàtica.

6.2.2 Sobre el nom en altres llengües

Convé comentar de pas que en anglés, la traducció automàtica s'anomena *machine translation* i s'abreuja MT, paral·lelament a l'alemany, que usa la denominació *maschinelle übersetzung*; en aquestes dues llengües s'expressa la noció d'automatisme mitjançant la referència a una *màquina*. En canvi, en francés o en rus es parla, com en català o en castellà, de *traduction automatique* o *avtomatitxeski perevod*, respectivament.

6.3 Història de la traducció automàtica

La major part del discutit en aquest apartat està pres dels treballs de John Hutchins, especialment de Hutchins (1995) i Hutchins (2001).

6.3.1 Els pioners, –1954

La traducció mitjançant màquines és una ambició humana des de fa segles que no es va fer realitat fins al XX. No feia molt que s'havia creat el primer ordinador, quan ja es va començar a pensar en la possibilitat d'usar-los per a traduir llenguatges humans.

Tot i que en els decennis dels 1930 i 1940 hi va haver alguns treballs precursors, és als primers cinquanta quan comença realment la recerca en TA en moltes universitats arreu al món, especialment als Estats Units. Els recursos de maquinari, programari i llenguatges de programació eren massa reduïts i la primera aproximació va ser la traducció mot per mot basada en diccionari amb algunes regles senzilles de reordenament (actualment anomenada *traducció directa* (vegeu l'apartat 9.2). Aquesta manca de recursos va fer que els primers objectius foren molt modests i, així, els primers investigadors van concentrar-se en el desenvolupament de llenguatges controlats (vegeu 7.5) i en l'ajuda humana en tasques de preedició i postedició (vegeu 7.4); era prou clar que els sistemes reals no podrien produir més que traduccions de molta baixa qualitat. El 1952 es va celebrar als Estats Units el primer congrés sobre TA on es van definir les línies fonamentals a seguir.

6.3.2 El decenni de l'optimisme, 1954–1966

La primera demostració pública d'un sistema de TA va ser desenvolupada per IBM i la Universitat Georgetown el 1954. Es va traduir a l'anglés un

conjunt de 49 frases en rus[5] usant un diccionari de només 250 mots i 6 regles gramaticals. Tot i que els resultats no eren massa bons, el públic i la indústria van creure que en uns anys es podrien aconseguir traduccions automàtiques de qualitat de documents científics i tècnics. Aquesta idea es va reforçar pel fet que van començar a aparèixer millores significatives en el maquinari, els primers llenguatges de programació i moltes millores en la lingüística formal (especialment en l'àrea de la sintaxi). L'entusiasme va fer que es finançaren un munt de projectes entre la meitat dels 50 i la meitat dels 60, projectes dins els quals van nàixer la major part de les tècniques actuals, com ara la traducció indirecta per transferència o la traducció per interlingua (vegeu el capítol 9).

L'objectiu era el desenvolupament de sistemes perfectes. Calia reduir al mínim la intervenció humana en el procés de TA, fins assolir la independència total i una qualitat comparable a la dels humans. Pràcticament ningú va considerar com es podria traure profit d'un sistema imperfecte: per què pensar-hi si aviat es disposaria de sistemes perfectes? Els traductors es van sentir amenaçats. No obstant això, algunes veus es pronunciaren en contra del perfeccionisme dominant i defensaren una aproximació més a llarg termini al problema i la construcció de sistemes que feren un ús efectiu de la interacció persona-màquina.

Un decenni després, i com que les expectatives eren tan altes, els avanços eren escassos i el futur pròxim no semblava poder millorar la situació. Molts investigadors començaven a trobar barreres de tot tipus, especialment semàntiques, que semblaven massa difícils de superar i que exigien mètodes més complexos. La Acadèmia Nacional de les Ciències dels Estats Units va publicar el 1966 l'informe ALPAC (Automatic Language Processing Advisory Committee) en el qual es recomanava que els nombrosos recursos que es dedicaven a la recerca en TA s'utilitzaren per tasques menys pretencioses i més bàsiques relacionades amb el processament del llenguatge natural i amb el desenvolupament d'eines de suport per als traductors com ara diccionaris automàtics. La conclusió era que només després de conéixer les arrels del problema, podria estudiar-se la realització d'un sistema de TA real. L'informe assegurava que la TA era més lenta i menys exacta que la feta pels humans, a més de ser el doble de cara, i que no hi havia cap indici de l'obtenció en el futur més o menys inmediat d'un sistema de TA útil. L'informe va fer que es reduira significativament el nombre de persones que es dedicaven a la TA i que els laboratoris començaren a treballar en el que es va conéixer com a lingüística computacional.

[5] Per raons polítiques i militars aquestes llengues van ser les elegides per als primers sistemes de TA.

6.3.3 Després de l'informe ALPAC (1966) als vuitanta

L'informe va acabar quasi virtualment amb la recerca en TA als Estats Units (també va tenir un impacte negatiu en els projectes desenvolupats a la resta del món) i durant molts anys la TA va ser percebuda com un autèntic fracàs. Tot i això, alguns grups van continuar treballant a Canadà i a Europa i van aparéixer els primers sistemes que funcionaven; el 1970 el sistema Systran va començar a ser usat per la USAF (United States Air Force) i el 1976 per la Comissió de la Comunitat Europea. També el 1976 apareix Metéo, desenvolupat per la Universitat de Montréal, que tradueix al francés els informes meteorològics. Per aquesta època, a més, els sistemes de TA comencen a ser demanats per empreses i administracions i no sols per traduir textos científics i tècnics.

Des de l'informe ALPAC el camp va patir una redefinició progressiva vers una concepció de la TA com un procés en el qual els traductors humans juguen un paper bàsic, i comencen a desenvolupar-se eines de traducció pensant en aquesta intervenció.

Els principals corrents dins la TA des dels 70 son, per tant: eines de suport a la traducció per a traductors, sistemes de TA amb intervenció humana i recerca teòrica vers un sistema completament automàtic de traducció.

6.3.4 Els primers vuitanta

Als 1980 apareixen nous sistemes de TA arreu al món amb expectatives més reals i l'interés en la TA resorgeix. Són especialment importants els resultats obtinguts a diverses empreses com Xerox on s'elimina quasi completament la postedició gràcies al control de la llengua origen; això permet la traducció senzilla dels manuals tècnics en anglés de la companyia a un gran nombre d'idiomes (francés, alemany, italià, espanyol, portugués i llengues escandinaves).

Durant aquest decenni els esforços es dirigeixen vers la traducció indirecta amb representacions intermèdies o sense (com la interlingua; vegeu l'apartat 9.4) mitjançant anàlisis morfològiques i sintàctiques i, de vegades, coneixements no lingüistics. Els projectes més notables són GETA-Ariane (Grenoble), SUSY (Saarbrücken), Mu (Kyoto), DLT (Utrecht), Rosetta (Eindhoven), el projecte de la Universitat Carnegie-Mellon (Pittsburgh) i dos projectes internacionals: Eurotra, suportat per la Comunitat Europea i el projecte japonés CICC amb participants a China, Indonesia i Tailàndia.

Eurotra és un dels projectes de traducció més coneguts del decenni dels 1980. El seu objectiu era la construcció d'un sistema de transferència multilingüe que permetera la traducció entre totes les llengües de la Comunitat Europea. Tot i que la traducció resultant tenia bastant qualitat, necessitava una gran quantitat de postedició. El projecte va estimular la investigació a tota Europa, però va ser abandonat finalment el 1992.

En aquests anys es consolida la idea que els sistemes de TA no són per a traductors; un traductor necessita eines que li faciliten el treball: diccionaris, bases de dades terminològiques (vegeu el capítol 5), sistemes de comunicació, memòries de traducció (vegeu el capítol 10), etc. De fet, actualment la postedició no s'encarrega a traductors (que no consideren açò com a part del seu treball), sinó a persones preparades específicament.

Tothom accepta ja en aquest decenni la importància dels llenguatges controlats i els subllenguatges en la TA, com ja havien defensat els precursors de la TA durant el decenni dels cinquanta.

El sistema comercial més sofisticat dels 1980 és Metal (1988), finançat per Siemens i que tradueix de l'alemany a l'anglés. Es tracta bàsicament d'un sistema per transferència indicat per a la traducció de documents relacionats amb el processament de dades i les telecomunicacions.

Al final dels 1980 comença l'aplicació de tècniques d'inteligència artificial al processament del llenguatge humà (sistemes experts i sistemes basats en el coneixement dissenyats per entendre els textos).

6.3.5 Els primers noranta

Tots els sistemes de TA dels vuitanta, tant els de transferència com els d'interlingua, funcionen bàsicament a partir de regles lingüístiques. Als 1990, però, apareixen noves estratègies conegudes com a mètodes basats en corpus. Els mètodes basats en corpus es poden dividir en dos grups: estadístics i basats en exemples.

Els mètodes estadístics ja van ser considerats als anys seixanta, però aviat van ser descartats perquè els resultats obtinguts no eren acceptables. Ara, però, el descobriment de noves tècniques va fer possible projectes com Candide a IBM. Candide usa mètodes estadístics per a l'anàlisi i la generació, però cap regla lingüística. Els treballs a IBM van utilitzar el gran corpus de textos en anglés i francés resultants de les sessions del Parlament de Canadà. El mètode consisteix a alinear en primer lloc les frases, els grups de mots i els mots en els dos textos i calcular després la probabilitat que un mot del text origen corresponga a un o més mots del text meta amb el qual ha estat alineat.

Els metodes basats en exemples (vegeu el capítol 10) s'aprofiten tanmateix de l'existència de grans corpora de textos traduïts (per això també s'en diu basats en memòria). La idea fonamental es que el procés de traducció es pot fer sovint consultant traduccions anteriors i identificant frases o grups de mots en el corpus ja traduït. Per poder dur a terme la traducció és necessari que els textos del corpus hagen estat alineats prèviament (mitjançant mètodes estadístics o mètodes basats en regles).

Tot i que la gran innovació dels noranta van ser els mètodes descrits abans, la recerca i el desenvolupament dels sistemes clàssics també va continuar: per exemple, el projecte Eurolang basat en el sistema de transferència

Metal pot traduir de l'anglès al francés, alemany, italià i espanyol, i viceversa. Ens els darrers 10 anys, un dels camps amb més investigacions ha estat el de traducció de la parla, una idea que evidentment ha estat present des de fa dècades, però que només ara es pot materialitzar parcialment. L'objectiu no és obtenir un sistema de traducció perfecta, sinó un sistema adequat per a aplicacions amb llenguatges, dominis i usuaris restringits. El principals són els desenvolupats a ATR, CMU i el projecte Verbmobil.

Una característica important dels primers 1990 és l'aparició de les primeres aplicacions pràctiques per a traductors: eines de suport a la traducció, diccionaris i bases de dades terminològiques, processadors de text multilingües, accés a glossaris i terminologies electròniques, eines de comunicació (escàners, OCRs, Internet; vegeu els capítols 3 i 4) o eines per a entorns restringits. La combinació d'algunes d'aquestes eines en un programari concret és el que es coneix com *estacions de treball per a traductors* (per exemple, el Translation Manager d'IBM o el Translator Workbench de Trados). La major part d'aquestes estacions de treball estan disponibles per a ordinadors personals.

6.3.6 Del darrers noranta a l'actualitat

La TA i les eines de suport a la traducció son cada vegada més usades per les grans empreses i per les administracions, principalment per a la traducció de documentació tècnica.

Al llarg dels darrers anys amb la generalització de l'ús d'Internet s'han desenvolupat serveis de traducció disponibles en línia i nombroses eines per a l'assimilació de continguts electrònics com ara documents HTML i missatges de correu electrònic.

Des dels seus inicis, quasi tota la recerca i quasi tots els sistemes comercials de TA s'han centrat en els principals idiomes internacionals: anglés, francés, espanyol, japonés, rus, etc. Encara resta molt a fer amb les altres llengues del món.

Capítol 7

Usos de la traducció automàtica

7.1 Utilitat de la traducció automàtica

En molts àmbits la traducció automàtica està encara molt lluny de poder competir en qualitat amb la realitzada per traductors professionals (vegeu el capítol 11). Per exemple, en molts casos és difícil aconseguir que l'ordinador sàpia elegir la interpretació correcta entre les possibles interpretacions d'un enunciat ambigu com

Els soldats van disparar als xiquets. Els vaig veure caure.

ja que això requereix l'ús de quantitats enormes de coneixement enciclopèdic sobre el funcionament del "món real". En aquest cas extrem, el sistema ha de saber: que els trets fereixen greument o maten les persones que els reben i que la condició de ferit greu o mort és incompatible amb mantenir-se dret, i que, per tot això, la interpretació més probable és que van caure els xiquets, no els soldats. Com a conseqüència de problemes com aquest i d'altres de naturalesa diversa, en moltes aplicacions, la traducció produïda per un bon programa s'ha de considerar com un esborrany que ha de ser revisat.

Però, per posar-nos en l'altre extrem, ha de quedar clar que la traducció automàtica pot ser molt útil en aquelles situacions en què l'ús d'un traductor professional siga impracticable o impossible econòmicament. Algunes d'aquestes situacions són:

- La traducció automàtica de correu electrònic entre les persones d'un grup de treball internacional amb la finalitat d'agilitzar les comunicacions; la traducció immediata de documents durant la *navegació* per Internet (de fet, hi ha programes especialment dissenyats per a aquesta finalitat), o la traducció automàtica de converses electròniques interactives (teclat i pantalla, *chat*).

- La traducció i el manteniment de tots els manuals tècnics d'una família de productes (per exemple, els manuals de manteniment d'una gamma d'automòbils).

En els epígrafs successius d'aquesta secció es donen alguns detalls sobre com es pot usar la traducció automàtica en situacions similars a les esmentades.

7.2 Assimilació i disseminació

Moltes de les aplicacions de la traducció automàtica es poden dividir en dos grans grups: l'*assimilació* d'informació (quan una persona usa la traducció automàtica per a obtenir informació a partir d'un document escrit en una altra llengua) i la *disseminació* —també anomenada *difusió*— d'informació (quan una persona usa la traducció automàtica per a produir documents que han de ser distribuïts a més d'un usuari). La traducció automàtica, tot i ser *imperfecta*, pot ser una eina molt útil en aquests dos grups d'aplicacions.

7.2.1 Assimilació

Per una banda, en situacions d'*assimilació* de la informació no sembla necessària una traducció perfecta, sinó més aviat una traducció ràpida i raonablement intel·ligible. És el cas de la traducció de correu electrònic, de documents d'Internet (durant la *navegació*) o de despatxos de premsa en altres idiomes. Una altra aplicació d'assimilació de la informació és l'anomenat *screening* o exploració de documents per a decidir quins són rellevants i mereixen una atenció més detallada; de fet, aquesta va ser una de les primeres aplicacions de la traducció automàtica als EUA: es volia tenir accés a la informació tecnològica present en documents de la Unió Soviètica. Els usos civils de l'*screening* han superat actualment l'ús tradicional, el militar. En el cas de l'*screening*, fins i tot una traducció incompleta a més d'incorrecta (per exemple, només dels mots terminològics) pot ser de gran utilitat. L'*screening* es pot fer usant un ordinador personal tot sol, o accedint a ordinadors servidors d'informació, en el cas de grans instal·lacions. És important indicar que en quasi totes les situacions d'assimilació el paper del traductor professional és inexistent, ja que el treball és de naturalesa molt diferent, i l'ús d'un traductor professional seria molt car i molt lent. Una altra aplicació interessant és la traducció automàtica de *converses electròniques* (usant el teclat i la pantalla d'ordinadors connectats entre si) entre persones que parlen dos idiomes diferents. Les imperfeccions de la traducció es poden compensar amb preguntes o dient les coses d'una altra manera fins que els dos interlocutors s'entenguen (és a dir, mitjançant una *negociació*).

Per rudimentari que siga un sistema de traducció automàtica, pot ser molt útil en tasques d'assimilació. Una de les aproximacions més simples a la TA és l'anomenada *traducció mot per mot*, en què el programa identifica

cada mot, el busca en un diccionari bilingüe i el substitueix per una traducció aproximada. A tall d'exemple, considereu el següent text en tok pisin[1] (el text està pres de Lyovin (1997)):

> *Long taim bifo, wanpela ailan, draipela pik i save stap ya, na em i save kaikai ol man. Em i save kaikai ol man nau; wanpela taim, wanpela taim nau ol man go tokim bikpela man bilong ol, bos bilong ol, ol i go tokim em nau, em i tok: "Orait yumi mas painim nupela ailan".*

Si prenem un diccionari i traduïm el text mot per mot, prenent la primera traducció possible en cada cas —pot haver-n'hi més d'una—, s'obté el text següent[2]:

> *En temps passat, un illa, enorme porc - soler viure esmentat i ell - soler menjar més-d'un home. Ell - soler menjar més-d'un home aleshores; un temps, un temps aleshores, més-d'un home anar parlar gran home en més-d'un, cap en més-d'un, més-d'un - anar parlar ell aleshores, ell - dir: "Molt-bé, vosaltres-i-jo haver-de trobar nou illa".*

I ara, veritat que s'entén una miqueta més? La traducció idiomàtica *correcta* seria poc més o menys:

> *Fa molt temps, en una certa illa, vivia un gran porc i se solia menjar la gent. Se solia menjar la gent, i una vegada, la gent va anar i va dir al seu gran home, al seu cap, va anar i van parlar amb ell. Ell va dir: "Molt bé, hem de trobar una nova illa".*

L'ordre dels mots no és molt diferent en tok pisin i en català i això fa que la traducció mot per mot siga prou llegidora. En canvi, si el text original està en basc, les coses no són tan senzilles. El text, pràcticament inservible per a qui no sàpia basc:

> *Bazkaria bukatu ondoren Koldo egunkarira joan zen eta Teoren foto bat hartu zuen. Gero, egunkariaren ale zaharrak irakurri zituen, boxeo txapelketako berriak aztertzeko. Boxealarien izenak apuntatu zituen.*

es pot traduir mot per mot com:

[1] Llengua de contacte que es parla a Papua Nova Guinea i que té 50.000 parlants que la parlen com a primera llengua i més de dos milions de parlants que la parlen com a segona llengua.

[2] És possible que ja us hàgeu adonat que el tok pisin té molt vocabulari pres de l'anglés, com a llengua de contacte que és.

> *El-dinar acabat després Koldo al-diari anat era i de-Teo foto una pres l'havia. Després, del-diari número els-vells llegit els-havia, boxa del-campionat les-notícies per-a-examinar. Dels-boxadors els noms apuntat els-havia.*

que és molt més difícil de llegir que el resultat de traduir el text en tok pisin mot per mot. Una traducció idiomàtica possible és:

> *Després de dinar Koldo va anar al diari i va prendre una foto de Teo. Després, va llegir [els] números vells del diari per a examinar les notícies del campionat de boxa. Va apuntar els noms dels boxadors.*

Fixeu-vos que fins i tot en aquest cas tan desfavorable el text traduït mot per mot dóna bastants pistes sobre el significat del text original.

7.2.2 Disseminació

Per altra banda, en situacions de *disseminació* de la informació, pot ser útil restringir la llengua d'origen (no permetre'n totes les realitzacions possibles, ni tot el lèxic, ni tots els registres) a un llenguatge que puga ser traduït automàticament amb el mínim possible de problemes, és a dir, amb el mínim esforç de postedició, o almenys, amb un esforç acceptable per un revisor.[3] Açò és especialment important quan es tracta de traduir manuals tècnics a diversos idiomes. Les restriccions es poden expressar sota la forma de missatges interactius dirigits a la persona que prepara el document original.[4]

La traducció automàtica per a la disseminació és especialment eficient quan només es tradueixen textos pertanyents a una part molt reduïda i ben regulada de l'idioma en qüestió (un *subllenguatge*). Un exemple n'és Météo, el sistema que produeix informes meteorològics simultanis en francés i en anglés al Canadà.

7.3 CAT, HAMT i MAHT

Moltes situacions de traducció automàtica es poden classificar com a situacions de CAT, *computer-aided translation* o traducció assistida per ordinador, també anomenada *traducció semiautomàtica*. Però cal precisar millor què volem dir amb això d'"assistida per ordinador". Ací són rellevants les nocions de MAHT (*machine-aided human translation*, traducció humana assistida per una màquina) i HAMT (*human-aided machine translation*, traducció automàtica assistida per un humà), que estableixen les dues situacions bàsiques d'interacció entre una persona i un ordinador a l'hora de fer la traducció. Els paràgrafs següents en donen alguns exemples.

[3] És a dir, quan la revisió no és més costosa que refer tota la traducció a mà.

[4] Vegeu l'apartat 7.5, on es discuteix un concepte molt relacionat, el de *llenguatge controlat*.

HAMT: Un programa de traducció automàtica pregunta a l'usuari quan té més d'una possible traducció per a un mot o per a una frase. Aquesta i altres situacions de *negociació* del text d'origen amb l'usuari del sistema impliquen una interacció que també pot ajudar a preparar un text més correcte, és a dir, a *preeditar-lo* perquè puga ser traduït automàticament. Altres voltes, el programa pot analitzar l'estructura profunda de la frase i presentar-ne les possibles interpretacions a l'autor, per tal que resolga alguna possible ambigüitat. En aquests sistemes interactius, cal tenir en compte dos factors: el primer, que un sistema que pregunta massa no és còmode d'usar (no és *ergonòmic*) i el segon, que pot passar que l'usuari siga monolingüe, circumstància que canvia molt la naturalesa de la interacció entre el programa i l'usuari. Els usuaris d'aquest tipus de sistemes es podrien classificar en tres grans grups: traductors ocasionals, traductors professionals individuals i traductors professionals que treballen per a empreses de traducció.

MAHT: L'usuari (un traductor competent o un professional independent) utilitza diccionaris bilingües, *thesauri*, conjugadors i declinadors, correctors ortogràfics, sintàctics i d'estil, i formularis o models de documents, com a ajuda mentre produeix una traducció de manera manual usant un processador de textos. Altres eines —d'ús comú entre diversos traductors, i accessibles normalment com a recursos remots— poden ser les bases de dades terminològiques i les bases de dades lèxiques multilingües (vegeu l'epígraf 5.1), o les memòries de traducció (vegeu el capítol 10).

7.4 Preedició i postedició

Altres dos conceptes rellevants en el cas de sistemes de traducció automàtica imperfectes són els de *preedició* i *postedició*. En el primer cas, el text origen es prepara o adapta (*preedita*) per a millorar el comportament del sistema de traducció, eliminant, per exemple, l'ambigüitat del text[5], o també per a marcar parts del text que no han de ser traduïdes, com ara una citació, o que han de ser tractades de manera especial per no ser frases completes, com un títol. En el segon cas, el text meta produït pel sistema es refina o revisa (*postedita*) perquè siga gramaticalment correcte o estiga escrit d'acord amb un registre determinat.

7.5 Llenguatges controlats

Quan la traducció automàtica s'usa per a la disseminació de documents tècnics de temàtica homogènia, pot ser interessant fer que els documents

[5]Per exemple, en anglés tècnic, el mot *replace* presenta una *ambigüitat lèxica* (vegeu l'apartat 8.1), ja que pot voler dir *exchange* (reemplaçar) o *put back* (tornar a col·locar).

originals estiguen escrits usant un lèxic estàndard sense ambigüitats semàntiques i seguint unes regles sintàctiques i d'estil ben determinades, és a dir, en un *llenguatge controlat* (Wojcik i Hoard 1996; Arnold et al. 1994) dissenyat de manera que el resultat de la traducció automàtica puga ser usat directament per a publicar-lo amb el mínim possible de postedició.

Un *llenguatge controlat* és "un subconjunt del llenguatge natural definit amb precisió, d'una banda restringit quant al lèxic, a la gramàtica i a l'estil, i d'una altra, possiblement estés amb terminologia i construccions gramaticals específiques d'un domini" (Huijsen 1998).

Un exemple de llenguatge controlat és l'anglés simplificat (*Simplified English*) de l'AECMA (Associació Europea d'Indústries Aeroespacials), que es caracteritza per "una sintaxi senzilla, un nombre limitat de mots, un nombre limitat de significats ben definits per mot (normalment un), i un nombre limitat de categories lèxiques[6] per mot (normalment una)", amb "l'objectiu de produir textos breus i no ambigus" (AECMA [Associació Europea d'Indústries Aerospacials] 1998).

Alguns dels avantatges dels llenguatges controlats (Huijsen 1998) es poden resumir com segueix:

- els textos són més senzills i intel·ligibles;
- el manteniment dels documents es facilita;
- se simplifica el tractament computacional dels documents, en particular la traducció automàtica.

Quant als desavantatges, podem dir que:

- el poder d'expressió d'un llenguatge controlat és sempre més restringit;
- l'escriptura de textos en llenguatge controlat és més lenta;
- és necessària una inversió addicional de temps en l'aprenentatge del llenguatge controlat per part dels autors.

Els dos últims desavantatges es poden reduir si es dota els autors d'eines informàtiques, com ara d'un editor de textos intel·ligent que els ajude a escriure en el llenguatge controlat.

Un exemple clàssic de llenguatge controlat és PACE (*Perkins-approved controlled English*), l'anglés controlat usat per la companyia de motors Diesel Perkins, concebut especialment per a facilitar la traducció automàtica dels

[6] Les *categories lèxiques* (o simplement *categories*) són conjunts de mots que tenen la mateixa funció sintàctica; hi ha categories *majors*, *lèxiques* o *de classe oberta* (substantiu, adjectiu, verb, etc.) que creixen quan s'afig nou lèxic a la llengua i categories *menors*, *gramaticals* o *de classe tancada* (articles, conjuncions, etc.), que no creixen i contenen mots amb funció gramatical. La sintaxi es defineix normalment, no en termes de mots, sinó en termes de categories lèxiques.

manuals que descriuen les característiques i el manteniment d'aquests motors (Newton 1992). Altres exemples de llenguatges controlats són l'*ScaniaSwedish* usat per la firma de camions Scania (Almqvist i Sågvall Hein 1996), o el *Caterpillar Technical English* de la companyia de maquinària d'excavació Caterpillar.

7.6 Qüestions, exercicis i problemes

1. (*) Elegiu un idioma qualsevol que conegueu bé, L. És ben segur que L té mots polisèmics que en altra llengua L' tenen més d'una traducció, segons el sentit que se'n prenga. Elegiu tres mots de L que tinguen aquest problema i descrigueu com els tractaríeu en un llenguatge controlat basat en L. Les regles que formuleu per als autors que escriguen en el llenguatge controlat han de estar escrites en L i no han de contenir referències a altres llengues.

2. En els sistemes de traducció automàtica, la preedició...

   (a) ... redueix la quantitat de postedició.

   (b) ... és una alternativa a la postedició, que elimina completament aquesta última fase.

   (c) ... imposibilita l'ús del sistema per a tasques de disseminació d'informació.

3. Indica en quina d'aquestes situacions de traducció automàtica és menys crítica la perfecció lingüística de la traducció.

   (a) Joan usa el Web Translator mentre navega per per les pàgines Internet de la Universität Mainz per a saber quina assignatura dóna el professor Karl-Hans Lehninger i quins són els seus interessos investigadors.

   (b) Joan usa el Web Translator per a fer una versió en alemany de la seua pàgina Web.

   (c) El personal d'IBM tradueix patents europees per a detectar possibles avanços en correcció d'errors de comunicacions digitals.

4. Imagineu que podem elegir entre dos sistemes de traducció automàtica diferent t_A i t_B per a traduir manuals de televisors de l'anglés al francés, i que s'ha de dissenyar un anglés controlat per a minimitzar la postedició. Les regles de l'anglés controlat, poden dependre del sistema de TA elegit?

   (a) No, perquè els llenguatges controlats s'han de dissenyar independentment dels sistemes de TA.

(b) Sí, perquè en cada cas s'han d'evitar problemes diferents.

(c) No, perquè la llengua meta dels dos sistemes és la mateixa.

5. Indica quina d'aquestes situacions de traducció automàtica és d'*assimilació* d'informació:

   (a) Narcís usa el programa traductor de l'anglés al castellà Spanish Assistant per a llegir els documents electrònics que troba en Internet sobre la influència de l'èuscar sobre el gascó.

   (b) Joan usa el Web Translator per a fer una versió en alemany de la seua pàgina Web abans de publicar-la en Internet.

   (c) L'empresa Into the Wind tradueix automàticament el seu catàleg de milotxes i catxerulos a diverses llengües.

6. Moltes voltes, la preedició la fa l'autor quan interacciona amb el programa de traducció automàtica. És possible dissenyar un sistema de preedició interactiva per a autors monolingües?

   (a) Sí.

   (b) No. Per a preeditar correctament cal conéixer l'idioma de destinació.

   (c) Només per a certs idiomes amb estructura gramatical senzilla com l'anglés.

7. Quin dels següents *no* és un avantatge dels llenguatges controlats?

   (a) S'evita la necessitat que una persona interaccione amb el programa de traducció automàtica per a resoldre ambigüitats.

   (b) Els textos meta resultants són molt més curts.

   (c) Els textos origen es fan més intel·ligibles.

8. Per què és necessària la preedició en els sistemes de traducció automàtica?

   (a) Per a evitar construccions o frases difícils de traduir.

   (b) Perquè el format quede més agradable a la vista.

   (c) És una alternativa a la postedició.

7.7 Solucions

1. (*) Per exemple, si L és el castellà, mots com *escondite* poden referir-se a un lloc on amagar-se (1) o a un joc (2) (en L'=català, *amagatall* (1) i *fet*, *amagar*, *fet a amagar* o *conillets a amagar* (2)). En el llenguatge controlat, es podria evitar el primer significat proposant els autors que

feren servir el mot alternatiu *escondrijo*. Les regles es podrien formular com segueix en castellà:

escondite *úsese sólo en el sentido de "juego del escondite"; úsese* escondrijo *si se quiere indicar el lugar donde se esconde alguna persona o cosa.*

registro *úsese sólo en el sentido de "transcripción" u "oficina de registro"; úsese* inspección *cuando se refiera, por ejemplo a la investigación detallada de un local por parte de la policía.*

explotar *úsese sólo en el sentido de "aprovechar económicamente"; úsese* estallar *en el sentido de "deflagrar" (una bomba, etc.) o "reventar" (un globo, etc.).*

2. (a)
3. (a)
4. (b)
5. (a)
6. (a). Les preguntes es poden plantejar com en el problema 1.
7. (b)
8. (a)

Capítol 8

Ambigüitat

Un dels obstacles més importants per a la traducció automàtica és l'ambigüitat inherent al llenguatge humà. Podem dir que un enunciat (una oració, un text) és ambigu quan és susceptible de dues o més interpretacions (Alcaraz Varó i Martinez Linares 1997)[1] i, per tant, pot tenir més d'una traducció a un altre idioma.[2] En aquest capítol ens fixarem molt especialment en l'ambigüitat de les oracions.

Una de les perspectives més interessants per a analitzar i ordenar els tipus d'ambigüitat descrits més amunt ens la proporciona l'anomenat *principi de composicionalitat* (Radford et al. 1999):

> *La interpretació d'una oració està determinada per la interpretació dels mots que apareixen en l'oració i per l'estructura sintàctica de l'oració.*

Aquest principi explica per què la interpretació de l'oració

(8.1) *El pare escura plats*

és diferent de la de l'oració

(8.2) *La mare llegeix llibres*

Les oracions (8.1) i (8.2) tenen la mateixa sintaxi però diferent interpretació perquè contenen mots diferents amb interpretacions diferents. També explica per què l'oració

(8.3) *El gos va mossegar l'home*

[1] Don et al. (1996) ho expressen dient que l'ambigüitat és "el fenomen pel qual una expressió té més d'un significat".

[2] De vegades no és així: hi ha una única traducció que conserva l'ambigüitat de la frase original; d'això, se'n sol dir *free ride* ("passi gratuït"). Per exemple, l'oració castellana *Aprendió a afeitarse en dos minutos* es pot traduir al català *Va aprendre a afaitar-se en dos minuts* sense resoldre l'ambigüitat següent: és el temps que va tardar a aprendre o el temps que empra per afaitar-se?

no té la mateixa interpretació que la frase

(8.4) *L'home va mossegar el gos*

Aquestes oracions no volen dir el mateix perquè, malgrat tenir els mateixos mots, l'estructura sintàctica no és la mateixa.

Per això, no és possible assignar una interpretació clara a oracions sintàcticament incorrectes, com ara

(8.5) **Llegeix mare llibres la*

encara que els mots tinguen interpretació independentment, ni tampoc a una oració sintàcticament correcta que continga algun mot al qual no podem assignar cap interpretació:

(8.6) *La mare *ingurpleix llibres*

Com veurem més endavant, trobem una complicació addicional: en algunes oracions, hi ha parts de l'estructura sintàctica que no es reflecteixen en cap mot, perquè generen *categories buides* que no tenen una representació fonètica o gràfica explícita. Per exemple, l'oració

(8.7) *Té molts amics*

té un subjecte buit (també anomenat el·líptic). En aquests casos podem considerar que les categories buides són mots que tenen una interpretació.

Si una oració és ambigua quan té més d'una interpretació possible, això pot tenir dues causes bàsiques:

- un o més mots de l'oració tenen més d'una interpretació possible (és a dir, són *lèxicament ambigus*).
- l'oració té més d'una estructura sintàctica possible (és a dir, és *estructuralment ambigua* o *sintàcticament ambigua*).

Les dues causes poden concórrer. De fet, estudiarem tres casos: l'ambigüitat purament deguda a l'ambigüitat dels mots (explícits o nuls); l'ambigüitat purament deguda a l'existència de més d'una estructura sintàctica, l'ambigüitat deguda a les dues causes alhora.

Hi ha també certs tipus d'ambigüitat que no es poden explicar de manera senzilla amb el principi de composicionalitat, com ara *l'ambigüitat en l'abast dels quantificadors*, que també estudiarem en aquest capítol (apartat 8.4).

8.1 Ambigüitat deguda a l'ambigüitat lèxica

En moltes llengües, els mots es flexionen i prenen formes diferents. Un mot (i, en general, una unitat lèxica de més d'un mot) es pot veure des de dues perspectives; d'una banda, la *forma superficial* del mot és la forma concreta que apareix en el text: *cantàvem*; d'altra banda, hi ha la *forma lèxica*, que consisteix en

- un *lema* o *forma canònica* (*cantar*),
- una *categoria lèxica*[3], classe de mot, o part de l'oració (verb) i
- uns *indicadors de flexió* que expressen les característiques morfològiques o flexives (primera persona, nombre plural, temps pretèrit imperfet, mode indicatiu).

Quan dues formes lèxiques diferents tenen la mateixa forma superficial; és a dir, s'escriuen de la mateixa manera, se sol dir que són *homògrafes*; a més, s'anomena símplement *homògraf* a la forma superficial a què correspon més d'una forma lèxica; el fenomen s'anomena *homografia*. Per exemple, el mot *riu* és homògraf perquè té tres formes lèxiques: *riu*, substantiu, masculí singular; *riure*, verb, 3a. persona del singular, present d'indicatiu, i *riure*, verb, 2a. persona del singular, imperatiu.

Però l'homografia no és l'única causa possible d'ambigüitat lèxica; hi ha mots que són ambigus tot i tenir la mateixa forma lèxica, perquè el que és ambigu és la interpretació del lema. Aquests mots s'anomenen habitualment *polisèmics*. Per exemple, el mot *estació* (forma lèxica: *estació*, substantiu, femení singular) és polisèmic perquè el lema corresponent té més d'una interpretació: indret on s'aturen temporalment els trens, part de l'any compresa entre un solstici i un equinocci, conjunt d'instal·lacions per a un propòsit determinat (per exemple, l'esquí), etc.

L'ambigüitat d'una oració pot ser causada per diversos tipus bàsics d'ambigüitat lèxica.

1. L'oració conté una o més unitats lèxiques (p.ex. mots) polisèmiques: si diem que algú

   (8.8) *Treballa en l'estudi que li van encarregar*

   podem referir-nos a un investigador o a un decorador, depenent de quina interpretació assignem al mot polisèmic *estudi*. De l'ambigüitat d'aquestes unitats lèxiques, també se'n sol dir *ambigüitat lèxica pura*. Aquesta oració l'hem de desambiguar si la volem traduir, per exemple, a l'anglés, perquè en el primer cas hauríem de dir *study* i en el segon, *studio*; per això l'efecte de la polisèmia en traducció causa l'anomenada *ambigüitat de transferència*. L'ambigüitat lèxica de transferència és especialment perillosa quan afecta un mot de la llengua origen no percebut com a ambigu. Per exemple, el mot castellà *destino* es pot traduir al català com a *destí* (sort futura) o *destinació* (punt d'arribada).

   Un altre exemple: l'oració

   (8.9) *Han posat un banc nou a la plaça*

[3] Vegeu la nota al peu de la pàg. pg:catgra

pot tenir dues interpretacions, segons la interpretació que s'assigne al mot polisèmic *banc* ("seient estret i llarg" o bé "institució financera").

Una ambigüitat que és molt semblant a l'ambigüitat lèxica pura es produeix quan una expressió idiomàtica es pren bé com a tal o bé en sentit literal. Per exemple, la interpretació d'*enviar algú a pastar fang* pot ser la idiomàtica de dir a algú que deixe de molestar (castellà *mandar a freír espárragos*) però podria ser també la literal en un taller de terrisseria.

2. L'ambigüitat d'oracions que contenen un homògraf que té dues o més interpretacions però la mateixa categoria lèxica, i no afecta, per tant, l'estructura de l'oració. Hi ha tres situacions possibles

   - canvia només el lema però no els indicadors de flexió: el mot castellà *creo* pot ser la 1a. persona del singular del present d'indicatiu del verb *creer* o del verb *crear*.
   - no canvia el lema però sí els indicadors de flexió: el mot castellà *cantamos* pot ser la 1a. persona del plural del present d'indicatiu o del pretèrit indefinit (perfet simple) del verb *cantar*.
   - canvien el lema i els indicadors de flexió: el mot castellà *salen* pot ser la 3a. persona del plural del present d'indicatiu del verb salir o del present de subjuntiu del verb *salar*.

3. L'oració conté una *expressió anafòrica*, com ara un pronom, adjectiu possesiu, etc., la qual pot tenir, en principi, més d'una possible interpretació, però aquesta interpretació està determinada per la relació de *co-referència* entre l'expressió i el seu *antecedent* (un sintagma que es pot trobar en la mateixa oració o en una altra oració del text) o perquè es refereix a algun objecte o concepte exterior al text. La relació que assigna una interpretació a una expressió anafòrica s'anomena *deïxi*: quan la interpretació és per *co-referència* amb un antecedent que apareix anteriorment en el text s'anomena *anàfora*, i *catàfora* si l'antecedent és posterior. En la frase

   (8.10) *Vaig obrir [la porta]$_i$ a [la cuinera]$_j$ i la$_{i/j?}$ vaig fer passar*

   els índexs (i, j, i/j?) indiquen que el pronom feble *la* es pot referir a la mateixa persona a la qual ens hem referit amb el sintagma nominal *la cuinera*, però no hi ha cap raó sintàctica perquè el referent no siga el mateix que el del sintagma *la porta*: aquesta pot ser una possible causa d'ambigüitat en la segona oració coordinada.

4. L'oració té constituents que no es reflecteixen com a mots però als quals cal assignar una interpretació. En algunes llengües romàniques (en italià, castellà i català però no en francés) és comuna l'absència del

subjecte quan és de tercera persona. En aquest cas, la posició on hauria d'anar el subjecte es pot suposar ocupada per un pronom sense forma superficial que dóna lloc a ambigüitat mitjançant mecanismes molt similars als de l'anàfora i per tant, se'ls pot considerar mecanismes lèxics. En el fragment

(8.11) *Felip va apunyalar Marta. Joan va veure com queia redolant*

qui va caure redolant, Felip o Marta? O alguna altra persona? El problema és que falta el subjecte de l'oració subordinada *com queia rodolant*. Aquesta omissió dóna lloc a una ambigüitat. Quan es tracta de l'omissió del subjecte, se sol postular en lingüística l'existència d'un pronom especial anomenat PRO, sense forma superficial, que fa de subjecte nul,

(8.12) *Joan va veure com PRO queia redolant*

i al qual s'assigna interpretació mitjançant processos deïctics o anafòrics com els descrits per a altres expressions anafòriques.

Aquesta classe d'ambigüitats se sol incloure dins d'un grup de fenòmens més generals anomenats *ambigüitats per el·lipsi*. Alcaraz Varó i Martinez Linares (1997) defineixen l'*el·lipsi* com l'omissió o l'absència d'alguna part d'una oració. Com veurem més avall, de vegades l'el·lipsi dóna lloc a l'existència de més d'un arbre d'anàlisi sintàctica per a l'oració i per tant aquests tipus d'el·lipsi no es poden incloure pròpiament en aquest apartat dedicat a l'ambigüitat purament lèxica.

8.2 Ambigüitat estructural pura

L'ambigüitat d'una oració també pot estar deguda al simple fet que tinga més d'un arbre d'anàlisi sintàctica. S'hi poden distingir diversos casos:

1. *Ambigüitat estructural d'origen coordinatiu*: Per exemple, si diem

   (8.13) *Posa els llençols i els cobertors nets a l'armari*

   hi ha dues possibles interpretacions; en una els llençols no estan nets, en l'altra sí, segons que es considere que l'adjectiu *nets* modifica els dos substantius coordinats o només l'últim (vegeu la fig. 8.1). L'ambigüitat estructural associada a les conjuncions coordinatives se sol anomenar *d'origen coordinatiu*.

2. *Ambigüitat estructural d'adjunció* (angl. *attachment ambiguity*): es tracta d'un cas típic d'ambigüitat estructural que es manifesta quan hi ha un *adjunt* (típicament un sintagma preposicional) que es pot inserir de diverses maneres en l'arbre d'anàlisi sintàctica de la frase. Per exemple, la frase

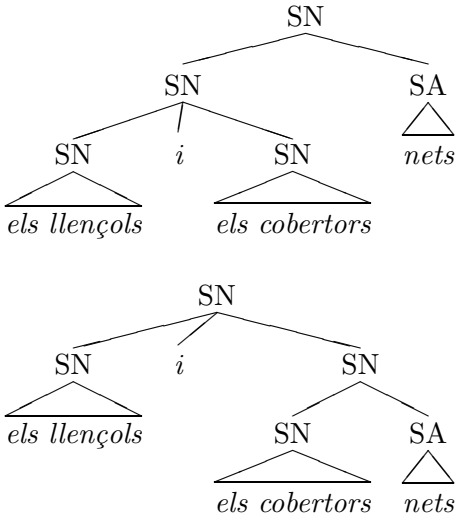


Figura 8.1: Dos arbres per a la frase "Posa els llençols i els cobertors nets a l'armari" (SN = sintagma nominal, SA = sintagma adjectival).

(8.14) *Joan va portar notícies de Grècia*

es pot interpretar de dues maneres: en una, el sintagma preposicional *de Grècia* modifica *notícies*; en l'altra, modifica *portar* (vegeu els arbres de la figura 8.2). Més exemples:

(8.15) *Va parlar amb l'encarregat de la neteja de la seua casa*

(8.16) *Hi ha una bossa de roba perduda en la Secretaria de l'Escola*

Tuson (1999) explica que aquesta última oració pot tenir fins a 12 interpretacions possibles.

3. L'*ambigüitat estructural deguda a l'el·lipsi* d'un o més constituents de l'oració, especialment quan aquesta oració hauria de tenir, si s'hagués produït en forma explícita, una estructura paral·lela a la d'una oració anterior (per exemple, en coordinacions, comparacions, etc). Considerem l'exemple següent, tret de Radford et al. (1999):

(8.17) *Els escocesos aprecien el whisky més que els gal·lesos*

L'oració té dues interpretacions:

(8.18)

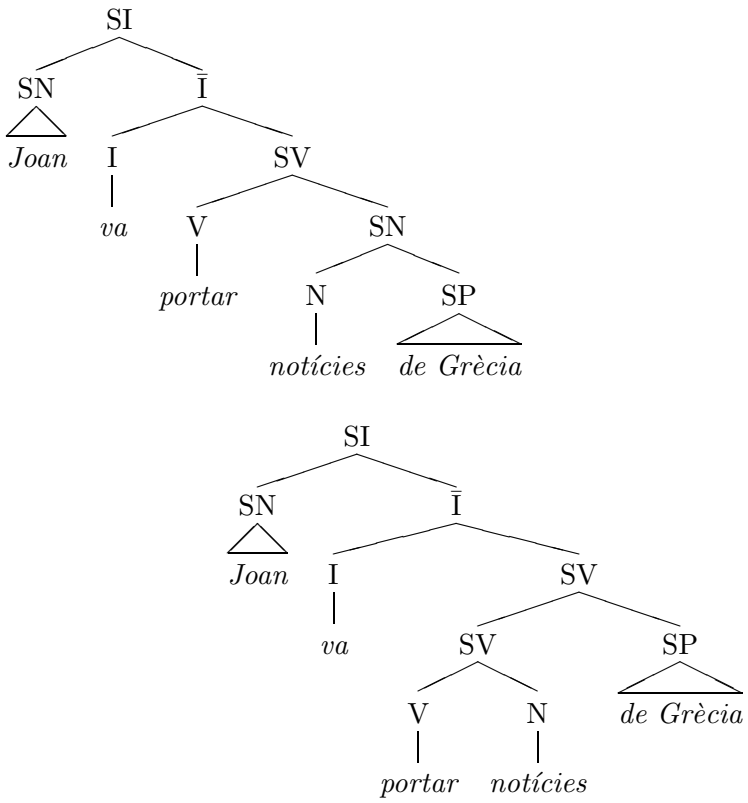


Figura 8.2: Dos arbres per a la frase "Joan va portar notícies de Grècia" (SI = sintagma inflexional, Ī = projecció intermèdia de la inflexió, I = inflexió, SV = sintagma verbal, V = verb, N = nom, SP = sintagma preposicional).

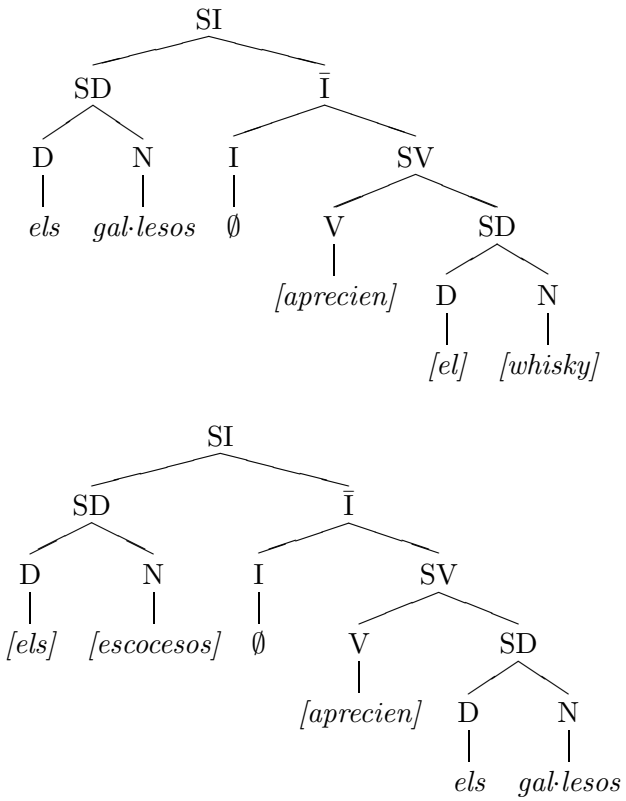


Figura 8.3: Dos arbres per a la segona part de la comparació "Els escocesos aprecien més el whisky que els gal·lesos" (SD =sintagma determinant, D = determinant).

(a) *Els escocesos aprecien el whisky més que els gal·lesos (aprecien el whisky)*

(b) *Els escocesos aprecien més el whisky que (els escocesos aprecien) els gal·lesos*[4]

En aquests dos casos, l'ambigüitat està causada pel fet que són possibles dues estructures sintàctiques per a la segona oració coordinada: en la primera estructura, el sintagma *els gal·lesos* és el subjecte mentre que en la segona estructura és l'objecte (vegeu els arbres de la fig. 8.3).

4. De vegades, l'anàlisi sintàctica d'una oració es complica per la presència de fenòmens de moviment de constituents. Considerem l'oració

[4]De fet, per a evitar aquesta ambigüitat, es considera convenient *però no obligatòria* en català la solució alternativa amb preposició *als gal·lesos* per a la segona interpretació.

(8.19) Qui diu que vindrà?

Aquesta oració té, bàsicament, dues interpretacions. Una és

(8.20) Qui diu que PRO vindrà?

i l'altra

(8.21) * PRO diu que qui vindrà?

És a dir, en la primera, el pronom interrogatiu *qui* és el subjecte de l'oració principal; en la segona, és el subjecte de l'oració subordinada, el qual ha experimentat el *moviment de Qu* (en anglés *Wh-movement*) típic dels mots amb funció interrogativa. En aquest cas, com en l'exemple 8.17, l'el·lipsi permet dos posicionaments diferents del pronom *qui* abans del moviment de Qu, però les ambigüitats causades pel moviment de Qu poden produir-se també sense el·lipsi, com en l'exemple

(8.22) *Com dius que ha explicat que vindria?*

on la posició inicial de l'adverbi interrogatiu *Com* pot resultar de la transformació per moviment de Qu de tres estructures hipotètiques diferents; en cada una d'elles, l'adverbi és adjunt d'un sintagma verbal diferent:

(8.23)

(a) **Dius com que ha explicat que vindria?*

(b) **Dius que ha explicat com que vindria?*

(c) **Dius que ha explicat que vindria com?*

En la primera interpretació es pregunta per la manera de dir-ho, en la segona per la manera d'explicar-ho i en la tercera per la manera de venir.

Es produeixen també moviments similars amb els relatius; per exemple, aquests es mouen cap a fora (cap amunt) des de les subordinades substantives completives amb verbs del tipus de *dir*, *explicar*, etc. En l'oració

(8.24) L'home que vas dir que vindria no ha arribat encara

el primer *que* és un pronom relatiu que fa de subjecte de *vindria* però ha estat mogut fora d'aquella perquè modifique *L'home*.

8.3 Ambigüitats mixtes

Hi ha oracions que són ambigües tant perquè contenen mots ambigus com perquè tenen més d'una estructura sintàctica possible. N'estudiarem dos casos:

1. L'oració conté mots afectats d'ambigüitat lèxica categorial; és a dir, conté un o més mots homògrafs que pertanyen a més d'una *categoria lèxica* (vegeu la pàg. pg:catlex i la nota a peu de la pàg. 40) diferent. Per exemple, el mot *deu* pot voler dir "nou més un" (numeral) o "ha de donar o pagar" (verb). O el mot *cap* pot ser un substantiu ("part superior del cos"), un verb (forma del verb "cabre"), un adjectiu o pronom ("no n'hi ha cap"), o part de la preposició composta "cap a".

   Aquest tipus d'ambigüitat lèxica pot provocar de vegades ambigüitat estructural, causada per la presència de més d'una anàlisi sintàctica acceptable (si, tot i els homògrafs, només n'hi ha una anàlisi acceptable, l'ambigüitat passa desapercebuda per al receptor; això és així perquè habitualment només es consideren estructures acceptables quan es vol assignar interpretació a una oració). Per exemple, la frase anglesa

   (8.25) *Time flies like an arrow*

   vol dir normalment *El temps vola (com una fletxa)* però també són possibles altres dues interpretacions (semànticament destrellatades però sintàcticament impecables): *A les mosques del temps els agrada una fletxa* o *Cronometra les mosques com una fletxa*. Aquesta varietat d'interpretacions es deu al fet que hi ha tres mots en la frase que poden pertànyer a dues categories lèxiques diferents: *time* pot ser verb i substantiu, *flies* pot ser verb i substantiu i *like* pot ser verb i conjunció. De les 8 ($2 \times 2 \times 2$) anàlisis morfològiques possibles de la frase, tres en resulten sintàcticament acceptables, amb interpretacions molt diferents. Aquest tipus d'ambigüitat se sol anomenar *ambigüitat estructural d'origen categorial*. En català —i en castellà— són molt comunes les ambigüitats degudes a la combinació d'un mot que pot ser pronom feble de tercera persona o article (*el*, *la*, *l'*, *els*, *les*) i un altre mot que pot ser substantiu o verb conjugat. Per exemple, l'oració

   (8.26) *La mata el vol*

   pot voler dir dues coses, segons l'elecció de categories lèxiques ("l'acte de volar li provoca la mort" o "la planta sent estima per ell").

2. Un altre tipus d'ambigüitat mixta succeeix quan l'ambigüitat lèxica categorial d'alguns mots es combina amb mecanismes d'el·lipsi com els descrits més amunt per a construccions coordinatives o comparatives. Per exemple, l'oració

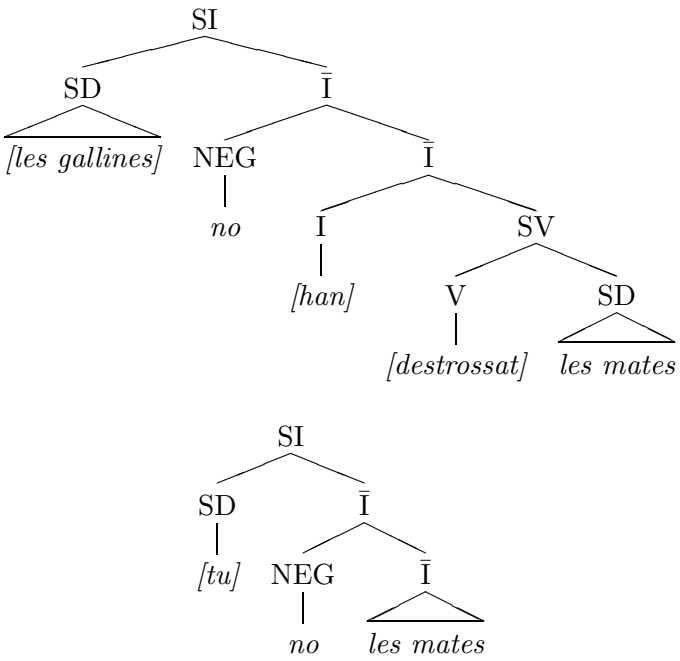


Figura 8.4: Dos arbres per a la segona oració coordinada en "Les gallines han destrossat el sembrat però no les mates" (NEG = negació).

(8.27) *Les gallines han destrossat el sembrat, però no les mates*

té dues interpretacions:

(8.28)

(a) *Les gallines han destrossat el sembrat, però (les gallines) no (han destrossat) les mates.*

(b) *[Les gallines]$_i$ han destrossat el sembrat, però (tu) no les$_i$ mates.*

En (8.28a) *les mates* és un sintagma determinant compost d'un article i un substantiu, que fa d'objecte del verb el·líptic *destrossat*, mentre que en (8.28b) *les mates* és un sintagma verbal compost d'un pronom (*les*) que es refereix a *Les gallines* i un verb (*mates*), sintagma que constitueix un sintagma verbal en la segona oració coordinada (vegeu la fig. 8.4).

8.4 Ambigüitat deguda a l'abast dels quantificadors

Els quantificadors són mots com *algun*, *tot*, *cada*. Quan l'abast d'un quantificador és imprecís, una oració pot tenir més d'una interpretació. Considerem l'exemple

(8.29) *Totes les dones no s'estimen els abrics de pell.*

(Hutchins i Somers 1992), el qual pot tenir dues interpretacions

(8.30)

(a) *No totes les dones s'estimen els abrics de pell.*

(b) *No hi ha cap dona que s'estime els abrics de pell.*

a pesar de no tenir cap ambigüitat lèxica ni estructural aparent. Aquest tipus d'ambigüitat es pot explicar pel fet que el principi de composicionalitat per si sol no és suficient per a especificar completament l'assignació d'interpretació a una oració;[5] la semàntica de les oracions que contenen quantificadors se sol explicar en termes de *formes lògiques* (Radford et al. 1999, p. 357) que contenen d'una banda, *variables* que poden referir-se a un rang d'objectes que cal considerar i, d'altra, a operacions sobre aquestes variables. Doncs bé, en aquestos casos, es pot assignar més d'una forma lògica a una oració.

8.5 Estratègies de resolució de l'ambigüitat

En general, els humans usem les nostres creences sobre el funcionament del món real (o d'un món fictici concret, com en una novel·la) per a elegir una de les interpretacions com a més versemblant (és a dir, per a *resoldre l'ambigüitat*); quan aquestes creences són compartides entre l'emissor i el receptor, es pot usar l'ambigüitat com un mecanisme molt eficient per a produir missatges més curts.

Com hem vist, les causes de l'ambigüitat són molt diverses; per això, també són molt diverses les estratègies de resolució. Aquest epígraf recull unes notes —no exhaustives— sobre les estratègies de resolució d'alguns tipus d'ambigüitat en sistemes automàtics de tractament del llenguatge humà.

[5] En paraules de Radford et al. (1999, p. 364) "hem de reconéixer [l'existència] d'un buit inacceptable entre el que proporciona la sintaxi i el que la semàntica necessita en el cas d'oracions que continguen sintagmes nominals quantificats"

8.5.1 Resolució de l'ambigüitat lèxica categorial

La resolució de l'ambigüitat lèxica categorial, altrament coneguda com etiquetatge de les parts de l'oració (en anglés, *part of speech (PoS) tagging*) està molt ben estudiada. L'ambigüitat es resol normalment usant bé el coneixement lingüístic, bé tècniques estadístiques basades en la freqüència d'aparició en els textos de determinades seqüències curtes (combinacions) de categories lèxiques, o bé les dues aproximacions alhora.

Considerem l'aproximació estadística. Per exemple, el mot castellà *ahorro* pot ser substantiu o verb. Si apareix entre un article i un adjectiu com ara en *el ahorro doméstico* no hi ha cap dubte que es tracta d'un substantiu: la seqüència article–nom–substantiu és molt més freqüent que no la seqüència article–verb–substantiu (el coneixement lingüístic ens permetria descartar la segona possibilitat amb seguretat quasi total). Si prenem un corpus (conjunt) suficientment gran de textos correctes i comptem quantes voltes apareixen totes les seqüències possibles de tres categories lèxiques, podem usar aquestes freqüències per a assignar la categoria d'un mot ambigu: de totes les seqüències de tres mots possibles que es puguen formar amb aquest mot, en prendrem la més freqüent.

8.5.2 Resolució de la polisèmia

La resolució de la polisèmia (en anglés habitualment *word sense disambiguation*) consisteix a assignar a un mot polisèmic, en un text o discurs, una interpretació concreta, diferent de les que podria tenir en altres textos (o contextos). La desambiguació s'efectua usant informació procedent de tres fonts: el *cotext* (intern al text o discurs) i el *context* (extern al text o discurs però relacionat amb ell) i fonts de coneixement addicionals. En traducció automàtica, estem interessats a elegir una de les interpretacions possibles, perquè és comú que els mots polisèmics tinguen diverses traduccions (l'*ambigüitat de transferència* esmentada en l'apartat 8.1).

S'accepta comunament que la major part dels mots polisèmics d'un text (o d'un fragment del text) solen tenir una única interpretació en un text donat, però aquest principi s'ha de concretar en un mètode concret per a resoldre la polisèmia.

La resolució de la polisèmia s'ha abordat des de perspectives molt diverses (vegeu Ide i Veronis (1998)); heus-ne ací dues:

- L'ús de *xarxes semàntiques* on els mots d'un vocabulari se situen en els nodes (nusos) de la xarxa i s'agrupen jeràrquicament en conceptes i superconceptes cada vegada més generals. Es prenen tots els possibles sentits de dos mots d'un text i se'ls assigna el sentit associat al concepte que hi ha en el camí més curt d'un mot a altre. La informació present en diccionaris electrònics preexistents pot servir per a

construir aquestes xarxes o ser usada directament per a la resolució de la polisèmia.

- L'estadística d'aparició conjunta de mots en corpus bilingües de textos pot ajudar a resoldre directament l'ambigüitat de transferència quan es disposa de diccionaris de transferència o quan els textos estan alineats. Per exemple, si en un corpus bilingüe castellà-català l'aparició de *destino* prop d'*incierto* en castellà coincideix amb l'aparició de *destí* en català, podem dir que el mot *destino* té en aquest cas la interpretació de "sort futura"; en canvi, si l'aparició de *destino* prop d'*estación* o *aeropuerto* en castellà coincideix amb l'aparició *destinació* en català, podem elegir el sentit de "punt d'arribada". Aquesta informació podria servir per a traduir després del castellà a l'anglés i elegir *destiny* o *destination* en cada cas amb molta probabilitat d'èxit.

8.5.3 Resolució de l'anàfora

La resolució de l'anàfora —és a dir, la determinació de l'*antecedent* d'un pronom o d'una altra expressió anafòrica— es pot basar en informació morfològica (com ara la concordança de gènere i de nombre entre un pronom i el seu antecedent), sintàctica, o fins i tot semàntica. La informació sintàctica pot ser més rellevant que no ho sembla: si diem

(8.31) *Marta va parlar amb ella*

l'antecedent d'*ella* no pot ser *Marta*. En canvi, si diem

(8.32) *Marta va parlar i ella va caure*

no es pot descartar completament que l'antecedent d'*ella* siga *Marta*. De fet, moltes vegades, només podem recórrer a la semàntica; en l'exemple (ja discutit en la secció 7.1)

(8.33) *[Els soldats]$_i$ van disparar [als xiquets]$_j$. Els$_{i/j?}$ vaig veure caure*

s'ha d'usar informació semàntica per a saber quin és l'antecedent d'*els* en la segona oració (*els soldats* o *els xiquets*).

8.5.4 Resolució de l'ambigüitat estructural

En principi, es podria dir que les persones resolem l'ambigüitat estructural —pura o d'origen categorial— elegint, usant les interpretacions assignades a cada una de les estructures possibles (principi de composicionalitat), quina és la més versemblant en una situació comunicativa determinada. Segons aquest model, les persones consideraríem sempre *totes* les estructures sintàctiques. Es podria argumentar fàcilment en contra dient que en frases complexes (per exemple, amb molts sintagmes preposicionals

com l'exemple 8.16) hi ha massa estructures a considerar. De fet, hi ha experiments psicolingüístics que indiquen que de vegades usem estratègies purament sintàctiques, elegint entre les possibles estructures fins i tot quan no hem sentit o llegit tota l'oració, potser per evitar un esforç intel·lectual excessiu, ja que hi pot haver moltíssimes interpretacions parcials. A canvi, hem de fer l'esforç (pressumiblement més lleuger) de predir una entre les possibles continuacions (sintàctiques) del que hem llegit; segons arriben mots, els anem encaixant en l'estructura predita i usem la sintaxi i la interpretació dels mots per a anar construint a poc a poc la interpretació de l'oració completa. L'experiència ens ajuda a fer prediccions que en general tenen èxit, però de vegades hi ha oracions "enganyoses" que "ens porten a l'hort" (anomenades, per això, en anglés *garden-path sentences*; de fet) ja que en cert punt del procés ens obliguen a descartar la predicció feta i reinterpretar el que havíem llegit fins a aquell punt (l'estudi dels moviments oculars durant la lectura donen pistes molt rellevants sobre l'existència d'aquests processos). Heus ací alguns exemples d'oracions que "ens porten a l'hort", amb una continuació inesperada en les notes a peu de pàgina:

(8.34) *Joan besà Maria i la seua germana...*[6]

(8.35) *Com que Joan sempre corre un parell de quilòmetres...*[7]

(8.36) *En el otro accidente murieron sesenta y cinco...*[8]

(8.37) *The horse raced by the barn...*[9]

Aquests processos de selecció purament sintàctica dónen com a resultat que hi ha certes estructures finals que són preferides a altres, potser perquè simplifiquen la comprensió. Per exemple, si llegim

(8.38) *Va aprendre a afaitar-se en dos minuts*

podríem considerar la interpretació que s'hi parla de la durada de l'afaitat com a més probable que la que interpreta que s'hi parla de la durada de l'aprenentatge, ja que en el segon cas potser hauria estat més natural dir

(8.39) *Va aprendre en dos minuts a afaitar-se*

La regla que afavoreix que els adjunts s'associen a l'últim sintagma que els admeta se sol anomenar regla de *clausura tardana* —angl. *late closure*—; per exemple, aquesta regla afavoreix el primer dels arbres de la figura 8.2. Altra regla que se sol usar és la d'*adjunció mínima* —angl. *minimal attachment*—

[6] ... el va recriminar per haver-ho fet.
[7] ... li semblen poc.
[8] ... resultaron heridos.
[9] ... fell down.

que afavoreix l'arbre sintàctic amb el mínim de nodes (punts de ramificació). Aquestes estratègies són d'utilitat en els sistemes de traducció automàtica per transferència sintàctica pura, ja que no s'hi fa cap processament semàntic.

El punt de vista purament sintàctic és clarament una simplificació excessiva; moltes vegades, les persones resolem l'ambigüitat estructural usant informació semàntica o fins i tot lèxico-semàntica [10] . A tall d'exemple, considereu aquestes dues frases estructuralment idèntiques afectades per ambigüitat d'adjunció:

(8.40) *Porta'm les claus de l'armari gran*

(8.41) *Porta'm les claus de la cadira verda*

En l'oració 8.40, podem dubtar, ja que no sabem si les claus són les que obrin l'armari o les que estan allà guardades. En canvi, en l'oració 8.41 no considerem la primera interpretació (encara que siga la preferida sintàcticament), perquè no és gens versemblant que les cadires tinguen pany (hem usat informació semàntica basada en les nostres creences sobre el món). Si el sistema de traducció automàtica és capaç d'usar informació semàntica, podria elegir correctament en aquest cas.

8.6 Qüestions, exercicis i problemes

1. Indiqueu quina classe d'ambigüitat presenten aquestes frases (justifiqueu molt breument la vostra resposta):

   (a) *Expulsaran l'alcalde de la ciutat* (1: "L'alcalde de la ciutat serà expulsat." 2: "L'alcalde serà expulsat de la ciutat").

   (b) *Hi havia un gat sota l'automòbil* (1: "...perquè acabaven de reparar una roda"; 2: "...i va eixir corrents quan el vaig posar en marxa")

   (c) *Maria va entrar amb una bossa gran. Jo la vaig posar damunt de la taula* (1: "Vaig posar Maria damunt de la taula"; 2: "Vaig posar la bossa damunt de la taula")

   (d) *Què vols, galetes o pa de la tia Pepa?* (Les galetes són també de la tia Pepa?)

   (e) *Posa una mà de paper en la impressora i connecta-la.* (Ha de connectar la mà de paper o la impressora?)

[10] Per exemple, el verb *vendre* admet un objecte directe i un d'indirecte, però el verb *menjar* només el directe, de manera que si diem "Va presentar l'home que venia taronges a Joan" es pot interpretar de dues maneres, però si diem la frase estructuralment idèntica "Va presentar l'home que menjava taronges a Joan" no hi ha més que una interpretació possible.

(f) *Ha aprés a afaitar-se en dos minuts.* (1: "Després de dos minuts, ja sabia afaitar-se"; 2: "Ara ja sap com afaitar-se en dos minuts")

(g) *Us han dit que vaja?* (Qui ha d'anar?).

(h) (cast.) *Vale más que las comas* (1: "...que els signes de puntuació"; 2:"...que les menges")

2. (*) Indiqueu breument almenys dues estratègies diferents que es podrien usar per a resoldre l'ambigüitat sintàctica d'adjunció. Per a inspirar-vos, fixeu-vos en els següents exemples:

   - *Va aprendre en dos minuts a afaitar-se*
   - *Va aprendre a afaitar-se en dos minuts*
   - *Porta'm les claus de l'armari gran*
   - *Porta'm de l'armari gran les claus*
   - *Porta'm les claus de la cadira verda*
   - *Toni comprarà les taronges que ha de vendre a Reme*
   - *Toni comprarà a Reme les taronges que ha de vendre*

3. Si una oració té només una ambigüitat lèxica pura...

   (a) ...té un únic arbre d'anàlisi sintàctica, pero més d'una anàlisi morfològica.

   (b) ...té un únic arbre d'anàlisi sintàctica i una única anàlisi morfològica, però dues interpretacions semàntiques diferents.

   (c) ...té més d'un arbre d'anàlisi sintàctica.

4. La frase "M'agrada més que la bata" pot tenir dues interpretacions; en la primera es parla d'una prenda de vestir; en la segona, d'una preferència a l'hora de preparar, per exemple, una salsa. Indiqueu de quina classe d'ambigüitat es tracta.

   (a) Estructural d'adjunció

   (b) Lèxica categorial

   (c) Estructural d'origen categorial.

5. La frase "baixa i puja amb ascensor" pot voler dir "(baixa) i (puja amb ascensor)" o "(baixa i puja) amb ascensor". De quin tipus d'ambigüitat es tracta?

   (a) Lèxica categorial

   (b) Estructural d'origen categorial

   (c) Estructural d'origen coordinatiu

6. En l'oració *El cotxe s'ha cremat amb el garatge i l'assegurança no el cobreix* no se sap quina de les dues coses està coberta per l'assegurança, el garatge o el cotxe. L'ambigüitat...

   (a) ... es deu a l'el·lipsi.

   (b) ... es deu a l'anàfora.

   (c) ... és estructural d'origen coordinatiu.

7. De quina classe és l'ambigüitat de l'oració "Va vendre les taronges que havia comprat a Maria"?

   (a) Estructural d'origen coordinatiu

   (b) Estructural d'adjunció

   (c) Extrasentencial per anàfora

8. Considereu l'homògraf castellà *vendo* ("Te vendo un coche" "Y yo, ¿para qué quiero un coche vendado?"). Es pot resoldre l'ambigüitat lèxica a què dóna lloc usant només informació sintàctica (és a dir, sobre les categories lèxiques que l'acompanyen en l'oració)?

   (a) No, perquè les dues formes *vendo* s'escriuen exactament igual.

   (b) No, perquè les dues formes *vendo* tenen la mateixa categoria lèxica i la mateixa anàlisi morfològica, tret del lema, i, per tant, poden fer exactament les mateixes funcions sintàctiques.

   (c) Sí, només mirant la categoria lèxica dels mots anteriors i la dels posteriors ja hi ha prou per a saber en quin dels dos casos ens trobem.

8.7 Solucions

1. (a) Ambigüitat sintàctica o estructural (pura) d'adjunció: el sintagma preposicional *de la ciutat* es pot inserir en dues posicions diferents de l'oració.

   (b) Ambigüitat lèxica pura (polisèmia) del mot *gat*.

   (c) Ambigüitat extrasentencial per anàfora: el pronom *la* pot tenir dos antecedents: *Maria* i la *bossa*.

   (d) Ambigüitat sintàctica o estructural (pura) d'origen coordinatiu: el sintagma *de la tia Pepa* pot modificar als dos elements coordinats o només al segon.

   (e) Ambigüitat extrasentencial per anàfora: el pronom *la* pot tenir dos antecedents: *mà [de paper]* i la *impressora*.

(f) Ambigüitat sintàctica o estructural (pura) d'adjunció: el sintagma preposicional *en dos minuts* es pot inserir en dues posicions diferents de l'oració.

(g) Ambigüitat extrasentencial per el·lipsi: el subjecte de *vaja* pot ser *jo*, *ell*, *ella*, etc.

(h) Ambigüitat sintàctica o estructural d'origen categorial deguda al fet que els mots *las* (article o pronom) i *comas* (substantiu o verb) són homògrafs afectats d'ambigüitat lèxica categorial. De les quatre combinacions possibles, dues són sintàcticament acceptables.

2. (*) Vegeu la secció 8.5.4. En el cas de les frases "Toni comprarà...", sembla lògic usar la regla de *clausura tardana*, ja que es correspon prou bé amb les interpretacions preferides per les persones.

3. (b)

4. (c). Els mots *la* i *bata* poden pertànyer cada un a dues categories lèxiques diferents. De les quatre combinacions resultants, dues són sintàcticament acceptables.

5. (c)

6. (b). El pronom feble *el* pot referir-se a *garatge* o a *cotxe*.

7. (b). El sintagma preposicional "a Maria" pot ser l'objecte indirecte de l'oració principal i de la subordinada.

8. (b)

Capítol 9

Tècniques de traducció automàtica

Aquest capítol descriu les tècniques o, dit d'una altra manera, les estratègies bàsiques usades pels programes de traducció automàtica[1], seguint aquest esquema:

- Traducció directa
  - és realment directa en els sistemes reals?
- Traducció indirecta
  - per transferència
    - per transferència morfològica (anomenada *directa*)
    - per transferència sintàctica
    - per transferència semàntica
  - per interlingua
    - interlingua *versus* transferència

9.1 Traducció directa i traducció indirecta

Les estratègies de traducció automàtica es poden dividir en dos grans grups, les *directes* i les *indirectes*. L'estratègia *directa* s'anomena així perquè la traducció d'una frase es produeix directament, sense que es genere una representació intermèdia de la frase; de vegades també se sol anomenar vagament

[1] Jacqmin (1993) també proposa una classificació general dels sistemes de traducció automàtica basada, entre altres criteris, en l'*arquitectura* (estratègia).

traducció *mot per mot*. L'estratègia *indirecta* produeix, a partir de la frase en la llengua d'origen (LO), una representació intermèdia de la frase que després s'usa per a traduir-la. Veurem més endavant de quina naturalesa són aquestes representacions intermèdies.

9.2 Traducció *directa*?

L'estratègia directa es correspon molt bé amb la que s'usa en els programes de traducció automàtica de primera generació (decenni de 1950). Com es veurà, molt freqüentment es qualifiquen de directes estratègies que sí que generen representacions intermèdies molt rudimentàries, i que, per tant, en sentit estricte, haurien de ser classificades com a indirectes. Els sistemes anomenats *directes* de primera generació tradueixen en tres fases (Hutchins i Somers 1992, secció 4.2):

- En la primera fase d'*anàlisi morfològica* rudimentària s'identifiquen les categories lèxiques i es redueixen les paraules a les formes no flexionades.
- En la segona, es consulta el diccionari bilingüe per a traduir les paraules de la frase a la llengua meta (LM).
- Finalment s'apliquen ajustos locals que afecten l'ordre dels mots en la frase (reordenaments locals) i es flexionen les paraules en la LM.

Si demanàrem a una persona no experta que dissenyara un sistema de traducció automàtica, el primer disseny no seria molt diferent del que s'ha descrit. El resultat (Hutchins i Somers 1992, section 4.2) és el que es podria esperar "d'una persona que comptara únicament amb un diccionari bilingüe molt barat i amb un coneixement molt rudimentari de la gramàtica de la llengua meta", amb "errors freqüents de naturalesa lèxica en la traducció i estructures sintàctiques inadequades" que reflecteixen "les estructures pròpies de la llengua d'origen".

Una aproximació a la traducció humana assistida per ordinador (és a dir, semiautomàtica) que està molt relacionada amb la traducció directa és la que s'usa en les anomenades *memòries de traducció* (vegeu el capítol 10).

9.3 Traducció indirecta per transferència

Molts dels sistemes indirectes són sistemes de *transferència*. Un *sistema de transferència* és el que fa les traduccions en tres fases ben diferenciades anomenades *anàlisi*, *transferència* i *generació*:

- La fase d'*anàlisi* produeix, a partir de la frase en la LO, una representació abstracta (RALO). En la RALO s'eliminen tots els detalls de la

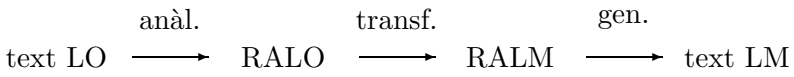


Figura 9.1: Fases d'anàlisi, transferència i generació en un sistema de traducció indirecta per transferència.

frase en LO que no són rellevants per a la traducció i se'n destaquen aquelles característiques i relacions que sí que ho són. Per exemple, podria convenir que les frases angleses "Sam gave a book to Leslie" i "Sam gave Leslie a book" (Arnold et al. 1993) tingueren la mateixa RALO.

- La fase de *transferència* converteix la RALO en una altra representació abstracta similar, però per a la llengua meta (RALM).
- La fase de *generació* (o *síntesi*) produeix la frase en la llengua meta a partir de la RALM.

Aquestes tres fases s'esquematitzen en la figura 9.1. L'arquitectura de transferència és el model estàndard per a la traducció automàtica contemporània, i ho ha estat per molts anys (Arnold et al. 1993).

Els sistemes de transferència es classifiquen segons la naturalesa de les representacions abstractes que utilitzen: es pot parlar, en ordre de profunditat de l'anàlisi, de sistemes de *transferència morfològica*, de *transferència sintàctica* o de *transferència semàntica*. L'elecció de la profunditat de l'anàlisi s'ha de fonamentar en la naturalesa i la profunditat de les divergències de traducció (Vandooren 1993) entre les llengües implicades.

9.3.1 Sistemes de transferència morfològica

Els sistemes de transferència morfològica no són molt diferents dels "directes" descrits més amunt: la fase d'*anàlisi* analitza morfològicament els mots de la frase però no identifica les relacions (sintàctiques) entre ells. La secció 9.3.2 dóna més detalls sobre l'anàlisi i la generació morfològiques.

De vegades l'anàlisi morfològica pot ser més difícil del que sembla, com per exemple, en el cas de la morfologia verbal de les llengües romàniques. I pot complicar-se més. Per exemple, fixeu-vos en l'imperatiu castellà *demos*: si va seguit del pronom enclític *le*, forma amb aquest un únic mot i rep un nou accent ortogràfic: *démosle*; si el pronom és *nos*, a més es perd una consonant: *démonos*; amb dos pronoms, pot ser *démonoslos*, etc. Altres vegades es

dóna el problema de l'ambigüitat lèxica categorial (vegeu l'apartat 8.1): és a dir, el mot pot pertànyer a dues categories lèxiques diferents i cal usar informació sobre les categories morfològiques dels mots anteriors i posteriors (en absència d'informació sintàctica) per a desfer l'ambigüitat.

La fase de *transferència* pot consistir en un reordenament local d'algunes seqüències de mots (per exemple, quan es tradueix de l'anglés al català, els parells adjectiu–substantiu es podrien reordenar a substantiu–adjectiu) i en la conversió de les formes lèxiques de la LO en les corresponents de la LM (mitjançant l'ús d'un diccionari bilingüe).

La fase de *generació* podria efectuar la substitució de les formes lèxiques de la LM per les corresponents formes superficials. Altres autors (Hutchins i Somers 1992, secció 6.4) situen els reordenaments en la fase de generació.

Com que els sistemes de transferència morfològica no identifiquen les relacions sintàctiques entre els mots de la frase en la llengua d'origen, per a fer els reordenaments han d'identificar les seqüències de mots que necessiten ser reordenats. La capacitat d'un sistema de transferència morfològica per a produir traduccions acceptables dependrà de la seua capacitat per a detectar seqüències de mots que es corresponguen amb els sintagmes que necessiten ser reordenats. Imaginem que volem traduir de l'anglés al català i hem decidit que s'han d'usar aquestes regles de reordenament:

$$R_1 \;\; adj\ subst \rightarrow subst\ adj$$

$$R_2 \;\; subst_1\ subst_2 \rightarrow subst_2\ prep\_de\ subst_1$$

Per exemple, la regla R_1 reordenaria "tall driver" en "conductor alt" i la regla R_2 reordenaria "truck driver" en "conductor de camió".

Ara, pensem què li succeiria a "tall truck driver". Si s'aplica primer la regla R_1 a "tall truck" ja no podem aplicar-hi la R_2. Si s'hi aplica primer la R_2 i després la R_1, s'obté la traducció correcta: "conductor alt de camió". Quan tenim més d'una regla, no sabem en quin ordre cal aplicar-les-hi. Si tenim "tall gasoline truck driver" ("conductor alt de camió de gasolina"), no hi ha cap ordre d'aplicació de R_2 i R_1 que done una traducció acceptable. Això suggereix la necessitat d'una nova regla que detecte i reordene el patró llarg adjectiu–substantiu–substantiu–substantiu, per exemple:

$$R_3 \;\; adj\ subst_1\ subst_2\ subst_3 \rightarrow subst_3\ adj\ prep\_de\ subst_2\ prep\_de\ subst_1$$

Aquesta regla podria reordenar correctament aquesta seqüència de quatre mots. Com es pot veure, les regles de reordenament intenten descobrir unitats sintàctiques (sintagmes) usant un nombre limitat de patrons que representen les seqüències de mots que poden formar aquestes unitats; el problema és que els sintagmes poden ser, en principi, indefinidament llargs[2],

[2] En el paradigma generatiu, si la mateixa regla de generació es pot aplicar repetidament a un sintagma, aquest sintagma es pot allargar indefinidament. Un exemple clàssic d'això

i el conjunt de regles de reordenament ha de ser forçosament limitat. Queda, a més, per determinar, en els casos en què es pot aplicar més d'una regla, quina s'hi ha d'aplicar abans[3].

La identificació de patrons de categories morfològiques que es corresponguen amb els sintagmes més freqüents pot servir, a més de per a fer reordenaments, per a resoldre la concordança de nombre i gènere. Per exemple, si usem el patró

subst adj

per a identificar una classe de sintagmes nominals senzills, podem fer que la traducció correcta al català del sintagma nominal castellà *postre buenísimo* siga *postres boníssimes*, ja que el gènere i el nombre de l'adjectiu que modifica a un substantiu ha de concordar-hi i el substantiu castellà *postre* (masculí singular) es correspon amb el substantiu català *postres* (femení plural). Com que una vegada identificada la classe de sintagma queda clar que el nucli és el primer element (el substantiu), ja es pot propagar el gènere i el nombre del primer element al segon (l'adjectiu).

Les figures 9.2, 9.3 i 9.4 il·lustren el funcionament de les fases d'anàlisi, transferència i generació, respectivament, d'un sistema de transferència morfològica avançada (és a dir, amb reconeixement de patrons senzills que representen sintagmes) de l'anglés al català.

9.3.2 Anàlisi i generació morfològiques

L'*anàlisi morfològica* és el procés que determina, per a cada mot d'un text (forma *superficial* o *flexionada*) una o diverses *formes lèxiques* , consistents en una *forma canònica* o *lema* i informació sobre la categoria lèxica del mot i la flexió (vegeu la secció 8.1). La *generació* morfològica fa l'operació inversa. Per exemple, l'anàlisi morfològica de la forma superficial *vius* (ambigua) donaria com a resultat dues formes lèxiques: `viure<verb><pres. ind.><2a pers.><sing.>` i `viu<adj.><masc.><pl.>`, on els lemes són, respectivament, *viure* i *viu*.

Un analitzador morfològic, per tant, ha de tenir la informació següent sobre la llengua dels textos que s'han d'analitzar: el vocabulari o conjunt de lemes, els paradigmes de flexió, i la correspondència entre lemes i paradigmes.

---

el donen les oracions adjectives de relatiu; la sèrie de sintagmes nominals "el cotxe", "el cotxe que va dur l'home", "el cotxe que va dur l'home que va vindre del poble", "el cotxe que va dur l'home que va vindre del poble que vam visitar durant el viatge", etc., demostra que no hi ha límits a la longitud d'un sintagma (nominal, en aquest cas).

[3] Una tècnica observada en alguns programes és la següent: (a) els reordenaments es van aplicant segons es recorre la frase d'esquerra a dreta; (b) els reordenaments de seqüències més llargues tenen prioritat, i (c) els mots afectats per un reordenament no tornen a estar involucrats en cap altre reordenament. Així només es visita una vegada cada mot de la frase.

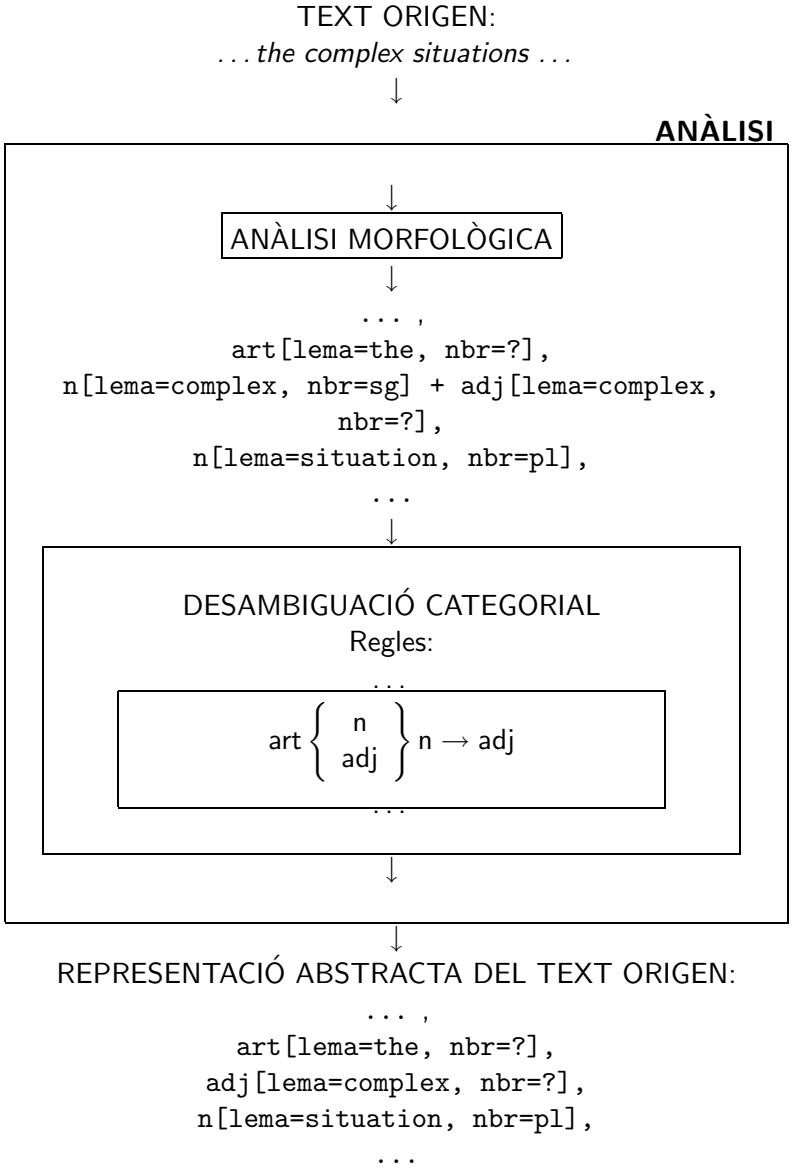


Figura 9.2: Fase d'anàlisi d'un sistema senzill de transferència morfològica avançada

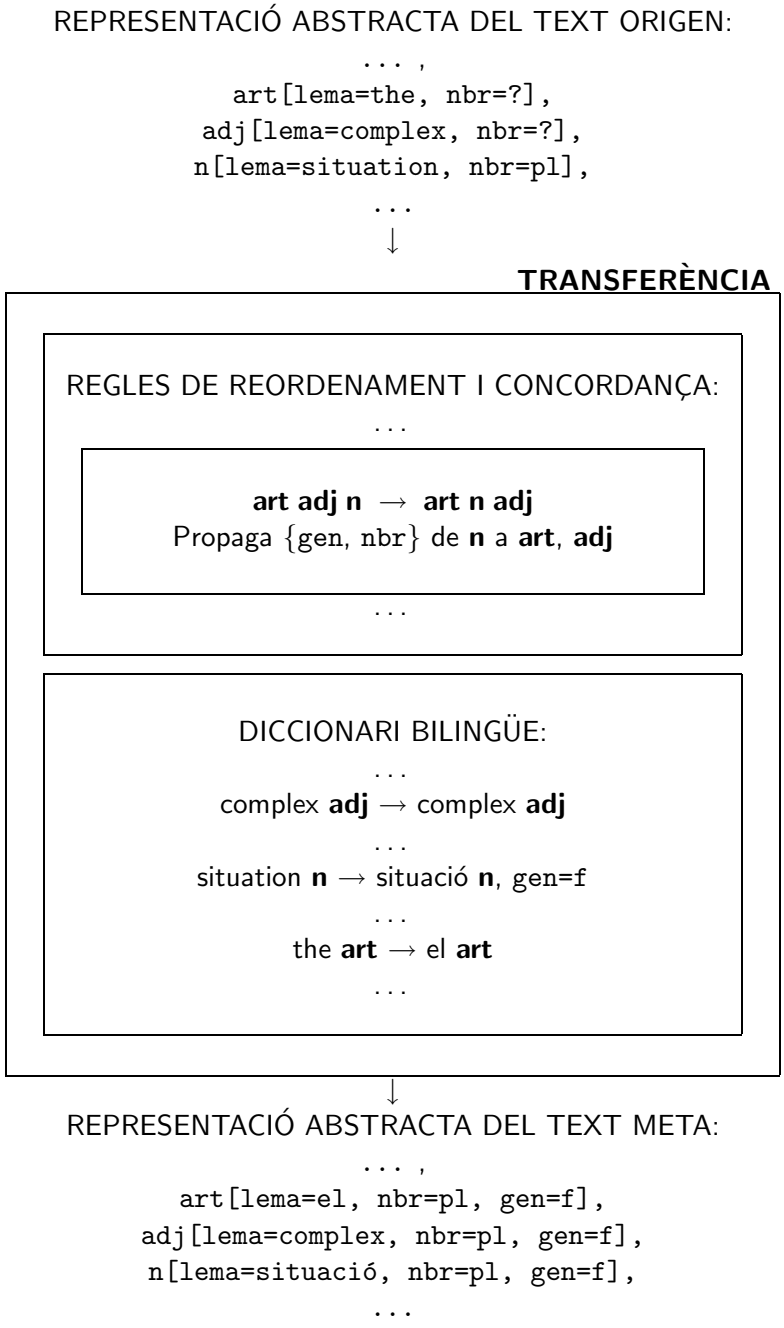


Figura 9.3: Fase de transferència d'un sistema senzill de transferència morfològica avançada, amb regles de reordenament i de concordança

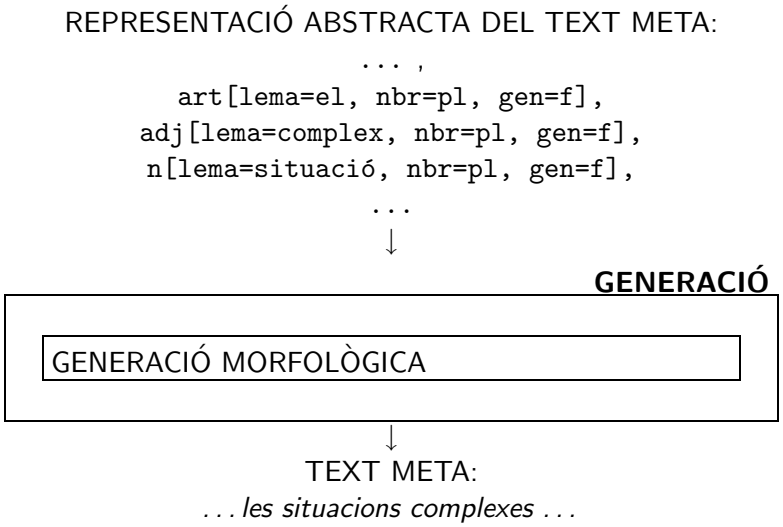


Figura 9.4: Fase de generació d'un sistema senzill de transferència morfològica avançada

Quan els mecanismes de flexió de les llengües són, com en la major part de les llengües indoeuropees, per modificació de les terminacions (*desinències*) dels mots, un mètode atractiu consisteix a processar la forma superficial lletra a lletra d'esquerra a dreta i produir la forma lèxica incrementalment, afegint a cada pas més informació. S'assumeix que les primeres lletres del mot en són l'*arrel* i, per tant, determinen el lema, i que les últimes lletres determinen la forma gramatical. Imaginem que seguim aquest mètode amb el mot *angoixaven*:

- Quan només hem vist *ang-* hi ha encara moltes possibilitats: pot ser, entre altres, qualsevol forma dels mots *angina*, *angle*, *anglés*, *angoixa*, *angoixar*, *angost*, *anguila* i *angula*. Com a màxim, podem dir que el lema comença per `ang`.
- Quan hem llegit *ango-* el ventall de possibilitats es fa més estret: pot ser una forma d'*angoixa*, d'*angoixar* o d'*angost*. Ja podem dir que el lema comença per `ango`.
- Quan hem llegit *angoi-* ja sabem que el lema és *angoixar* o *angoixa*, és a dir, que comença per `angoixa`.
- Llegir *angoix-* o *angoixa-* no ens permet determinar amb seguretat més informació sobre la forma lèxica; en el cas d'*angoixa-* es poden

descartar algunes formes del verb angoixar com *angoixem* o *angoixí*, però encara en queden moltes. Encara pot ser nom o verb.

- Quan hem vist *angoixav-* ja podem dir que el lema és *angoixar*, que es tracta d'un verb, i que estem, amb tota seguretat, davant d'una forma de l'imperfet d'indicatiu. L'analitzador ens pot dir ja `angoixar<verb>-<impind>`.

- Després de veure *angoixave-* encara no sabem la persona del verb (pot ser la segona del singular o la tercera del plural).

- Finalment, quan veiem *angoixaven* ja sabem que és la tercera del plural. L'analitzador produeix: `angoixar<verb><impind><3><pl>`.

El procés es pot resumir en l'alineament següent:

```
a n g o i   x a v               e n
a n g o ixa - - r<verb><impind> - <3><pl>
```

Si la forma superficial haguera estat *angostes*, tot el procés fins a *ango-* hauria estat idèntic. En el cas d'*angoixem*, el procés hauria estat idèntic fins a *angoixe-*:

```
a n g o i   x e m
a n g o ixa - - r<verb><presind><1><pl>
              r<verb><pressub><1><pl>
              r<verb><imper><1><pl>
```

En aquest cas, el mot té tres anàlisis morfològiques diferents. Per altra banda, la part final del processament d'*angoixaven* i de *cantaven* seria molt similar, ja que els dos verbs es conjuguen segons el mateix paradigma.

Tot això permet representar l'analitzador morfològic com el que els matemàtics anomenen *graf dirigit acíclic* (GDA). Un *graf dirigit* té dues parts: un conjunt de nodes, nusos o *vèrtexs* (representats gràficament com a punts o cercles) i un conjunt de fletxes cada una de les quals va d'un vèrtex a altre. El graf és *acíclic* si, seguint les fletxes, no es pot passar dues voltes pel mateix *vèrtex*. El GDA d'un analitzador morfològic té un vèrtex inicial únic (del qual només poden eixir fletxes) que representa l'estat inicial de l'analitzador abans de començar a llegir un mot. Els altres vèrtexs representen l'estat de l'analitzador després d'haver llegit una o més lletres. Les fletxes que ixen d'un vèrtex qualsevol del graf tenen dues etiquetes separades per ":". La primera (d'entrada) indica la lletra que es llig; la segona (d'eixida) indica el que s'ha de produir (zero, un o més símbols). Els estats finals (representats amb dos cercles concèntrics) indiquen estats en els quals l'analitzador determina que ha llegit una forma superficial completa si no queden més lletres a llegir; en alguns estats finals s'escriuen símbols addicionals per a completar l'anàlisi i, en el cas d'un homògraf (ambigüitat lèxica), totes les

opcions. L'analitzador llig la forma superficial lletra per lletra, va d'estat en estat, i va produint la forma lèxica, fins que llig tot el mot; si arriba a un estat final, accepta el mot i en retorna la forma lèxica. En la figura 9.5 es pot veure una part de l'analitzador morfològic, corresponent als mots que comencen per *ang-*. En informàtica teòrica, les màquines idealitzades que representen aquesta classe de GDA s'anomenen *transductors d'estats finits p-subseqüencials sense cicles*, on p indica el nombre de possibles opcions en els estats finals. Els *generadors morfològics* es poden organitzar de forma molt similar: lligen símbol a símbol la forma lèxica i produeixen la forma superficial.

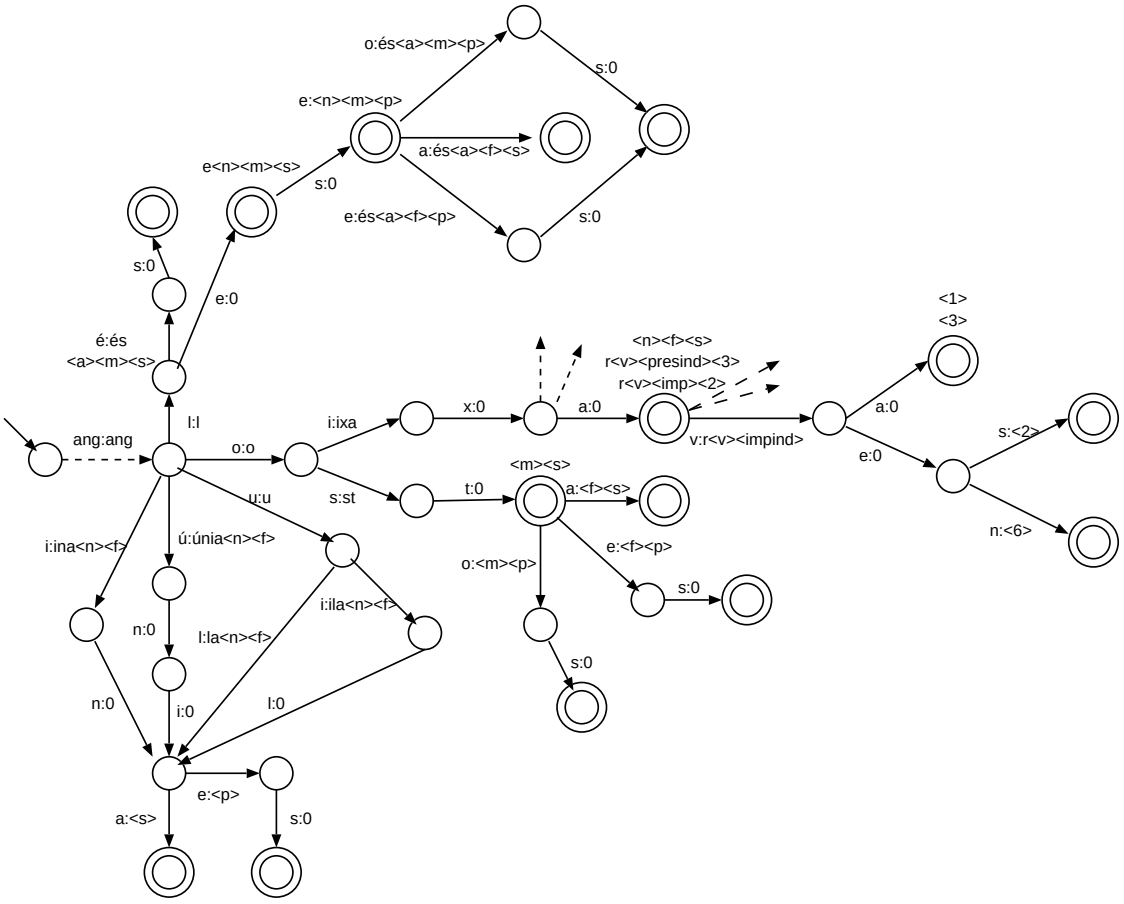


Figura 9.5: Detall de l'analitzador morfològic. L'estat inicial, indicat amb una fletxa entrant, és el situat més a l'esquerra en el graf.

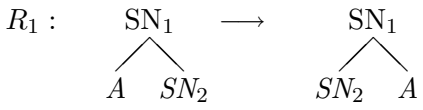
9.3.3 Sistemes de transferència sintàctica

En aquests sistemes, la representació abstracta (RALO) que s'obté en l'anàlisi inclou un arbre d'anàlisi sintàctica de la frase en la LO (o una entitat

equivalent), que descriu les relacions sintàctiques existents entre les parts de la frase, a més de la informació morfològica necessària per a fer-ne la traducció. És a dir, es fa una anàlisi morfològica i una anàlisi sintàctica (anglés *parsing*) de la frase en LO. L'anàlisi sintàctica s'explica amb més detall en l'apartat 9.3.4. En la fase de transferència, s'apliquen regles que transformen la representació sintàctica de la frase d'entrada (la RALO) en una representació sintàctica de la traducció (la RALM). Finalment, la fase de generació transforma aquesta representació en la frase en LM.

L'estratègia de transferència sintàctica resol una bona part dels problemes dels sistemes directes i dels de transferència morfològica, ja que és capaç de determinar l'extensió i l'estructura de cada un dels sintagmes de la frase en LO i manipular cada sintagma com una unitat, independentment de l'estructura o de la longitud. De fet, com ja s'ha dit, els sintagmes *tenen estructura*; és a dir, les relacions entre els elements d'un sintagma no són purament lineals, sinó jeràrquiques: els sintagmes estan fets de sintagmes. Els sistemes de transferència morfològica no poden tenir en compte aquesta estructura interna, i, com a resultat, necessiten moltíssimes regles per a reordenar adequadament els mots de les frases[4].

Imaginem el sintagma nominal següent en anglés: "The old professor's smart girlfriend's red car"; el sintagma és massa llarg per a la major part dels programes de transferència morfològica, perquè involucra una seqüència massa llarga de reordenament. Si, en canvi, tenim un sistema de TA per transferència sintàctica i suposem que la gramàtica conté les regles següents[5]: $SN \rightarrow SG\ SN$, $SN \rightarrow A\ SN$, $SN \rightarrow N$ i $SG \rightarrow SN\ GS$, on SN és un sintagma nominal, SG un "sintagma de genitiu", A un adjectiu, N un substantiu i GS la partícula de genitiu saxó (*'s*, *'*), l'estructura d'aquest sintagma nominal es podria representar, usant un arbre d'anàlisi sintàctica (sense tenir en compte els determinants, per a simplificar) com es veu en la fig. 9.6. Un sistema de transferència sintàctica de l'anglés al català podria usar dues regles per a transformar subarbres (parts de l'arbre), una per a moure els adjectius:[6]



[4]Els sistemes de transferència morfològica per reordenament de patrons assumeixen que una oració és una seqüència lineal de sintagmes d'estructura lineal; aquest model de la sintaxi d'una oració pot ser molt limitat en moltes aplicacions de traducció automàtica.

[5]La sintaxi generativa actual (vegeu Chomsky (1996)) prediu molts aspectes de les llengües naturals postulant l'existència d'un conjunt de regles universals molt senzilles o *principis* amb variacions que en cada llengua estan determinades per *paràmetres*; la gramàtica que es considera en aquesta discussió no és, en principi, la postulada per aquesta formulació, sinó una adequada per a la tasca concreta.

[6]Evidentment, no sempre s'han de moure els adjectius; la traducció correcta de "the last car" és "l'últim cotxe", sense canviar l'ordre. Un sistema real de transferència sintàctica hauria de considerar aquests casos de manera especial.

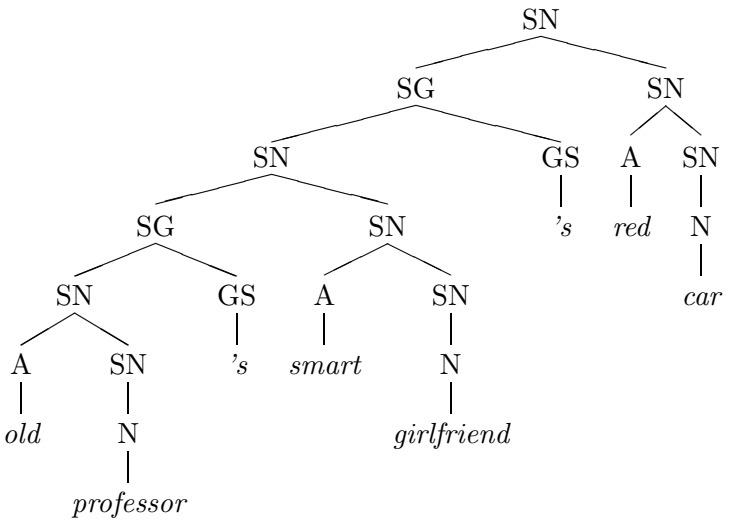
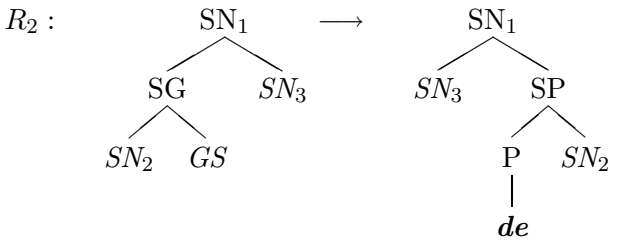


Figura 9.6: Arbre d'anàlisi sintàctica del sintagma nominal "The old professor's smart girlfriend's red car"

i una altra per a reordenar els sintagmes nominals que contenen un genitiu saxó:



L'aplicació de la regla R_2 al SN en l'arrel de l'arbre d'anàlisi sintàctica de la frase dóna com a resultat l'arbre que es veu en la figura 9.7. Després caldria aplicar la regla R_1 al SN que genera "red car", després la regla R_2 al SN que genera "old professor's smart girlfriend", etc. El resultat final es mostra en la fig. 9.8 i es correspon amb l'arbre d'anàlisi sintàctica de la traducció, "El cotxe roig de l'amiga intel·ligent del professor vell", la qual es podria generar directament a partir de l'arbre.

Quan es tradueix del castellà al català, poques vegades es donen situacions com aquesta, ja que, en general, l'ordre dels mots no sol canviar tan radicalment; una construcció castellana que sí que podria requerir la identificació i el desplaçament d'un sintagma nominal complet per a traduir-lo al català seria la construcció de relatiu possessiu amb *cuyo*, ja que el català no en té. Una possible solució usa frases preposicionals del tipus de *del qual*

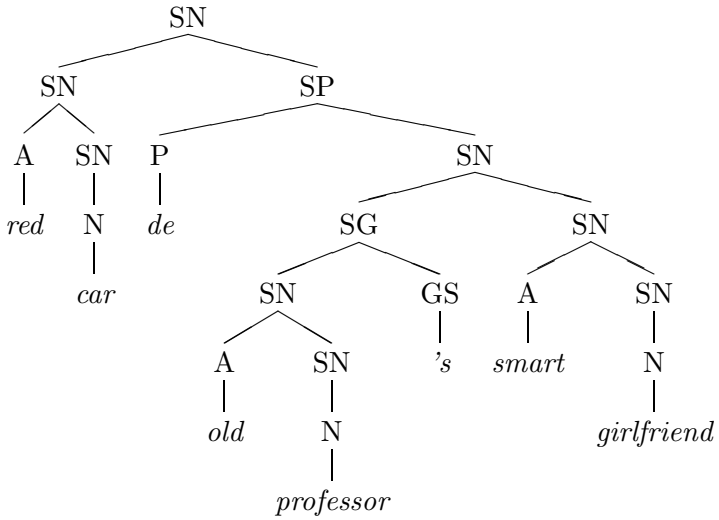


Figura 9.7: Arbre d'anàlisi sintàctica del sintagma "The old professor's smart girlfriend's red car" després d'aplicar-hi la regla R_2 al SN principal o arrel (vegeu el text i la figura 9.6).

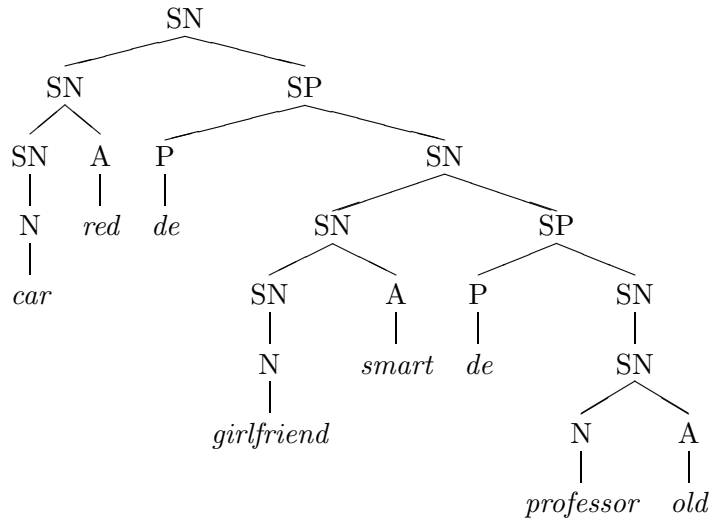


Figura 9.8: Arbre d'anàlisi sintàctica del sintagma "The old professor's smart girlfriend's red car" després d'haver-hi fet tots els reordenaments possibles amb les regles R_1 i R_2 (vegeu el text i les figures 9.6 i 9.7).

postposades a la traducció del sintagma nominal que segueix a *cuyo*. Fixeu-vos en aquests dos exemples, on s'ha fet una anàlisi sintàctica parcial[7] per a marcar els sintagmes nominals amb claudàtors:

$$\begin{array}{rl} & [_{SN} \text{ las hijas }][_{SN} \text{ cuyo } [_{SN} \text{ padre }]] \\ \rightarrow & [_{SN} \text{ les filles }][_{SN} \text{ el pare } [_{SP} \text{ de les quals }]]\ldots]] \end{array}$$

$$\begin{array}{rl} & [_{SN} \text{ la comunidad }][_{SN} \text{ cuyas } [_{SN} \text{ señas de identidad básicas }]] \\ \rightarrow & [_{SN} \text{ la comunitat }][_{SN} \text{ les senyes d'identitat bàsiques } [_{SP} \text{ de la qual }]]\ldots]] \end{array}$$

El sintagma nominal que segueix a *cuyo* (i hi concorda en gènere i nombre) pot ser, en principi, indefinidament llarg; cal identificar-lo correctament, moure'l com una unitat i afegir-li la frase relativa *del qual* de manera que concorde ara (en gènere i nombre) amb l'antecedent (el sintagma nominal anterior al *cuyo*). Aquestes operacions es poden resoldre de manera molt natural en un sistema de transferència sintàctica.

9.3.4 Anàlisi sintàctica

L'anàlisi sintàctica pressuposa l'existència d'una *gramàtica* de la LO, és a dir, d'un conjunt de regles que descriuen com es construeixen, sintagma per sintagma, totes les frases vàlides de la LO:[8] l'analitzador sintàctic actua després de l'analitzador morfològic, i obté l'*arbre d'anàlisi sintàctica* (o els arbres, si n'hi ha més d'un) a partir de la seqüència de categories lèxiques de la frase; cada arbre indica una possible combinació i ordre d'aplicació de regles que dóna lloc a la frase en qüestió.[9] Els algorismes d'anàlisi sintàctica poden ser *ascendents* (anglés *bottom-up*) quan construeixen l'arbre començant a partir de les fulles —les categories lèxiques de cada mot— anant cap a l'arrel —el qual correspon al sintagma complet—, o *descendents* (anglés *top-down*), en cas contrari.[10] Si la frase és estructuralment ambigua (vegeu l'apartat 8.2), té més d'un arbre d'anàlisi sintàctica: alguns analitzadors produeixen tots els arbres possibles; d'altres, n'elegeixen un (usant alguna estratègia de desambiguació sintàctica com les descrites en 8.5.4).

Perquè siga pràctic, l'algorisme d'anàlisi sintàctica ha de ser ràpid i eficient; per exemple, és convenient que funcione de manera que puga construir progressivament els arbres llegint l'oració d'esquerra a dreta, ja que així (a) pot començar a treballar abans que l'analitzador morfològic haja analitzat

[7] És a dir, sense construir l'arbre complet.

[8] Escriure una gramàtica que cobrisca completament totes les possibles frases sintàcticament correctes d'una llengua és una tasca que està molt lluny de ser trivial.

[9] Les regles són, en el fons, els subarbres bàsics amb els quals es construeixen els arbres d'anàlisi sintàctica de totes les frases sintàcticament acceptables; és a dir, aquestes regles especifiquen com es pot construir un sintagma o *constituent* a partir d'altres sintagmes i de categories lèxiques.

[10] Recordeu que els arbres d'anàlisi sintàctica estan "cap per avall": l'arrel és a dalt i les fulles, a baix.

tota l'oració i (b) pot proveïr el mòdul de transferència amb anàlisis parcials que li poden servir per a anar preparant la traducció parcial de les parts ja analitzades.

Exemple d'anàlisi ascendent

Com a exemple, descriurem un tipus bastant estés d'analitzador ascendent, anomenat usualment GLR (*generalized LR* o LR generalitzat, on LR és l'abreviació de *left-to-right, rightmost derivation*, "d'esquerra a dreta i amb derivació per la dreta"). Els analitzadors GLR lligen les categories lèxiques d'esquerra a dreta, les *desplacen* a un tipus de memòria especial anomenat pila (és a dir, les hi *apilen*) i quan el cim de la pila conté elements que segons la gramàtica es poden agrupar en un subarbre, els desapila, els *redueix* a un subarbre, i deixa el subarbre en el cim de la pila. Per a saber quan ha de desplaçar una categoria lèxica a la pila o quan i com ha de reduir el cim de la pila, té en compte una o més categories lèxiques de les que està a punt de llegir i un o més elements del cim de la pila, i fa l'acció que li indica una *taula d'anàlisi sintàctica* que es construeix a partir de la gramàtica que s'haja proposat per a la llengua origen i que el analitzador consulta en cada pas de l'anàlisi.[11]

Un exemple servirà per a il·lustrar tots aquests conceptes. Imaginem la següent gramàtica simplificada que accepta un bon nombre d'oracions simples en català:

$$
\begin{array}{rcl}
O & \longrightarrow & SN\ SV \\
SN & \longrightarrow & \textbf{det}\ \bar{N} \\
SN & \longrightarrow & \bar{N} \\
SN & \longrightarrow & SN\ SP \\
\bar{N} & \longrightarrow & \textbf{n adj} \\
\bar{N} & \longrightarrow & \textbf{n} \\
SV & \longrightarrow & \textbf{v} \\
SV & \longrightarrow & \textbf{v}\ SN \\
SV & \longrightarrow & SV\ SP \\
SP & \longrightarrow & \textbf{prep}\ SN
\end{array}
$$

La gramàtica es ambigua, és a dir, capaç de generar dos arbres d'anàlisi sintàctica per a oracions com ara "L'home porta la clau de l'armari gran".

Usant un algorisme estàndard que no ve al cas detallar ací, la gramàtica es transforma en la taula d'anàlisi sintàctica corresponent (taula 9.1), que indica què cal fer en cada pas de l'anàlisi. En la taula, el símbol • indica tant el principi com el final de la oració (l'anàlisi d'una oració comença apilant • en la pila). Quan la situació a la qual s'arriba no està prevista en la taula, és perquè l'oració no és correcta d'acord amb la gramàtica donada; aquesta

[11] L'ús de la taula permet construir el programa analitzador independentment de la gramàtica concreta: si canvia la gramàtica, només canvia la taula.

situació d'error s'ha de resoldre de manera que l'anàlisi puga continuar, encara que el resultat siga una anàlisi parcial, ja que ens interessa produir una traducció aproximada; el tractament de les situacions d'error és complex i cau fora de l'abast d'aquest llibre. Quan en una situació hi ha més d'una acció possible —circumstància que pot ser deguda, com en l'exemple, al fet que la gramàtica és ambigua— es pot fer una d'aquestes dues coses: elegir sempre una acció fixa o bé "duplicar" l'analitzador de manera que cada còpia continue l'anàlisi per cada un dels camins.

Si el cim de la pila és...	I hi ha a la vista...	L'acció pertinent és...
•	**det** o **n**	apilar-lo
• $[_O \cdots]$	•	anàlisi finalitzada
• $[_{SN} \cdots]$	**v** o **prep**	apilar-lo
$\cdots$ **det** $[_{\bar{N}} \cdots]$	**v**, **prep** o •	reduir a $[_{SN}$ **det** $[_{\bar{N}} \cdots]]$
$\cdots [_{\bar{N}} \cdots]$ (altres)	**v**, **prep** o •	reduir a $[_{SN} [_{\bar{N}} \cdots]]$
$\cdots$ **det**	**n**	apilar-lo
$\cdots$ **n**	**adj**	apilar-lo
	v, **prep** o •	reduir a $[_N$ **n**$]$
•$[_{SN} \cdots][_{SV} \cdots]$	•	reduir a $[_O [_{SN} \ldots][_{SV} \ldots]]$
	prep	apilar-la
$\cdots$ $[_{SN} \cdots]$ $[_{SP} \cdots]$	**v**, **prep** o •	reduir a $[_{SN} [_{SN} \cdots][_{SP} \cdots]]$
$\cdots$ **v**	**prep** o •	reduir a $[_{SV}$ **v**$]$
	det o **n**	apilar-lo
$\cdots$ **prep**	**det** o **n**	apilar-lo
$\cdots$ **n adj**	**v**, **prep** o •	reduir a $[_{\bar{N}}$ **n adj**$]$
$\cdots$ $[_{SV} \cdots]$ $[_{SP} \cdots]$	**prep** o •	reduir a $[_{SV} [_{SV} \cdots][_{SP} \cdots]]$
$\cdots$ **v** $[_{SN} \cdots]$	•	reduir a $[_{SV}$ **v**$[_{SN} \cdots]]$
	prep	CONFLICTE: reduir a $[_{SV}$ **v**$[_{SN} \cdots]]$ o apilar-la
$\cdots$ **prep** $[_{SN} \cdots]$	**prep**, **v** o •	reduir a $[_{SP}$ **prep**$[_{SN} \cdots]]$

Taula 9.1: Taula d'anàlisi sintàctica per a les oracions senzilles descrites per la gramàtica de la p. pg:gramsenz. Les situacions no previstes indiquen que l'oració d'entrada no és acceptada per la gramàtica i donen lloc a una situació d'error.

Vegem com es faria l'anàlisi de l'oració (no ambigua)

(9.1) L'home porta la clau.

D'aquesta oració, l'analitzador sintàctic, només en veu la seqüència de categories lèxiques:

(9.2) **det n v det n** •

L'anàlisi, pas a pas, es mostra en la taula 9.2.

Pila	Entrada restant...	Acció
•	**det n v det n** •	apilar **det**
• **det**	**n v det n** •	apilar **n**
• **det n**	**v det n** •	reduir a $[_{\bar{N}}\mathbf{n}]$
• **det** $[_{\bar{N}}\mathbf{n}]$	**v det n** •	reduir a $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$	**v det n** •	apilar **v**
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v**	**det n** •	apilar **det**
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v det**	**n** •	apilar **n**
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v det n**	•	reduir a $[_{\bar{N}}\mathbf{n}]$
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v det** $[_{\bar{N}}\mathbf{n}]$	•	reduir a $[_{SN}\mathbf{det}[_{\bar{N}}\ldots]]$
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v** $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]]$	•	reduir a $[_{SV}\mathbf{v}[_{SN}\ldots]]]$
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ $[_{SV}\mathbf{v}[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]]$	•	reduir a $[_{O}[_{SN}\ldots][_{SV}\ldots]]$
• $[_{O}[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]][_{SV}\mathbf{v}[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]]]$	•	acceptar

Taula 9.2: Anàlisi sintàctica de la seqüència de categories lèxiques **det n v det n** • segons les accions especificades en la taula 9.1.

9.3.5 Sistemes de transferència semàntica

Els sistemes de traducció automàtica basats en la transferència sintàctica solen funcionar bé en casos senzills, però en general es fa necessària una anàlisi més profunda (Hovy 1993), ja que la correspondència entre les relacions sintàctiques (subjecte, objecte, etc.) i les relacions semàntiques (agent, pacient, etc.) dels constituents d'una frase poden variar d'una llengua a una altra. Per exemple, en la frase catalana "m'agraden els llimons", qui produeix el plaer ("els llimons") fa de subjecte en l'oració, mentre que en l'equivalent anglesa ("I like lemons") hi fa d'objecte directe. És cert que aquest cas es podria tractar de manera particular en un sistema de transferència sintàctica, fent la transformació corresponent en l'arbre d'anàlisi sintàctica quan el verb fos "agradar", però hi ha casos en què també convé oblidar l'estructura sintàctica concreta de la frase en LO i fixar-se més aviat en la semàntica, com per exemple quan cal resoldre ambigüitats causades per l'anàfora o a l'el·lipsi (vegeu les pàgines pg:anafora i pg:ellipsi). Per exemple, les dues frases angleses esmentades més amunt "Sam gave a book to Leslie" i "Sam gave Leslie a book" tenen una sintaxi diferent però volen dir exactament el mateix: "Sam va donar un llibre a Leslie"; el que més compta és qui fa l'acció, quin objecte afecta i qui n'és el destinatari, però no l'ordre en què aquestes entitats apareixen en la frase en LO (Arnold et al. 1993). Els sistemes de transferència semàntica construeixen representacions intermèdies més profundes; l'anàlisi i la generació són més complexes, però la transferència se simplifica.

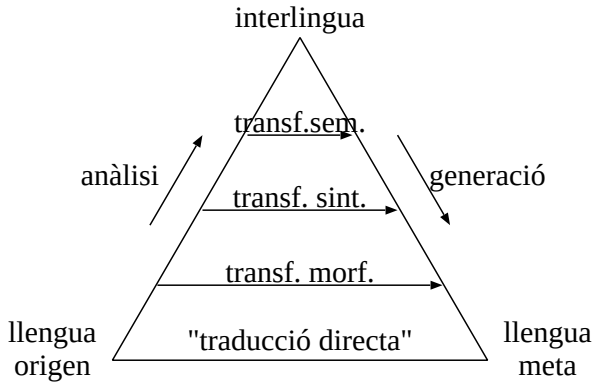


Figura 9.9: Com més profunda és l'anàlisi del text origen, més senzilla és (menys esforç comporta) la transferència a la representació corresponent de la llengua meta i més complexa la generació. L'anàlisi del text origen és tan profunda en els sistemes d'interlingua que no és necessària la transferència.

9.4 Sistemes sense transferència: els sistemes d'*interlingua*

Els sistemes anomenats d'*interlingua* apareixen en el cas extrem en què l'anàlisi de la frase d'origen és tan profunda que la traducció es pot generar directament a partir d'aquesta sense fer-hi transferència (vegeu la fig. 9.9). En particular, es parla d'*interlingua* quan la representació interna que s'obté de l'anàlisi és independent de quines siguen la LO i la LM, és a dir, la interlingua és *lingüísticament neutral*.

Les interlingües poden ser de molts tipus. Alguns sistemes com DLT (Hutchins i Somers 1992, cap. 17) usen una llengua "natural" com l'esperanto; altres sistemes usen representacions estructurals més o menys complexes per a representar les relacions semàntiques entre els elements de la frase. En l'intent de representar els significats de totes les frases de totes les llengües, les interlingües acabarien per ser "models del món". Això fa que, actualment, només s'hagen desenvolupat sistemes d'interlingua per a àmbits temàtics molt concrets.

Un dels avantatges més importants dels sistemes d'interlingua respecte dels sistemes de transferència és la facilitat amb què es pot afegir una llengua nova a un sistema de traducció automàtica multilingüe. Imaginem tres llengües que anomenarem L_1, L_2 i L_3. Un sistema complet de transferència que traduïra entre aquestes tres llengües en els dos sentits tindria tres mòduls d'anàlisi (que anomenarem A_1, A_2 i A_3), tres mòduls de generació (que anomenarem G_1, G_2 i G_3) i sis mòduls de transferència (que anomenarem T_{12},

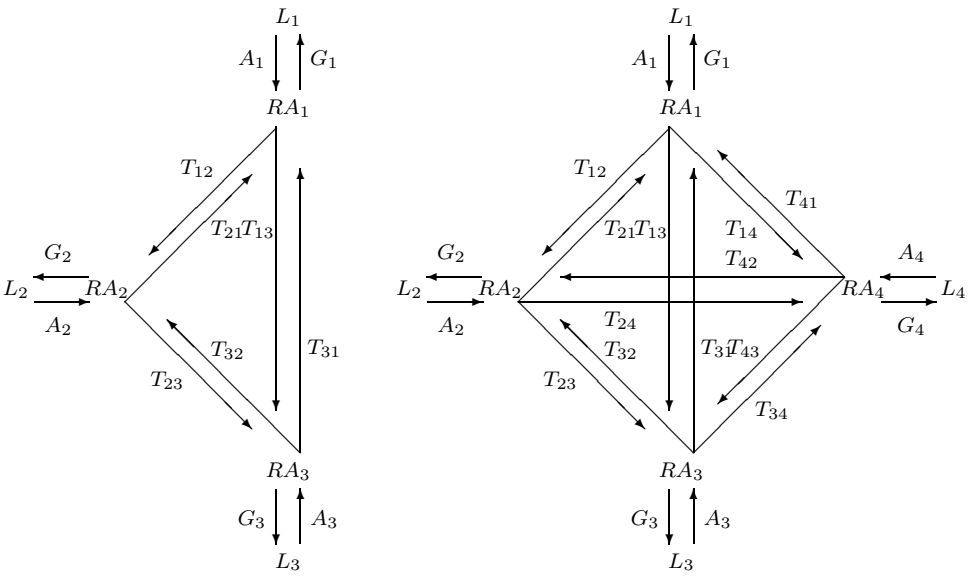


Figura 9.10: Cost d'afegir una quarta llengua L_4 a un sistema de transferència. Les entitats RA_1 a RA_4 són les representacions abstractes (tant RALO com RALM) que usen els mòduls de transferència.

T_{13}, T_{23}, T_{31}, T_{32} i T_{21})[12]. Afegir un quart idioma L_4 al sistema comporta:

- Crear un nou mòdul d'anàlisi (A_4).
- Crear un nou mòdul de generació (G_4).
- Construir 6 nous mòduls de transferència (T_{14}, T_{24}, T_{34}, T_{41}, T_{42} i T_{43}). Notem que per a aquesta última fase són necessaris diversos experts bilingües en sistemes de transferència[13].

La figura 9.10 il·lustra el cost d'afegir L_4 al sistema de transferència; En canvi, en un sistema d'interlingua no hi ha mòduls de transferència; un sistema trilingüe basat en una interlingua tindria només sis mòduls: tres d'anàlisi (A'_1, A'_2 i A'_3) i tres de generació (G'_1, G'_2 i G'_3). Queda clar que els mòduls d'anàlisi i de generació en aquests sistemes són més complexos que en el cas de transferència (ja que han de fer transformacions més profundes),

[12] En general, per a N llengües $L_1, L_2, \ldots, L_N$ hi hauria N mòduls d'anàlisi, N mòduls de generació i $N(N-1)$ mòduls de transferència.

[13] En el cas general d'afegir una llengua a un conjunt de N llengües, calen $2N$ nous mòduls de transferència.

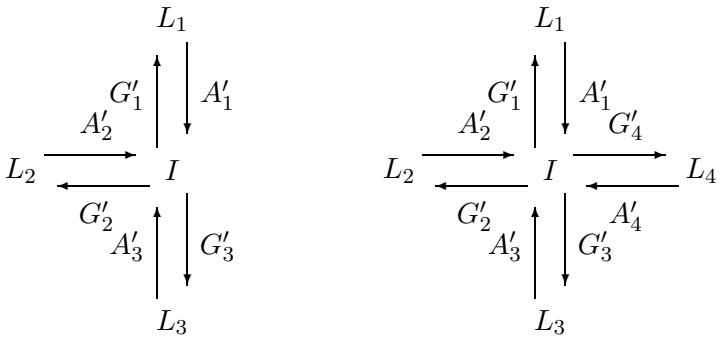


Figura 9.11: Cost d'afegir una quarta llengua L_4 a un sistema d'interlingua.

però també és clar l'avantatge del sistema d'interlingua a l'hora d'afegir-hi la llengua L_4: només cal dissenyar dos mòduls nous, A'_4 i G'_4, i per a dissenyar-los només necessitem una persona que conega bé la llengua L_4 i la interlingua I que usa el sistema. La figura 9.11 il·lustra el cost d'afegir L_4 al sistema.

9.5 Qüestions, exercicis i problemes

1. Els sistemes de traducció mot per mot poden cometre, per exemple, errors en la concordança de gènere o de nombre. Elegiu dues llengües L_1 i L_2 i poseu almenys dos exemples de traduccions mot per mot de L_1 a L_2 amb problemes de concordança.

2. CasCat és un sistema de traducció automàtica del castellà al català que usa regles que reordenen seqüències de categories morfològiques. Les regles s'apliquen de la manera usual: d'esquerra a dreta, reordenant la seqüència més llarga possible, i sense que se solapen les àrees reordenades. Heus ací algunes frases castellanes amb *cuyo*, les traduccions produïdes per CasCat, i, on la traducció és incorrecta, una alternativa acceptable.

   (a) *La chica cuyos compañeros murieron es china*
   La noia els companys de la qual van morir és xinesa

   (b) *La chica cuyos compañeros de clase murieron es china*
   La noia els companys de classe de la qual van morir és xinesa

   (c) *La chica cuyos compañeros mayores murieron es china*
   La noia els companys grans de la qual van morir és xinesa

(d) *La chica cuyos compañeros de clase de francés murieron es china*
**La noia els companys de classe de la qual de francès van morir és xinesa*
(*La noia els companys de classe de francès de la qual van morir és xinesa*)

(e) *La chica cuyos compañeros mayores de classe murieron es china*
**La noia els companys grans de la qual de classe van morir és xinesa*
(*La noia els companys grans de classe de la qual van morir és xinesa*

(f) *La chica cuyos compañeros mayores de clase de francés murieron es china*
**La noia els companys grans de la qual de classe de francès van morir és xinesa*
(*La noia els companys grans de classe de francès de la qual van morir és xinesa*)

Les traduccions inacceptables estan marcades amb un asterisc. Proposeu un conjunt de regles de reordenament que expliquen el conjunt de traduccions observat. En quins casos es "trenquen" sintagmes?

3. La multinacional WorldTrans ha decidit ampliar el seu sistema de traducció automàtica multilingüe LetTrans (que tradueix correspondència comercial entre qualssevol dues llengües d'un grup de quinze) i afegir-hi la capacitat de traduir del swahili a les quinze llengües i de les quinze llengües cap al swahili. En una oferta de treball, WorldTrans demana experts en swahili però no demana cap expert en traducció entre swahili i cap de les quinze llengües. Quina classe de sistema de traducció automàtica és LetTrans? Justifiqueu la resposta.

4. Quina és l'operació inversa de l'anàlisi morfològica?

   (a) L'obtenció de la forma lèxica d'un mot a partir de la forma superficial.
   (b) La generació morfològica.
   (c) La transferència morfològica.

5. La traducció automàtica per transferència és sempre...

   (a) ... morfològica.
   (b) ... directa.
   (c) ... indirecta.

6. Dues traduccions possibles del mot valencià *cap* al castellà són *cabe* o *cabeza*. Com podria fer l'elecció adequada un sistema de traducció automàtica?

(a) Posant-hi la traducció més probable, basada en les freqüències d'ús dels mots.

(b) Usant informació morfosintàctica, ja que en la posició concreta de la frase podria anar només un verb o un substantiu.

(c) No podria, perquè les dues traduccions són sempre possibles en qualsevol frase.

7. Quines de les següents representacions intermèdies son més costoses d'obtenir a partir de les frases?

(a) Els arbres d'anàlisi sintàctica corresponents.

(b) Les seqüències de categories morfològiques corresponents.

(c) Les estructures semàntiques superficials corresponents.

8. Només una d'aquestes operacions és esperable en una situació normal de traducció automàtica:

(a) La postedició en un sistema de traducció automàtica usat per a l'assimilació.

(b) Una fase complexa de transferència en un sistema d'interlingua.

(c) La preedició en un sistema de traducció automàtica usat per a la disseminació.

Quina és?

9. L'anàlisi morfològica pren una oració i...

(a) ... produeix un arbre d'anàlisi.

(b) ... produeix, per a cada mot, totes les formes superficials corresponents.

(c) ... produeix, per a cada mot, totes les parelles lema–informació morfològica possibles.

10. Quines són les fases bàsiques d'un sistema de traducció automàtica indirecta?

(a) Anàlisi, generació i traducció

(b) Anàlisi, transferència i generació

(c) Anàlisi i transferència.

11. Quin dels següents tipus de sistema de traducció automàtica faciliten més l'addició d'una nova llengua?

(a) Els sistemes de transferència morfològica.

(b) Els sistemes de transferència semàntica superficial.

(c) Els sistemes d'*interlingua*.

12. Quin dels següents tipus de sistema de traducció automàtica té una fase de *generació* més complexa?

    (a) Els sistemes de transferència morfològica

    (b) Els sistemes de transferència semàntica superficial

    (c) Els sistemes d'*interlingua*

13. Quin dels següents tipus de sistema de traducció automàtica tenen la fase de transferència més senzilla possible?

    (a) Els sistemes de transferència morfològica

    (b) Els sistemes de transferència semàntica superficial

    (c) Els sistemes d'*interlingua*

9.6 Solucions

1. Per exemple, L_1=castellà i L_2=català: *un buen postre* → **un bon postres* (*unes bones postres*); *una señal inequívoca* → **una senyal inequívoca* (*un senyal inequívoc*).

2. Les traduccions observades es poden explicar amb les tres regles següents:

   - R_1: **cuyo n** → **art n de art qual**
   - R_2: **cuyo** $\mathbf{n_1}$ **de** $\mathbf{n_2}$ → **art** $\mathbf{n_1}$ **de** $\mathbf{n_2}$ **de art qual**
   - R_3: **cuyo n adj** → **art n adj de art qual**

   Les regles que s'apliquen en cada cas són:

   (a) R_1

   (b) R_2

   (c) R_3

   (d) R_2; no abarca el segment *de francés* i trenca el sintagma;

   (e) R_3; no abarca el segment *de clase* i trenca el sintagma;

   (f) R_3; no abarca el segment *de clase de francés* i trenca el sintagma.

3. LetTrans és un sistema d'interlingua: per a afegir el swahili només es necessiten experts en swahili i en la interlingua de LetTrans. Si fóra un sistema de transferència seria necessària la participació d'experts bilingües en swahili i cada una de les quinze llengües que ja hi ha en el sistema.

4. (b)

5. (c)
6. (b)
7. (c)
8. (c)
9. (c)
10. (b)
11. (c)
12. (c)
13. (c)

Capítol 10

Memòries de traducció

Una aproximació a la traducció humana assistida per ordinador (és a dir, semiautomàtica) que està molt relacionada amb la traducció directa és la que s'usa en les anomenades *memòries de traducció*[1]. La noció bàsica (Somers i Rutzler 1996; Samuelson-Brown 1996) és la utilitat de tenir a mà, quan s'està traduint un text nou, exemples de frases similars i de les traduccions corresponents, provinents de traduccions realitzades anteriorment. De fet, certs tipus de textos com ara documents tècnics, informes anuals o manuals d'instruccions, els quals se solen revisar freqüentment, sovint tenen moltes repeticions. En aquests casos, la comparació de versions diferents del que és essencialment el mateix text i la traducció repetitiva de textos similars és innecessàriament laboriosa. A més, moltes vegades el treball de traducció comporta un esforç creatiu considerable, com ara quan es tracta de trobar una equivalència adequada a alguna expressió especialment difícil de traduir; les memòries de traducció permetrien no haver de repetir aquest esforç en el futur.

Les memòries de traducció són programes que permeten automatitzar (almenys en part) aquest procés de reutilització o *reciclatge* de traduccions antigues. L'èxit d'una memòria de traducció depén de la seua capacitat per a realitzar automàticament o assistir la persona usuària a fer les tasques següents:

- *Alinear* els textos i les traduccions existents[2] per a identificar fragments o *unitats de traducció* que es puguen reutilitzar posteriorment. La longitud d'aquests fragments pot anar des dels mots fins a les oracions senceres, però s'ha d'elegir bé, ja que es produeix un compromís: si els fragments són menuts, és més probable que puguen repetir-se en

[1] En comptes de traduir mot a mot fent una simple substitució de cada mot origen pel(s) mot(s) meta corresponent, les memòries de traducció fan substitucions de fragments de més d'un mot, i, en lloc d'usar un diccionari bilingüe, usen una base de dades de fragments de més d'un mot prèviament traduïts.

[2] Tots dos en format informatitzat, és clar!

textos nous, però poden donar lloc a correspondències múltiples entre les quals s'hauria d'elegit; en canvi, si els fragments són massa llargs és molt menys probable que es repetisquen exactament en textos futurs, però és improbable que siguen ambigus. És a dir, per una banda, es dona un compromís entre la *cobertura* (fracció d'un text nou que podria ser traduït usant els fragments alineats) i la *precisió* (correcció de les traduccions resultants).

Però és que, d'altra banda, l'alineament automàtic de textos traduïts no és una tasca senzilla; els coneixements que es tinguen sobre les llengües involucrades (per exemple, correspondències entre mots o alineaments prèviament validats per una persona experta) poden ajudar molt l'alineador automàtic. En vista de la dificultat de l'alineament, i en absència del coneixement necessari per a fer un alineament més fi, es pot dir que la major part dels alineadors usats amb les memòries de traducció actuals examinen el format i la puntuació dels textos amb algorismes que divideixen (aproximadament) els textos origen i meta en unitats que podríem classificar com a *oracions* o *frases* i exploten aquesta divisió per a alinear els textos. Com que els alineaments automàtics poden no ser correctes, la majoria dels programes actuals ofereixen a la persona usuària la possibilitat de validar o modificar (usant una representació gràfica, comandes senzilles i el ratolí) l'alineament automàtic inicial abans d'incorporar-lo a la memòria de traducció.

- *Organitzar* aquests fragments en una base de dades de manera que el programa o la persona usuària hi puguen accedir eficientment. S'ha de tenir en compte que la utilitat de les memòries de traducció millora considerablement amb la grandària del corpus de traduccions usades per a omplir-les; per tant, no és estrany que una memòria de traducció haja de gestionar una gran quantitat de fragments. Molts programes marquen els fragments amb un codi que indica la temàtica o la naturalesa del document del qual s'ha extret el fragment, de manera que la temàtica del nou document servisca per a localitzar el fragment més adequat en cada cas.

- *Recuperar* de la memòria de traducció, quan s'està traduint un text nou, els fragments més adequats i *construir* amb ells la traducció. Aquests fragments poden ser idèntics o similars. En cas de trobar fragments *idèntics* per als quals només hi haja una traducció disponible, només cal inserir-ne la traducció directament, però això succeeix poques vegades. L'exit d'aquesta fase depén en gran part de la capacitat del sistema per a proposar traduccions per a segments *similars* (i per a això s'han de definir i usar criteris adequats de *similitud*). Normalment, els sistemes, quan troben un fragment similar, destaquen gràficament les diferències perquè la persona usuària faça la modi-

ficació necessària perquè la traducció resultant siga correcta; alguns sistemes, fins i tot, són capaços d'usar bases de dades lèxiques o terminològiques per a proposar traduccions per als nous mots. També es pot donar el cas que una frase nova es puga traduir de més d'una manera perquè hi haja més d'una combinació de fragments pretraduïts que la cobrisca; en aquest cas, es pot usar un sistema d'avaluació (per exemple, estadístic) per a elegir la millor fragmentació o donar a la persona usuària la possibilitat de triar entre les possibles fragmentacions.

En l'actualitat, les memòries de traducció de propòsit general disponibles comercialment (*Trados*, *Déjà Vu* d'Atril, *IBM Translation Manager*, *Transit* de Star, *SDLX* de SDL International, etc., per nomenar algunes de les més conegudes) no són completament automàtiques, però no és imposible una automatització total del procés, especialment quan es disposa d'una gran quantitat de documents pretraduïts. De fet, els esborranys de l'edició bilingüe (castellà–català) d'*El Periódico de Catalunya* (vegeu l'epígraf 13.4) es preparen diàriament amb un mètode completament automàtic que funciona de manera similar a una memòria de traducció.

L'intercanvi de memòries de traducció: Freqüentment, els traductors formen equips que col·laboren a l'hora de produir les traduccions; quan usen memòries de traducció, és possible que hi haja traductors que preferisquen un programa i uns altres que en preferisquen un altre. Vol dir això que no podran compartir les memòries de traducció que hagen anat construint? Per sort, no. En agost de 1998 es va aprovar la versió 1.1 d'un format estàndard anomenat TMX (*Translation Memory eXchange*, "intercanvi de memòries de traducció"); quasi tots els programes gestors de memòries de traducció poden escriure i llegir memòries en aquest format, que actualment (setembre de 2001) ha arribat a la versió 1.3. El format TMX segueix les especificacions XML (vegeu l'apartat 4.1.6); és a dir, les memòries TMX són un tipus de document XML, definit, per tant, per una DTD concreta.[3]

De qualsevol manera, l'intercanvi de memòries de traducció entre traductors o equips de traducció diferents no està exempt de problemes. D'una banda, es poden produir incoherències terminològiques i d'estil entre els fragments procedents de grups diferents; les decisions en cas de conflicte comporten mecanismes complexos de reconeixement d'autoritat o de prestigi, que poden ser difícils de consensuar. D'altra banda, l'organització, el manteniment i la explotació de grans memòries de traducció distribuïdes (en les diverses màquines d'una xarxa) està lluny de ser trivial. Per exemple, en el cas del castellà i el català, una gran memòria de traducció alimentada amb

[3]Tota la informació sobre TMX es troba en el servidor de LISA (*Localisation Industry Standard Organization*), un consorci d'empreses de *localització* (adaptació del programari a les llengües locals dels usuaris) per a la creació d'estàndards com ara el TMX. L'URL és `http://www.lisa.org/tmx/`.

les traduccions fetes només en l'àmbit de les administracions autonomiques i locals estalviaria grans quantitats de temps i diners a l'hora de mantenir la documentació bilingüe d'aquestes institucions, però encara no s'ha substanciat un recurs d'aquesta mena, malgrat la gran quantitat de veus que n'han expressat la necessitat i la conveniència.

Capítol 11

L'avaluació dels sistemes de traducció automàtica

Aquest capítol pretén enunciar i descriure molt breument alguns dels aspectes rellevants de l'avaluació dels sistemes de traducció automàtica i donar algunes referències que puguen ser d'interés per a qui vulga aprofundir en aquest tema.

11.1 Qüestions bàsiques

Quan ens plantegem l'avaluació dels sistemes de traducció automàtica (TA), hi ha algunes preguntes bàsiques que cal respondre. Arnold et al. (1994) plantegen el problema així:

- Com es pot decidir si un sistema de TA és *bo*?
- Com es pot decidir si un sistema de TA és *millor* que un altre?

i afegeixen la pregunta clau: "Què vol dir *bo* o *millor* en aquest context?" La resposta a totes aquestes preguntes és molt difícil, com diu Minnis (1994): "el fet que no s'haja proposat cap métode d'avaluació o de mesurament estàndard és un bon indicador de la magnitud del problema".

11.2 Tipus d'avaluació

La naturalesa de l'avaluació d'un sistema de TA depén de diversos factors:

1. *Per a què* es fa l'avaluació? Hutchins (1996) distingeix entre tres tipus bàsics d'avaluació:
   (a) l'*avaluació d'adequació*, que serveix per a "determinar la idoneïtat dels sistemes de TA en un context operacional especificat" —per

exemple, per a decidir si el sistema de TA és bo per a traduir el correu comercial d'una empresa alimentària—;

(b) l'*avaluació diagnòstica*, que serveix per a "identificar limitacions, errors o deficiències, les quals poden ser corregides o millorades" —per exemple, defectes en el tractament de la concordança verbal de les oracions subordinades—, i

(c) l'*avaluació de funcionament*, "per a valorar l'estat de desenvolupament del sistema o les diferents realitzacions tècniques" —per exemple, si el programa és robust, ràpid, fa un ús racional de la memòria del sistema, etc.

Quant a l'*avaluació diagnòstica*, vegeu en la secció 11.4 la discussió sobre la noció d'*avaluació predictiva*.

2. *Qui* fa l'avaluació? L'avaluació la poden fer:

   (a) les persones que presumiblement usaran el sistema o l'adquiriran per a una empresa (avaluació d'adequació);

   (b) els investigadors, equips de desenvolupament, programadors (avaluació diagnòstica), molt especialment durant el desenvolupament d'un sistema de TA;

   (c) qualsevol dels dos grups anteriors (avaluació del funcionament).

3. *Com* es fa l'avaluació? Quan s'avalua un sistema de TA es tenen en compte:

   (a) *La qualitat de les traduccions en brut* produïdes pel sistema. La qualitat és una combinació (en proporcions difícils de determinar[1]) de diversos factors, com ara: l'*intel·ligibilitat* dels documents traduïts per part dels usuaris; la *precisió* o *fidelitat* amb què el text traduït comunica el significat del document original (les quals han de ser jutjades per part de persones bilingües coneixedores de la temàtica dels documents); l'adequació de l'estil o del registre dels documents traduïts, etc.

   Aquesta avaluació es pot fer mitjançant l'ús de col·leccions de documents típics o representatius (com se sol fer en les avaluacions d'adequació) o mitjançant sèries de proves objectives (angl. *test suites*), usades en les avaluacions diagnòstiques[2] i dissenyades per a abraçar conjunts complets de fenòmens lingüístics que es manifesten en la traducció.[3]

[1] Minnis (1994) diu: "La raó per la qual el mesurament de la qualitat és difícil és, per descomptat, el fet que la qualitat siga un concepte tan polifacètic i intangible".

[2] Però no únicament, com indica Lewis (1997), ja que també poden servir perquè els usuaris jutgen l'adequació de l'eixida produïda pel sistema.

[3] Per exemple, el reordenament dels mots dels sintagmes nominals quan es tradueix de l'anglés al castellà (Mira i Gimènez i Forcada 1998; Forcada 2000).

Una possible mesura quantitativa de la qualitat d'un text produït per un sistema de TA és el nombre mínim d'insercions, esborraments i substitucions de mots necessaris (per exemple, per cada 100 mots de text) per a transformar-lo en un text que siga una traducció acceptable per[4] de l'original. A més, la determinació de la qualitat d'una traducció en brut per còmput del nombre de correccions necessàries no està exempta de problemes:

- Aquest mètode dóna la mateixa importància a totes les operacions de correcció, independentment del mot. Això pot no ser adequat.
- Si suposem que existeix una única traducció acceptable del text origen i l'usem com a referència, hi ha més d'una manera de corregir la traducció en brut de manera que el resultat siga idèntic al de referència. Per a poder fer comparacions, estem interessats en la correcció produïda el nombre mínim d'operacions d'inserció, esborrament i substitució de mots; aquest nombre mínim es pot considerar una *distància*.[5] La recerca d'aquesta manera òptima de corregir pot no ser trivial per a una persona, especialment si els errors apareixen junts i agrupats.
- Però és que, a més, la traducció de referència pot no estar disponible; a més, en la majoria dels casos no hi ha una única traducció acceptable. De nou, si volem comparar, voldríem trobar la traducció acceptable més pròxima a la traducció en brut, és a dir, la que s'obté amb el mínim de correccions possibles. És a dir, avaluar (corregir) la traducció en brut suposa per tant fer una doble recerca: la persona que corregeix ha de buscar mentalment la traducció acceptable més propera (tot tenint en compte els criteris que fan acceptable una traducció, els quals poden no ser fàcils d'aplicar), però la *distància* entre els dos textos també es calcula fent una recerca mental del nombre mínim de correccions necessàries.

El fet que és possible que l'avaluació per recompte de correccions no siga òptima en vista d'aquests problemes fa que, a més, siga especialment difícil comparar les avaluacions fetes per persones diferents.

D'altra banda, sempre s'ha de tenir en compte que els mètodes d'avaluació de la qualitat depenen de l'ús que es pensa donar al sistema de TA (Arnold et al. 1993); per exemple, l'avaluació d'un

[4] De qualsevol manera, queda el problema que un text que per a un lector és acceptable pot no ser-ho per a un altre.

[5] De fet, matemàticament, ho és: s'anomena *distància d'edició* (angl. *edit distance*).

sistema que s'usa per a la *disseminació* de material és ben diferent de l'avaluació d'un sistema usat per a l'*assimilació* d'informació.[6]

(b) *La qualitat del sistema de TA mateix*: per exemple, "la facilitat amb què es poden crear i actualitzar diccionaris, posteditar els textos, controlar el llenguatge d'entrada" o "l'extensibilitat [del sistema] a parells nous d'idiomes o a noves temàtiques" (Hutchins 1996).

(c) *La comparació dels costos i dels beneficis* d'usar un sistema de TA en comptes dels serveis de professionals de la traducció: per exemple, si costa més (en despeses de personal) la postedició (revisió) dels textos meta produïts pel sistema que la traducció completa dels textos origen per part de professionals, l'adopció del sistema de TA no convé a una empresa.

11.3 Sobre la comparació entre traducció automàtica i traducció humana

Una visió predominant de l'avaluació dels sistemes de TA és l'anomenada *metàfora del traductor humà*, segons la qual (Krauwer 1993) la tasca consisteix a "determinar fins a quin punt els constructors del sistema han aconseguit imitar el comportament d'un traductor humà". Sager (1993) ho formula dient que "S'ha argumentat que la qualitat dels documents produïts mitjançant traducció automàtica s'hauria d'avaluar en termes de la identitat amb productes humans".

Tant Krauwer (1993) com Sager (1993) qüestionen aquesta visió; aquest últim argumenta que "s'ha d'acceptar que no hi ha cap situació que puga servir com a punt de comparació entre la traducció humana i l'automàtica, i que potser no hi ha cap situació en la qual la traducció humana i l'automàtica siguen igualment adequades" i proposa que, en canvi, les traduccions poden ser comparades per a veure "si satisfan, i fins a quin punt, les expectatives de l'usuari final [dels documents traduïts]", ja que la traducció és una "activitat de mediació, la forma particular de la qual està determinada tant pel text com per les circumstàncies comunicatives que requereixen aquesta mediació". En concret, la traducció automàtica pot ser la més adequada en algunes circumstàncies, en vista de l'enorme demanda general existent i, més concretament, de la demanda de traduccions ràpides i barates que no poden ser produïdes per professionals.

[6] De nou, com en el capítol 6, la noció central és la de *propòsit* de la traducció.

11.4 Avaluació predictiva

Hi ha un tipus d'avaluació que es pot considerar com a cas particular de l'avaluació diagnòstica definida més amunt, encara que no s'use estrictament per a millorar el funcionament d'un sistema, sinó només per a predir el comportament del sistema en situacions noves. L'anomenarem ací *avaluació predictiva*.

Per a poder fer l'avaluació predictiva, És crucial que els avaluadors tinguen, en primer lloc, un model que descriga aproximadament el funcionament del sistema de traducció automàtica (relacionat amb la tipologia del sistema), i, en segon lloc, un conjunt de textos o frases d'avaluació (*test suite*) que els permeta obtenir detalls concrets sobre les dades (p.ex., regles) que usa aquell model. Les prediccions serien de l'estil de "com que sembla usar regles patró–acció de l'estil de "si troba un patró X farà l'acció Y" i en una sèrie de casos troba el patró X_1 i fa l'acció Y_1, podem predir que sempre que trobe aquest mateix patró farà la mateixa acció". Com que la majoria dels sistemes comercials no ens donen suficient informació sobre la naturalesa del model, els haurem de tractar com com una *caixa negra*; la intuïció de la persona avaluadora, el seu coneixement d'altres sistemes o de la història de les empreses involucrades (per exemple, quant a l'adquisició de tecnologia d'altres empreses) i la seua habilitat per a elegir exemples reveladors li permetran determinar aspectes bàsics del model de traducció. En particular, qualsevol avaluació predictiva necessita tenir una idea clara sobre el nivell d'anàlisi que es fa en el sistema de TA, ja que el nivell d'anàlisi és el que determina més la naturalesa d'un sistema (vegeu l'apartat 9.3). L'avaluació predictiva pot tenir, a més, un paper fonamental en l'educació dels futurs traductors (Mira i Gimènez i Forcada 1998; Forcada 2000), en vista de la disponibilitat creixent de sistemes comercials (per exemple, a través d'internet).

D'altra banda, perquè l'avaluació siga útil els conjunts de prova haurien d'estar dissenyats de manera que abraçaren conjunts complets de fenòmens lingüístics que es manifesten amb freqüència rellevant en les situacions reals de traducció que es volen avaluar, ja que es vol predir el comporatment del sistema en aquestes situacions concretes.

Capítol 12

Problemàtica de la TA castellà–català

12.1 Introducció

Les aplicacions potencialment més interessants de la TA castellà–català s'emmarquen dins de l'anomenada *normalització lingüística*, és a dir, l'esforç de les societats de parla catalana per promoure'n l'ús normal en tots els àmbits; un exemple actual el constitueixen els servidors d'Internet d'institucions públiques i d'empreses privades dels Països Catalans, on la presència del català és encara minoritària. Quan la llengua original dels documents és el castellà, es podria usar un sistema de TA per a generar esborranys de documents catalans (o, fins i tot, documents correctes si els documents castellans estan escrits en un llenguatge controlat).

En el cas concret del castellà i el català, la proximitat lingüística entre les dues llengües fa que siga abordable el disseny de sistemes de traducció automàtica que generen textos d'un nivell de correcció tal que resulte més rendible revisar el resultat en brut produït pel programa que fer la traducció completa.

En aquest capítol es presenten alguns dels problemes més importants amb què es pot trobar qui vulga dissenyar un sistema de traducció automàtica per a traduir textos del castellà al català. A la vista de la notable similitud lingüística existent entre les dues llengües, es podria pensar que la tasca de traducció automàtica podria, en la majoria dels casos, ser tan senzilla com substituir un a un els mots castellans pels seus equivalents catalans. De fet, el model de traducció automàtica *mot per mot* (definit en la pàg. pg:mpm, i que no s'ha de confondre amb el que els traductors humans anomenen *traducció literal*) és el model de referència que usarem en aquest capítol: els tres grups de *problemes* que es presenten en aquesta capítol són alguns —no tots— dels que no resol el model mot per mot: la segmentació del text origen, l'homografia i les divergències sintàctiques.

12.2 Segmentació del text origen

La segmentació d'un text en mots sol ser normalment molt senzilla: el programa pot usar els blancs, els tabuladors, els finals de línia o els signes de puntuació com a fronteres entre mots. Però de vegades no és tan fàcil: el castellà uneix moltes vegades diversos mots en un sol mot, sense que s'hi puguen distingir els mots components; en català, llevat de contraccions com *al* i *del*, sempre queda alguna indicació d'aquesta unió, com ara un apòstrof o un guionet. Per exemple, en castellà, els pronoms enclítics s'uneixen a l'imperatiu, a l'infinitiu i al gerundi, i moltes voltes fan que en canvie la forma (vegeu l'epígraf 9.3.1). Per sort, aquests problemes es poden resoldre de manera senzilla usant analitzadors morfològics com els que es descriuen en l'epígraf 9.3.2.

12.3 Homografia

L'homografia pot produir ambigüitat lèxica categorial o fins i tot aparéixer entre mots de la mateixa categoria lèxica. L'homografia apareix quan un mot (anomenat usualment *homògraf*) té més d'una anàlisi morfològica possible (vegeu l'apartat 8.1). El castellà —com les altres llengües romàniques— té molts homògrafs. Una de les fonts més importants d'homografia és la coincidència entre algunes terminacions de la flexió verbal i algunes terminacions de la flexió nominal i adjectival (*-a*, *-as*, *-o*, *-e*, *-es*), ja que involucra categories lèxiques obertes amb molts membres[1]. Però hi ha, a més, altres fonts menys productives d'ambigüitat, com ara la coincidència d'algunes de les terminacions del present d'indicatiu dels verbs en *-ar* amb les del present de subjuntiu dels verbs en *-er* i *-ir* i al revés. Finalment, hi ha algunes homografies fortuïtes (com ara *para*: preposició i verb).

Per a il·lustrar aquest fet, es presenta un assaig de classificació —no exhaustiva— dels homògrafs castellans:

1. Homografia verb conjugat–substantiu:
   (a) En *-a*:
      - Pres. ind., 3a pers. sing. (1a conj.) / subst. fem. sing.: *casa*, *pinta*, *sala*, *toma*, *entrega*, *osa*.
      - Pres. subj., 1a i 3a pers. sing. (2a i 3a conj.) / subst. fem. sing.: *bata*, *tema*, *meta*
      - Altres: *era* (verb *ser*, 1a i 3a pers. sing. pretèrit imperf. i subst fem sing.).
   (b) En *-as*:

[1] De vegades, els mots homògrafs comparteixen una semàntica relacionada, com *ahorro*, i altres vegades no, com *oso*.

- Pres. ind. 2a pers. sing. (1a conj.) / subst. fem. pl.: *casas*, *salas*, *tomas*, *entregas*, *osas*;
- Pres. ind. 2a pers. sing (2a i 3a conj.) / subst. fem. pl.: *batas*, *temas*, *metas*.
- Altres: *eras*.

(c) En *-e*:

- Pres. subj., 1a/3a pers. sing. (1a conj.) / subst. masc. i fem. sing.: *cante*, *deje*, *sobre*, *pose*, *apunte*
- Pres. ind., 3a pers. sing. (1a conj.) / subst. masc. i fem. sing.: *vale*.
- Altres: *traje* (verb *traer*, 1a pers. sing. pretèrit indefinit i subst. masc. sing.)

(d) En *-es*:

- Pres. subj., 2a pers. sing. (1a conj.) / subst. masc. i fem. pl.: *sales* (verb *salar*), *ases* (verb *asar*), *cantes*, *dejes*, *sobres*, *poses*, *apuntes*
- Pres. ind., 2a pers. sing. (1a conj.) / subst. masc. i fem. pl.: *vales*, *sales* (verb *salir*), *ases* (verb *asir*).

(e) En *-o*:

- 1a pers. del present d'indicatiu / subst. masc. sing.: *oso*, *remiendo*, *riego*, *mando*, *canto*, *cardo*, *recibo*, *abono*, *saldo*;
- altres: *vino*.

(f) En *-os*: marchamos (1a pers. pl. present i pretèrit perfet simple d'indicatiu i subst. masc. pl.).

(g) Altres terminacions: *sal* (verb *salir*) *mentís*, *pagaré*.

2. Homografia verb conjugat–adjectiu:

(a) En *-a*:

- Pres. ind., 3a pers. sing. (1a conj.) / adj. fem. sing.: *pinta*, *monda*, *baja*, *linda*.
- Pres. subj., 1a i 3a pers. sing. (2a i 3a conj.) / adj. fem. sing.: *viva*.

(b) En *-as*:

- Pres. ind. 2a pers. sing. (1a conj.) / adj. fem. pl.: *pintas*, *bajas*, *mondas*, *lindas*;
- Pres. ind. 2a pers. sing (2a i 3a conj.) / adj. fem. pl.: *vivas*.

(c) En *-e*:

- Pres. subj. 1a/3a. pers. sing. (1a conj.) / adj. masc. i fem. sing.: *leve*, *ausente*, *presente*.

(d) En *-es*:
- Pres. subj. 2a pers. sing. (1a conj.) / adj. masc. i fem. sing.: *leves*, *ausentes*, *presentes*.

(e) En *-o*:
- 1a pers. del present d'indicatiu / adj. masc. sing.: *pinto*, *mondo*, *bajo*, *lindo*, *vivo*.

3. Homografia verb conjugat–verb conjugat (molt difícil de resoldre):

(a) Entre verbs de la 1a conj. i verbs de la 2a o 3a conj.:
- *sentir/sentar*: *siento*, *sientes*, *siente*, *sienten*, *sienta*, *sientas*, *sientan*.
- *mentir/mentar*: com *sentir/sentar*
- *vendar/vender*: *vendo*, *venda*, *vendas*, *vendamos*, *vendáis*, *vendan*, *vende*, *vendes*, *vendemos*, *vendéis*, *venden*.
- *salir/salar*: *sales*, *sale*, *salen*
- *asir/asar*: como *salir/salar*
- *poder/podar*: *podamos*, *podáis*, *podemos*, *podéis*.
- *vengar/venir*: *vengo*, *vengas*, *venga*, *vengamos*, *vengáis*, *vengan*.

(b) Entre la 1a pers. pl. del present d'indicatiu i del pretèrit perfet simple ("pretèrito indefinido") dels verbs regulars de la 1a i 3a conjs.: *amamos*, *cantamos*, *conseguimos*, etc.

(c) Altres casos: *amase*, *amasen*, *amases* (*amar*, *amasar*); fui, fuiste, ... (*ir* i *ser*), *ven* (*ver* i *venir*), etc.

4. Homògrafs verb conjugat–preposició: *bajo*, *cabe*, *entre*, *para*, *sobre*.

5. Homògrafs adjectiu–preposició: *bajo*.

6. Homògrafs substantiu–preposició: *ante*, *sobre*.

7. Homògrafs verb conjugat–determinant: *uno*, *una*, *unas* (*unir*)

8. Homògrafs verb conjugat–adverbi: *así* (*asir*), *fuera* (*ser*, *ir*), *arriba* (*arribar*), *adelante* (*adelantar*), *cerca* (*cercar*).

9. Homògrafs adjectiu–adverbi: *mucho*, *poco*, *fuerte*...

10. Homògrafs substantiu–adverbi: *antes*, *tanto*, *mal*, *bien*...

11. Homògrafs adjectiu–substantiu: *complejo*, *impreso*, *derecho*...

12. Homògrafs determinant–pronom *la*, *los*, *las*, *lo* (en "lo que", "lo grande")

13. Altres homògrafs: *como* (conjunció i forma de *comer*), *ora* (conjunció i forma de *orar*), *bien* (conjunció, substantiu i adverbi)

12.4 Divergències de traducció

Imaginem que hem pogut segmentar el text castellà i que hem resolt correctament les ambigüitats lèxiques; si encara decidim fer la traducció mot per mot, ens trobarem que hi ha certes construccions per a les quals la traducció no és correcta, ja que els mots catalans no es corresponen mot per mot amb els castellans. Vegem quins són alguns dels problemes[2]:

Concordança de gènere i nombre: De vegades el gènere i el nombre d'un ítem lèxic varien del castellà al català. La dificultat per a un sistema de traducció automàtica apareix a l'hora de propagar el gènere i el nombre del nucli d'un sintagma als modificadors que hi hagen de concordar: *su único amparo → la seua única empara*; *un buen postre → unes bones postres*. Els problemes augmenten si la concordança s'ha de produir entre sintagmes distants: *el calor producido por el motor ha resultado ser nefasto → la calor produïda pel motor ha resultat ser nefasta*. Un problema similar el presenta l'establiment de la concordança del participi, inexistent en castellà en situacions com ara *todavía no la hemos estudiado con profundidad → encara no l'hem estudiada amb profunditat*.

L'article neutre: El castellà posseeix l'anomenat *article neutre*, que no té correspondència en català (*lo que me dijiste → el que em vas dir*); presenten dificultat especial les construccions usades per a expressar l'abstracció o la intensitat: *recibirá el informe lo más pronto posible → rebrà l'informe el més aviat possible*; *me asusta lo grande que es → m'espanta com és de gran*.

Els possessius: De vegades, el català usa articles determinats i construccions amb el pronom feble *en* on el castellà usa possessius: *cuando hagas cosas así debes valorar sus consecuencias → quan faces coses així n'has de valorar les conseqüències*.

Els relatius: El principal problema apareix quan es volen traduir oracions que contenen el relatiu possessiu *cuyo*, inexistent en català, on el més senzill és usar una construcció amb *qual*, que, a més, presenta un esquema de concordança diferent (*qual* ha de concordar amb l'antecedent, mentre *cuyo* concorda amb el nom que el segueix): *el contribuyente cuyos informes hemos solicitado llegará tarde → el contribuent els informes del qual hem sol·licitat arribarà tard* (vegeu el final de l'apartat 9.3.3).

Els pronoms febles: Els principals problemes es troben en la traducció de *lo*, ja que pot correspondre en català a alguna forma del pronom

[2] Aquesta secció es basa parcialment en un document generat per Sandra Montserrat.

masculí singular *lo* o a alguna forma del pronom neutre *ho*; en la traducció de *se*, el qual correspon normalment al reflexiu català *se* però en les combinacions castellanes *se la*, *se lo*, etc. pot correspondre de vegades a alguna forma de *li* o *els*, i en el fet que el castellà no té equivalents dels pronoms catalans adverbials *en* i *hi* (*me* ∅ *dio uno* → *me'n va donar un*); ∅ *había dos salidas* → *hi havia dues eixides*; *no* ∅ ∅ *dejó una* → *no n'hi va deixar cap*).

Règim preposicional: Hi ha diferències notables entre els règims preposicionals castellà i català: les preposicions castellanes davant de *que* completiu no apareixen en català (*el hecho de que me hable* → *el fet que em parle*); algunes preposicions no són possibles en català davant d'infinitiu (*el juego consiste en ganar...* → *el joc consisteix a guanyar...*), etc.

12.5 Qüestions i exercicis

1. Quina d'aquestes tres tasques és més difícil en un sistema de traducció automàtica castellà–català?
   (a) Decidir la traducció del pronom castellà *se* (pot ser *se*, *li* o *els*).
   (b) Detectar les formes de *tener que* i traduir-les per *haver de*.
   (c) Fer l'anàlisi morfològica de verbs seguits d'enclítics com ara *estudiémonoslos* o *dándoselo*.

2. Indica quina d'aquestes tres és la font més important d'homografia (ambigüitat lèxica categorial) del castellà:
   (a) Les coincidències d'algunes formes d'alguns noms i d'alguns adjectius amb certes formes conjugades d'alguns verbs.
   (b) Les coincidències d'algunes formes de noms amb preposicions.
   (c) Les coincidències d'algunes formes de noms amb adverbis.

3. El català no té cap construcció equivalent al *cuyo* castellà. En traducció automàtica del castellà al català, una alternativa interessant és posar primer el sintagma nominal que segueix al *cuyo* i després, una forma de *del qual* que concorde amb l'antecedent. Es pot fer sempre correctament aquesta operació en un sistema de traducció automàtica que no faça anàlisi sintàctica?
   (a) Sí, hi ha prou amb fer l'anàlisi morfològica.
   (b) No, perquè cal determinar bé la longitud del sintagma nominal que segueix a *cuyo* per a poder posar *del qual* en la posició correcta.
   (c) No, perquè *cuyo* no té un equivalent morfològic en castellà.

12.6 Solucions

1. (a)
2. (a)
3. (b)

Capítol 13

Experiències de TA castellà–català

En aquest capítol es descriuen breument cinc experiències de traducció automàtica del castellà al català: SALT, Ara, Es-Ca, el Traductor de *El Periòdico de Catalunya* i una altra, interNOSTRUM, amb una miqueta més de detall.

13.1 SALT, de la Generalitat Valenciana

El programa SALT porta el nom del *Servei d'Assessorament Lingüístic i Traducció* de la Conselleria de Cultura Educació i Ciència de la Generalitat Valenciana; es tracta d'un programa per al sistema operatiu Windows que ha desenvolupat un equip de programadors dirigit per Rafael Pinter sota la direcció lingüística de Josep Lacreu, responsable d'aquest servei. La disponibilitat del programa fins fa molt poc ha estat més aviat reduïda; actualment es pot descarregar gratuïtament de diversos servidors d'Internet (per exemple, http://sofia.uji.es/software/gva/salt/index.html) i el distribueixen els serveis de normalització lingüística d'algunes universitats. SALT tradueix textos (ASCII o RTF) castellans al valencià —l'estàndard dels textos meta es pot regular usant un menú molt senzill— o corregeix una bona part de les errades típiques dels textos valencians. El programa és interactiu, és a dir, moltes vegades pregunta a l'usuari com ha de resoldre una ambigüitat, i dialoga sempre en català; a més, l'usuari pot seguir visualment el procés de traducció (mot a mot amb modificacions locals) en dues passades. Els resultats són molt interessants. El programa està bàsicament concebut com una ajuda a les persones que volen començar a generar documents en valencià (entre altres eines, inclou una completíssima guia interactiva de gramàtica i estil).

13.2 Ara, d'Autotrad

El programa Ara, llançat l'any 2000 per l'empresa Autotrad de València[1] —el gerent de la qual és Rafael Pinter, responsable informàtic de SALT— és bàsicament una versió bastant millorada del SALT, amb una aparença molt similar però amb algunes diferències: produeix textos en català oriental estàndard, pot dialogar amb la persona usuària en castellà i en català, admet més formats de text (HTML, Microsoft Word 97 i Microsoft Word 2000), deixa els diàlegs de resolució d'ambigüitats per al final a fi de no interrompre el procés i permet programar més d'una tasca de traducció.

13.3 Es-Ca, de Incyta

El sistema de traducció automàtica Es-Ca ha estat desenvolupat per l'empresa Incyta de Cornellà (recentment adquirida per la multinacional Sail Labs), en col·laboració amb la Universitat Autònoma de Barcelona; es tracta d'un sistema de transferència sintàctica estàndard, hereu del sistema METAL de l'empresa Siemens. El sistema no es distribueix com a programa, sinó que es troba en Internet (http://www.incyta.es): l'usuari inscrit envia el text i el servidor li'l retorna traduït; el cost en 1999 era de 3 pessetes per paraula. El servidor dóna accés a una versió gratuïta de demostració que tradueix textos curts. Els resultats són molt acceptables en la major part dels casos.

13.4 El traductor d'*El Periódico de Catalunya*

Una experiència interessant de traducció castellà–català per a la disseminació és l'edició bilingüe del diari *El Periódico de Catalunya*[2]; el text original —en castellà la major part de les vegades— es tradueix usant una tècnica similar a les *memòries de traducció* (vegeu el capítol 10) i després és revisat per un equip de posteditors abans de ser publicat. El programa usat per *El Periódico de Catalunya* també es pot provar en Internet[3].

13.5 interNOSTRUM

Un equip d'investigadors de la Universitat d'Alacant, finançat per la Caja de Ahorros del Mediterràneo i la mateixa Universitat, està desenvolupant actualment sota la direcció de Mikel L. Forcada un sistema de traducció automàtica castellà–català anomenat interNOSTRUM. Més concretament l'objectiu del projecte (vigent des de novembre de 1998) és desenvolupar un

[1] URL http://www.ara-autotrad.com
[2] Disponible per Internet: http://www.elperiodico.es.
[3] http://www.softly.es/transacciones/traductor/index.htm

sistema de traducció automàtica del castellà a les variants estàndards del català i el sistema invers corresponent.

La versió actual d'interNOSTRUM (accessible de forma gratuïta a través d'Internet, http://www.internostrum.com) no és un producte acabat, però ja pot ser usat per a generar, gairebé instantàniament, esborranys de traduccions al català llestes per a ser corregides (posteditades).

Actualment, interNOSTRUM tradueix textos en formats ANSI, HTML i RTF del castellà al català oriental i al revés (sobre formats, vegeu l'epígraf 4.1) i permet la navegació traduïda per Internet (és a dir, permet la traducció instantània dels documents que es vagen visitant sense haver d'invocar explícitament el traductor). Al final del projecte, es disposarà de traductors que produïsquen i accepten altres variants estàndard del català. El traductor català–castellà està menys desenvolupat que el traductor castellà–català.

13.5.1 Característiques informàtiques

El traductor s'executa actualment sobre el sistema operatiu Linux i és accessible, com ja s'ha dit, a través d'un servidor d'Internet;[4] está constituït per 8 subprogrames independents que s'executen simultàniament (en paral·lel), elaboren la traducció per etapes i es comuniquen mitjançant canals de text ASCII legible[5] (cosa que facilita enormement el diagnòstic dels errors de traducció). Sis dels vuit subprogrames es generen automàticament a partir de les dades lingüístiques corresponents,[6] mitjançant ferramentes informàtiques desevolupades en el projecte; els altres dos mòduls (el primer i l'últim de la cadena de muntatge) s'encarreguen, el primer (*desformatador*), de separar els codis de format HTML o RTF del text, i l´últim (*reformatador*), de tornar a combinar-los per tal que la traducció conserve el format del text origen. La velocitat actual del sistema és de l'ordre de milers de mots per segon sobre un PC estàndard (molt més ràpida que SALT o Ara, que tradueixen a velocitats de l'ordre de desenes de molts per segon).

13.5.2 Característiques lingüístiques

interNOSTRUM és un sistema clàssic de traducció indirecta per transferència morfològica avançada (vegeu l'epígraf 9.3.1), amb les fases lingüístiques següents:

1. ANÀLISI:
    - Anàlisi morfològica

[4] També hi ha disponible una versió per a servidors basats en el sistema operatiu Windows.

[5] El sistema operatiu Linux permet construir una canonada (*pipeline*) o cadena de muntatge en la qual el text d'eixida d'un programa s'envia com a entrada a un altre (en comptes d'enviar-lo a la pantalla), i així successivament.

[6] Característica que permet estendre fàcilment el producte a altres idiomes

- Desambiguació lèxica categorial

2. TRANSFERÈNCIA:
   - Consulta del diccionari bilingüe
   - Tractament de patrons (concordança, reordenament, canvis lèxics)
3. GENERACIÓ:
   - Generació morfològica
   - Postgeneració (apostrofació, etc.)

Subprogrames basats en tècniques d'estats finits

Els subprogrames d'*anàlisi morfològica*, *consulta del diccionari bilingüe*, *generació morfològica* i *postgeneració* estan basats en *transductors d'estats finits*, molt similars als descrits en l'epígraf 9.3.2. Aquesta tecnologia permet velocitats de processament de l'ordre de 10.000 mots per segon, velocitats que pràcticament no depenen de la grandària dels diccionaris. Els *transductors d'estats finits* usats en interNOSTRUM lligen l'entrada símbol a símbol; cada vegada que es llegeix una lletra canvien d'estat i van produint, també lletra a lletra, una o més sortides.

Anàlisi morfològica: El subprograma d'anàlisi morfològica, que es genera automàticament a partir d'un *diccionari morfològic* de la llengua origen (LO), el qual conté els lemes, els paradigmes de flexió i les connexions entre ells. L'entrada són les formes superficials del text i la sortida, formes lèxiques consistents en lema, categoria lèxica i informació de flexió.

Consulta del diccionari bilingüe: El subprograma de consulta del diccionari bilingüe és invocat pel subprograma de tractament de patrons (vegeu més avall); es genera automàticament a partir d'un fitxer que conté les correspondències bilingües. L'entrada és la forma lèxica de la LO i la sortida, la forma lèxica corresponent en la llengua meta (LM).

Generació morfològica: El generador morfològic fa l'operació inversa a l'analitzador morfològic però amb formes de la LM i es genera automàticament a partir d'un diccionari morfològic de la LM.

Postgeneració: Les formes superficials que estan implicades en processos d'apostrofació i guionatge (pronoms febles, articles, algunes preposicions, etc.) activen aquest subprograma, que normalment es troba inactiu. El postgenerador es genera a partir de regles senzilles d'apostrofació, guionatge i combinació de pronoms febles.

Com ja s'ha discutit en el capítol 12, la divisió d'un text en mots presenta alguns aspectes no trivials; se n'esmenten dos: les *locucions* (o *girs*) i els pronoms enclítics.

Locucions i girs: Hi ha nombroses locucions i girs que es poden tractar com a *unitats multimot* i s'estan incorporant gradualment als diccionaris morfològics de les dues llengües i al diccionari bilingüe:

- *con cargo a* → *a càrrec de*
- *por adelantado* → *per endavant*, *a la bestreta*
- *el abajo firmante* → *el sotasignat*
- **echar** *de menos* → **trobar** *a faltar*

En l'últim exemple, el gir no és invariable sinó que té un element que es flexiona (en negretes).

Pronoms enclítics: El subprograma d'anàlisi morfològica també és capaç de resoldre les combinacions de verbs i pronoms febles enclítics en castellà, les quals presenten variacions ortogràfiques com ara canvis d'accentuació o pèrdua de consonants:

- *dámelo* = *da* + *me* + *lo* → *dóna* + *me* + *lo* = *dóna-me'l*
- *pongámonos* = *pongamos* + *nos* → *posem* + *nos* = *posem-nos*.

El sistema interNOSTRUM tracta aquests dos problemes amb l'analitzador morfològic, el qual és capaç de decidir quan un grup de mots s'ha de tractar conjuntament o per separat.

El subprograma de desambiguació lèxica categorial

Aquest subprograma usa un model de llenguatge basat en trigrames (seqüències de tres categories lèxiques), de l'estil del descrit en la secció 8.5.1. Aquest model es basa en les freqüències observades per a aquests trigrames en un corpus de referència, i assigna una probabilitat a cada possible desambiguació de la frase que conté mots amb ambigüitat categorial. La desambiguació més probable (la més versemblant) és l'elegida. En l'actualitat, les prestacions d'aquest subprograma són molt millorables perquè els corpus de referència actualment en ús no són encara suficientment representatius.

Els pocs errors en homògrafs *difícils* i *freqüents*, com ara *una* (article/verb, freqüència 0,77%), *para* (verb/prep., freqüència 0,77%) i *como* (conj./verb, freqüència 0,43%) degraden actualment molt la qualitat de la traducció. Altres homògrafs freqüents no són tan difícils de desambiguar.

Les ambigüitats lèxiques no categorials s'aborden amb estratègies *ad hoc* provisionals. En el futur, molts d'aquests homògrafs s'inclouran en la definició d'un *llenguatge controlat*, alternativa a la preedició consistent en l'aplicació de restriccions lèxiques, sintàctiques i d'estil als textos de la LO. La CAM ha encarregat també el disseny d'un castellà controlat per a textos financers i dels *assistents d'estil* corresponents per als autors.

El subprograma de tractament de patrons

Malgrat la gran semblança entre el castellà i el català, hi ha divergències gramaticals considerables:

- perífrasis modals: *tienen que firmar* → *han de firmar*;
- canvis de gènere i nombre: *la deuda contraída* → *el deute contret* (masc.);
- caiguda de preposicions: *la intención de que el cliente* → *la intenció* ∅ *que el client*;
- construccions relatives: *la cuenta cuyo titular es* → *el compte el titular del qual és*.

Aquestes divergències s'han de tractar amb les regles gramaticals escaients.

La solució elegida (estàndard en sistemes comercials, vegeu Mira i Gimènez i Forcada (1998), Forcada (2000)) es basa en la detecció i el tractament de seqüències predefinides de categories lèxiques (anomenades *patrons*), és a dir, una mena de sintagmes rudimentaris, com ara **art**–**nom** o **art**–**nom**–**adj**. Les seqüències considerades pel subprograma en formen el *catàleg* de patrons. El funcionament del subprograma es basa en un esquema patró–acció:

- Llegeix el text (analitzat i desambiguat) d'esquerra a dreta, categoria lèxica a categoria lèxica.
- Busca, en la posició actual de la frase, el patró més llarg que concorda amb un patró del seu catàleg (per exemple, si en la posició actual es llegeix "un senyal inequívoc…", tria **art**–**nom**–**adj** en comptes de **art**–**nom**).
- Opera sobre aquest patró (propagació de gènere i nombre, reordenament, canvis lèxics) seguint les regles associades a ell.
- Continua immediatament darrere del patró tractat (no torna a visitar els mots sobre els quals ha operat).

Quan no es detecta cap patró en la posició actual, es tradueix literalment un mot i es torna a iniciar el procés. Els fenòmens "a la llarga" com la concordança subjecte–predicat són una mica més difícils de tractar; s'usa un registre *estat* o *memòria* que recorda certes informacions al llarg del procés.

El subprograma de tractament de patrons es genera automàticament a partir d'un fitxer de regles que especifica els patrons i les accions associades. Aquest és molt probablement el subprograma més lent (uns 3 000 mots/segon).

13.5.3 Eines de suport a interNOSTRUM

Es projecta construir les eines següents:

- Un assistent d'estil que permetrà l'autor d'un text en castellà evitar moltes ambigüitats difícils de resoldre usant regles lèxiques, sintàctiques i d'estil (un *llenguatge controlat*, vegeu la secció 7.5).
- Un assistent de preedició, que permetrà una desambiguació manual de mots i estructures problemàtiques (simplement fent-hi clic per accedir als menús corresponents) quan els mètodes estadístics indicats més amunt siguen incapaços de fer les tries correctes.
- Un assistent de postedició, que permetrà fer clic sobre un mot sospitós de ser una traducció incorrecta i substituir-lo per altres alternatives tenint en compte el text original i farà possible en general qualsevol canvi del text meta.

Bibliografia

AECMA [Associació Europea d'Indústries Aerospacials] (1998). AECMA Simplified English. `http://www.aecma.org/sebrc.html`.

Alcaraz Varó, E. i Martinez Linares, M. (1997). *Diccionario de lingüística moderna*. Ariel, Barcelona.

Almqvist, I. i Sågvall Hein, A. (1996). Defining ScaniaSwedish — a controlled language for truck maintenance. En *CLAW 96, Proceedings of the First International Workshop on Controlled Language Applications (Leuven)*, pages 159–164.

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., i Sadler, L. (1994). *Machine Translation: An Introductory Guide*. NCC Blackwell, Oxford. Available as `http://clwww.essex.ac.uk/~doug/MTbook/`.

Arnold, D., Sadler, L., i Humphreys, R. (1993). Evaluation: an assessment. *Machine Translation*, 8:1–24.

Chomsky, N. (1996). *The minimalist program*. MIT Press, Cambridge, Massachusetts.

Don, J., Kerstens, J., Ruys, E., i Zwarts, J. (1996). Lexicon of linguistics. `http://www-uilots.let.ruu.nl/~Hans.Leidekker/lexicon/ll.html`.

Flamand, J. (1983). *Écrire et traduire: sur la voie de la création*. Editions du Vermillion, Ottawa.

Forcada, M. (2000). Learning machine translation strategies using commercial systems: discovering word-reordering rules. En *Proceedings of MT 2000 (Exeter, November 2000)*.

Hovy, E. (1993). How MT works. *Byte*, (gener):167–176.

Huijsen, W. (1998). Introduction to controlled languages. `http://www-uilots.let.ruu.nl/Controlled-Languages/faq.html`.

Hutchins, J. (1995). Machine translation: a brief history. En Koerner, E. i R.E.Asher, editors, *The concise history of the language sciences: from the Sumerians to the cognitivists*, pages 431–445. Pergamon.

Hutchins, J. (1996). Evaluation of machine translation and translation tools. En Cole, R., editor, *Survey of the State of the Art in Human Language Technology*. (disponible per internet: `http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html`).

Hutchins, J. (2001). Machine translation over fifty years. *Histoire Epistémologie Langage*, 23(1):7–31.

Hutchins, W. i Somers, H. (1992). *An introduction to machine translation*. Academic Press. (hi ha una traduucció al castellà, *Introducción a la traducción automática*, editada por Visor en 1995).

Ide, N. i Veronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24:1–40.

Jacqmin, L. (1993). Classification générale des systèmes de traduction automatique. En P., B. i Clas, A., editors, *La traductique*. Presses Univ. Montréal, Montréal.

Jakobson, R. (1966). On the linguistic aspects of translation. En Brower, R., editor, *On translation*. Oxford Univ. Press, Oxford.

Krauwer, S. (1993). Evaluation of MT systems: a programmatic view. *Machine Translation*, 8.

Lewis, D. (1997). MT evaluation: science or art? *Machine Translation Review*, 6:25–36.

Lyovin, A. (1997). *Languages of the world*. Oxford Univ. Press, Oxford.

Minnis, S. (1994). A simple and practical method for evaluation machine translation quality. *Machine Translation*, 9:133–149.

Mira i Gimènez, M. i Forcada, M. L. (1998). Understanding PC-based machine translation systems for evaluation, teaching and reverse engineering: the treatment of noun phrases in Power Translator. *Machine Translation Review (British Computer Society)*, 7:20–27. (available at `http://www.dlsi.ua.es/~mlf/mtr98.ps.Z`).

Newton, J. (1992). The Perkins experience. En *Computers in Translation: a practical appraisal*. Routledge, Londres.

Nida, E. (1966). Principles of translation exemplified by Bible translation. En Brower, R., editor, *On translation*. Oxford Univ. Press, Oxford.

Radford, A., Atkinson, M., Britain, D., Clahsen, H., i Spencer, A. (1999). *Linguistics: an introduction*. Cambridge Univ. Press, Cambridge.

Sager, J. C. (1993). *Language engineering and translation: consequences of automation*. Benjamins, Amsterdam.

Samuelson-Brown, G. (1996). New technology for translators. En Owens, R., editor, *The translator's handbook*. Aslib, Londres, 3a. edició.

Somers, H. i Rutzler, C. (1996). Machine translation. En Owens, R., editor, *The translator's handbook*. Aslib, Londres, 3a. edició.

Tuson, J. (1999). *¿Com és que ens entenem? (si és que ens entenem)*. Empúries, Barcelona.

Vandooren, F. (1993). Divergences de traduction et architectures de transfert. En P., B. i Clas, A., editors, *La traductique*. Presses Univ. Montréal, Montréal.

Wojcik, R. i Hoard, J. (1996). Controlled languages in industry. En Cole, R., editor, *Survey of the State of the Art in Human Language Technology*. (disponible per internet: `http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html`).