Manual d'informàtica i de tecnologies per a la traducció

Mikel L. Forcada
Felipe Sánchez Martínez
Juan Antonio Pérez Ortiz

Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant
E-03071 Alacant

{mlf,fsanchez,japerez}@dlsi.ua.es

http://www.dlsi.ua.es/~mlf
http://www.dlsi.ua.es/~fsanchez
http://www.dlsi.ua.es/~japerez

Edició 0.9
Febrer de 2016

Índex

Capítol 1

Introducció

Aquestes pàgines cobreixen la major part dels continguts[1] de l'assignatura *Tecnologies de la Traducció* que cursarà l'alumnat de segon curs del grau en Traducció i Interpretació de la Universitat d'Alacant; també poden ser útils per a assignatures similars en altres universitats (per això s'hi ha inclòs material més avançat que no s'estudia en Tecnologies de la Traducció). La lectura d'aquest manual —que pot fins i tot contenir algun error no detectat— no pot mai substituir l'estudi d'altres llibres sobre la matèria, alguns dels quals se citen en aquest text i es llisten en la bibliografia.

Veureu que els continguts d'aquest manual es poden dividir en dues 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), i la segona és una introducció a alguns aspectes generals de la traducció automàtica (capítols 6 a 9) i a la traducció assistida per ordinador amb memòries de traducció (capítol 10). Finalment, un apèndix discuteix la problemàtica de la traducció espanyol–català i alguns dels sistemes existents per a aquest parell de llengües; aquesta informació pot servir com a il·lustració en un cas concret del que s'ha estudiat sobre traducció automàtica. Els continguts d'aquesta tercera part són, per tant, complementaris.

De fet, aquest manual es pot millorar molt, i n'anirem fent versions noves. A més, és segur que hi deu haver errades que s'han 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 vulguen saber sobre el tema. De fet, aprofitem per donar les gràcies a totes les persones (alumnat, professorat, etc.) que, amb els seus comentaris crítics, han anat millorant aquest text.[2] Són massa gent per a esmentar-los tots, però

[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.

[2] Aquest llibre està basat en una obra anterior, (Forcada i Pérez-Ortiz 2009), usada per a

no volem acabar sense agrair les aportacions de Raül Canals i Marote, que va corregir errades de versions anteriors i va fer aportacions en la part de conceptes bàsics de la informàtica, de Gema Ramírez Sánchez, particularment en el capítol de memòries de traducció, i de Sandra Montserrat, en la discussió sobre divergències lingüístiques espanyol–català de l'apèndix.

Els fitxers font (LaTeX, .eps, etc.) necessaris per a tornar a generar el llibre estan disponibles en un *repositori* públic,[3] de manera que, si ho desitgeu, els podeu modificar per a generar un text nou i publicar-lo vosaltres, però sempre d'acord amb les condicions de la versió 3 de la Llicència General Pública de GNU[4] o de la llicencia Creative Commons Reconeixement-CompatirIgual 4.0 Internacional.[5] Aquestes llicències us obliguen a publicar qualsevol treball derivat d'aquest amb la mateixa llicència. Així garantim que el nostre treball està sempre accessible per a qualsevol persona que el considere útil per a l'ensenyament o l'estudi personal. Si aquest és el vostre cas, us estarem molt agraïts si ens envieu un missatge de correu electrònic dient-nos per a quina assignatura l'esteu usant.

la llicenciatura en Traducció i Interpretació.

[3] Repositori GitHub: `https://github.com/mlforcada/llibre-tecnol-trad`

[4] Descrita en `http://www.gnu.org/licenses/gpl-3.0.txt`

[5] Descrita en `http://creativecommons.org/licenses/by-sa/4.0/`

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, el ratolí, 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 a la persona usuària o per a un altre *programa*. Per exemple, un processador de textos com LibreOffice o Microsoft Word pot estar compost per més d'un *programa*. Un *programa* és una seqüència (llista o conjunt ordenat) d'instruccions que són seguides o executades pel maquinari, de tal manera que realitzen alguna tasca determinada.[3] Normalment, els ordinadors estan organitzats al voltant d'un *processador central* (vegeu més endavant) que és capaç de comprendre i executar instruccions bàsiques preses d'un conjunt determinat (el *conjunt d'instruccions* del processador). Els programes poden estar guardats en un disc o carregats en la memòria de l'ordinador mentre són executats pel processador.

A continuació es consideren el maquinari i el programari amb més detall.

[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.

[3] L'ús de la paraula *programa* en informàtica (seqüència d'operacions o esdeveniments) és paral·lel a molts usos d'aquest mot en la vida quotidiana: programa *de festes*, *d'un concert*, *de la llavadora*, etc.; encara que per a la persona usuària un programa d'ordinador és més similar a una espècie de caixa d'eines per a fer una tasca determinada, com, per exemple, editar un document de text.

2.1 Maquinari

Tots els sistemes informàtics tenen maquinari de les classes següents:

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 el 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. A més, la majoria dels sistemes actuals contenen també una GPU (*graphic processing unit*, unitat de processament de gràfics), una CPU especialitzada en el tractament d'imatges però que també es pot usar per a altres tasques computacionalment molt intensives.

La velocitat a la que una CPU executa les instruccions bàsiques d'un programa es mesura en megahertzs (MHz) o gigahertzs (GHz; un gigahertzs són 1000 megahertzs). Un megahertzs equival a un milió de hertzs (Hz), és a dir, un milió de cicles de processament d'informació per segon. Cada cicle de processament d'informació es correspon amb un *tic* del rellotge que tots el dispositius de processament tenen per sincronitzar tots el circuits de l'ordinador. Normalment una instrucció requereix d'uns pocs cicles de processament per ser executada, tot i que alguns sistemes són capaços de processar més d'una instrucció al mateix temps. La velocitat típica de la CPU d'un ordinador en l'actualitat és de 3 Ghz.

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; si els programes i les dades no caben en la memòria primària, el sistema operatiu —vegeu l'apartat 2.2— s'encarrega de copiar-los de la memòria al disc dur quan no s'estan usant i copiar-los de tornada del disc dur a la memòria quan són necessaris, operació que s'anomena *intercanvi*.[4] La memòria primària normalment consisteix en xips RAM (*random-access memory*, memòria d'accés aleatori[5]) de silici.

[4] En anglés *swapping*. Com que el disc dur és més lent que la memòria primària, l'intercanvi fa que l'ordinador vaja més lent; per això, ampliar la memòria primària sol fer que l'ordinador vaja més ràpid.

[5] Noteu la diferència entre *accés aleatori* (a voluntat) i *accés seqüencial*. Un CD-ROM de

Memòria secundària: memòria de llarg termini, permanent. Exemples: els antics disquets, discos fixos o durs interns i externs, memòries USB (també anomenades llapis o *pendrive*) i diverses formes de ROM (*read-only memory*, memòria de lectura només), com els xips ROM, els CD-ROM o els DVD.

Els discos fixos (i els antics disquets) són dispositius d'emmagatzematge magnètic, poc més o menys com ho eren les antigues cassets. La informació s'emmagatzema fent servir les propietats magnètiques de determinats materials magnetitzables. En la actualitat la grandària típica d'un dic fix és de 500 GB o 1 TB (vegeu l'apartat 2.3 per a assabentar-vos de les mesures d'emmagatzemament de la informació).

La memòria USB és un dispositiu de memòria flaix, un xip de memòria que manté el seu contingut en absència d'alimentació, que es connecta al port USB de l'ordinador. La grandària d'aquestes memòries pot arribar fins a 1 TB, tot i que les grandàries més típiques són 16, 32 i 64 MB.

La memòria ROM sol estar feta de xips de silici. Els CD-ROM (*compact dics read-only memory*) —idèntics en aparença i similars en molts aspectes als CD de música— emmagatzemen la informació òpticament.[6] La grandària d'un CR-ROM sol ser de 650 MB o de 700 MB.

El DVD (*digital versatil dics*)[7] és un tipus més avançat de sistemes d'emmagatzematge basat en discs òptics; bàsicament, es tracta d'un CD més ràpid i amb més capacitat, que ha desplaçat quasi completament els CD-ROM. La grandària d'un DVD depén del tipus de DVD i sols estar entre 4,7 GB i 17 GB.

La manera més comuna d'emmagatzemar les dades en memòria secundària és organitzar-les en *fitxers* o *documents* organitzats en *directoris* o *carpetes*; la secció 2.4 explica aquests conceptes amb detall.

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 finalitat d'*introduir-hi* dades o informació. Els dispositius d'entrada més comuns són el teclat, el ratolí, la pantalla tàctil, la maneta de jocs,

música és d'accés aleatori perquè podem accedir a la setzena cançó directament; en canvi, un casset (cinta magnètica) és d'accés seqüencial perquè per a accedir a la setzena cançó hem de passar per les 15 cançons anteriors.

[6] Altres termes habituals són CD-R (*compact disc recordable*) —que identifica els CD en què es pot escriure informació només una vegada amb l'ajuda de dispositius coneguts com a enregistradores— i CD-RW (*compact disc rewritable*) —utilitzat per als CD que poden ser esborrats i reescrits un nombre il·limitat de vegades.

[7] Com en el cas dels CD, podem parlar de DVD-ROM, DVD-R i DVD-RW.

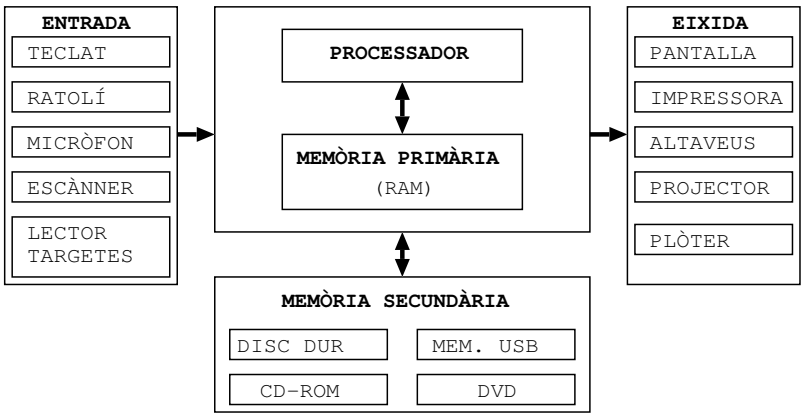


Figura 2.1: Esquema del maquinari d'un ordinador.

les càmeres de fotos i *webcams* o l'escànner —un dispositiu que llegeix una imatge impresa i la converteix en un fitxer (vegeu l'apartat 2.4) que conté la imatge digitalitzada.[8]

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, els altaveus, els vibradors dels dispositius mòbils, etc.

En la figura 2.1 es resumeix esquemàticament el maquinari d'un ordinador.

2.2 Programari

Hi ha tres classes bàsiques de programari:

Sistemes operatius i *firmware*: són els programes que permeten el funcionament bàsic de l'ordinador. S'anomena *firmware* el programari del sistema que s'usa tan freqüentment que s'emmagatzema permanentment en xips ROM. Aquest programari oferix al sistema operatiu serveix bàsics d'accés als dispositius d'entrada i d'eixida més habituals.

Quan connectem l'ordinador, el primer programa a executar-se és el *firmware*, el qual s'encarrega de fer algunes comprovacions, com ara

[8] 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 4.6), generalment amb alguns errors tipogràfics menors.

que hi ha un teclat enganxat a l'ordinador o que la memòria RAM no te defectes, i de carregar el sistema operatiu.

El sistema operatiu, d'una banda, permet que la persona usuària hi execute programes i gestione els fitxers de dades, etc.; per a això, ofereix una *interfície d'ús* (vegeu més endavant). D'altra banda, el sistema operatiu ofereix serveis bàsics (vegeu més avall) als programes d'aplicació que s'executen en l'ordinador (els quals poden tenir la seua pròpia interfície d'ús).

Quant a la *interfície d'ús*, és a dir, l'aparença i la forma d'interaccionar amb l'usuari, la majoria dels sistemes operatius són *gràfics*, és a dir, basats en ratolí o pantalla tàctil, punters, finestres, etc. (GNU/Linux, Windows, MacOS, iOS, Android); antigament, els sistemes operatius eren de *línia d'ordres*, és a dir, basats en text (Unix primigeni, MS-DOS).

Els sistemes més antics eren de vegades *monousuari* (MS-DOS, Windows 3.11), és a dir, només podien donar suport a una persona usuària, o *monotasca*, és a dir, no podien executar més d'un programa al mateix temps. La major part dels actuals sistemes operatius són *multiusuari*, és a dir, poden donar accés i suport a més d'una persona usuària alhora, i *multitasca* (GNU/Linux, versions recents de Windows, MacOS). La major part dels sistemes operatius actuals estan a més preparats per a interaccionar amb altres dispositius través de diferents tipus de *xarxes*.[9]

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

- Controlar el maquinari de l'ordinador on s'executen.
- 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.
- Establir connexions amb altres ordinadors o dispositius en xarxa.

De fet, els programes d'aplicació solen estar escrits per a ser executats *sobre un sistema operatiu*, és a dir, els programes d'aplicació *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,

[9] L'organització dels ordinadors en una xarxa local 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.

cosa que simplifica enormement l'escriptura dels programes per part dels programadors. Per això, quan s'especifiquen les característiques d'un programa d'ordinador s'ha de dir per a quin sistema operatiu està escrit, ja que cada sistema operatiu ofereix serveis diferents i interacciona de manera diferent amb els programes d'aplicació.

Programes d'aplicació: programari dissenyat específicament per a satisfer les necessitats dels usuaris (de vegades s'anomenen simplement *aplicacions*). 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, fet a mida per a ell.

Programari específic per a professionals de la traducció: sistemes de traducció automàtica (capítols 6 a 9), sistemes de traducció assistida basats en memòries de traducció (capítol 10) i bases de dades terminològiques (capítol 5).

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

Editors i processadors de text per a preparar, modificar, emmagatzemar i imprimir documents de text (vegeu la secció 4.6).

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).

Navegadors d'Internet: programes que permeten accedir de manera senzilla als documents d'Internet en màquines connectades a aquesta xarxa.[10] Vegeu la secció 3.5.

Jocs de moltes classes.

Els programes d'aplicació els activa l'usuari per mitjà del sistema operatiu, utilitzen el sistema operatiu per a accedir als recursos (maquinari i altres programes) del sistema i interaccionen amb l'usuari mitjançant el dispositius d'entrada i d'eixida (vegeu la figura 2.2).

[10]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.

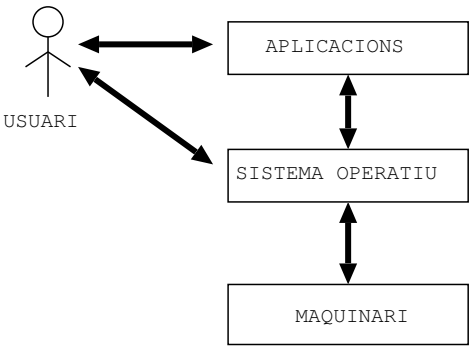


Figura 2.2: Esquema de la interacció entre la persona usuària, el sistema operatiu i els programes d'aplicació.

Per saber més sobre 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 d'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
```

Processadors de llenguatges de programació: les instruccions que executa el processador central d'un ordinador són massa senzilles perquè 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, Java, C, C++, Pascal, Perl o Python) i usen programes especials —els processadors de llenguatges— per a traduir-los a les instruccions senzilles que entén la màquina.[a] 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 quedaria així en el llenguatge Pascal:

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

[a] Hi ha dues 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.

2.3 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,[11] 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 diferents, però amb 10, ja en podem fer suficients, perquè $2^{10} = 1\,024$ (en quedarien $1\,024 - 778 = 246$ 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 es guardaven històricament 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 va ser el primer codi estàndard 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 que executa el processador central també s'emmagatzemen en octets.[12]

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.
- *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:

$$1 \text{ GB} = 1.024 \text{ MB} = 1.048.576 \text{ kB} = 1.073.741.824 \text{ octets}.$$

[11] Com, per exemple, els signes d'algun sistema d'escriptura no alfabètic

[12] En l'exemple de la secció anterior, la instrucció `inc A`, que incrementa el valor emmagatzemat l'acumulador en 1, podria ser l'octet `11010110`

- *teraoctets* o *terabytes* (TB), o bilions (milions de milions) —de nou, una mica més— d'octets:

$$1\ \text{TB} = 1.024\ \text{GB} = 1.048.576\ \text{MB} = 1.073.741.824\ \text{kB} = \\ = 1.099.511.627.776\ \text{octets}.$$

Com que els prefixos *k*, *M*, *G* i *T* s'usen en la resta de les disciplines científiques per expressar a potències exactes de 10 (de 1.000), hi ha qui prefereix parlar de *kibioctets* o *kibibytes* (kiB), *mibioctets* o *mibibytes* (MiB), *gibioctets* o *gibibytes* (GiB) i *tebibytes* o *tebioctets* (TiB) per a referir-se a les unitats de capacitat d'emmagatzematge basades en múltiples de 1.024.

2.4 Fitxers i directoris

Com ja s'ha dit en la pàgina 5, és comú que les dades —de qualsevol classe: textos, instruccions de programa, dades gràfiques, de so, de vídeo, etc.— emmagatzemades en memòria secundària estiguen organitzades en *fitxers*, també anomenats *documents* o *arxius*. Els fitxers són conjunts de dades amb un nom que els identifica i que es manipulen —s'obrin, es tanquen, es copien, s'esborren— com un tot. En discos grans, seria molt incòmode tenir tots els fitxers un darrere l'altre; per això, és comú que els fitxers estiguen organitzats en *directoris*, també anomenats *carpetes*. Els directoris són fitxers especials que agrupen els noms i les característiques d'altres fitxers; de fet, els directoris poden contenir zero o més fitxers o també zero o més directoris (sense restriccions de quantitat), i així successivament, de manera que la persona usuària pot establir una estructuració jeràrquica o arbòria dels seus fitxers en el disc.

Normalment, cada disc té un *directori principal* o *directori arrel* (el més elevat en la jerarquia de directoris), dins del qual es troba tota la resta 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, `alacant.txt`). El nom sol ser normalment lliure, però l'extensió sol ser curta (entre una i quatre lletres) i el sistema operatiu la sol usar per a 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 fitxer de text pla, vegeu l'apartat 4.1; l'extensió `.exe` s'usa per als programes d'ordinador, etc.).

La seqüència dels noms de les carpetes que cal anar obrint fins que arribem a un fitxer s'anomena la *trajectòria* o la *ruta* del fitxer. De fet, convé considerar la trajectòria com a part del nom del fitxer, cosa que ens permetria dir, senzillament, que en un disc no pot haver-hi dos fitxers amb el mateix nom.

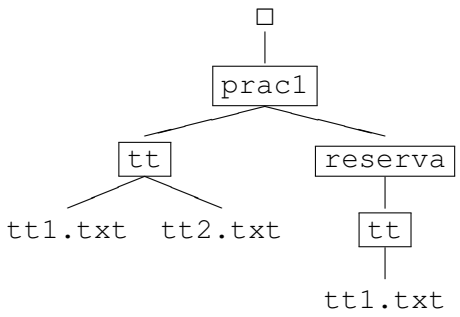


Figura 2.3: Exemple d'estructura de fitxers i directoris en un dispositiu d'emmagatzemament. El directori principal o arrel està representat pel símbol □.

Tots aquests conceptes es veuen potser més clars amb l'exemple de la figura 2.3 en què es mostra l'estructura de fitxers i directoris en un dispositiu d'emmagatzemament qualsevol. En aquest dispositiu, el directori principal o arrel (representat amb el símbol □) conté un únic (sub)directori `prac1`; aquest directori conté dos (sub)directoris, `tt` (que conté els fitxers `tt1.txt` i `tt2.txt`) i `reserva`. El directori `reserva` conté un subdirectori `tt` (que conté l'arxiu `tt1.txt`). Fixeu-vos que dues carpetes diferents contenen arxius amb el mateix nom `tt1.txt`; això no és problema si considerem la trajectòria completa com a nom del fitxer. Si el disc es diu `C:` (típic en el cas del disc dur d'un PC amb sistema operatiu Windows), les trajectòries d'aquests dos fitxers serien `C:\prac1\tt\tt1.txt` i `C:\prac1\reserva\tt\tt1.txt`, i, per tant, serien diferents. En el cas d'el sistema operatiu GNU/Linux les trajectòries d'aquests dos fitxers serien `/prac1/tt/tt1.txt` i `/prac1/reserva/tt/tt1.txt`. Fixeu-vos que cada sistema operatiu fa servir un símbol diferent per al directori principal o arrel i per a separar els noms dels directoris i arxius dins de la ruta o trajectòria.

2.5 Tipus d'ordinadors

Una classificació no gaire exhaustiva dels diferents tipus d'ordinadors que podem trobar-nos avui dia és la següent:

De sobretaula (en anglés *desktop*): Estan formats per una *caixa* amb dispositius de processament i emmagatzematge i un conjunt de perifèrics (dispositius d'entrada o d'eixida) com ara el teclat, el ratolí o la pantalla. Són, amb diferència, els ordinadors més habituals.

Portàtils (en anglés *laptop*, encara que quan són menuts se'n diu de vega-

des *notebook* o *netbook*): Tenen una grandària menor que la d'un maletí i un pes lleuger que permet dur-los sense massa esforç d'un lloc a un altre. Però, el seu volum reduït limita les possibilitats de fer-hi ampliacions i, per tant, el seu temps de vida pot ser més curt que el dels ordinadors de sobretaula.

Tauletes i *smartphones* : les tauletes (en anglés *tablets*) i els telèfons mòbils més moderns, anomenats *smartphones*, són veritables ordinadors portàtils, amb una pantalla tàctil i sense teclat, amb càmera, connectivitat Wi-Fi i Bluetooth, receptor GPS, etc. Solen venir amb un sistema operatiu gràfic: el més comú és Android, però els de la marca Apple usen iOS.[13]

Servidors: Els servidors són ordinadors que contenen i gestionen informació que s'utilitzarà en altres ordinadors ("clients") connectats a ells a través d'una xarxa interna o a través d'Internet; no són massa diferents d'un ordinador de sobretaula, típicament més potents quant a memòria, disc i processador, però, com que ningú ha de seure davant d'ells no solen tenir pantalles, teclats o ratolins i sovint estan pensats per a ser disposats horitzontalment en armaris especials anomenats *racks*. Aquests ordinadors es poden presentar en grups connectats entre si per a oferir major potència i capacitat.

Quant als ordinadors de taula i els portàtils, sovint es fa la distinció entre els ordinadors de tipus PC i els Macintosh (sovint anomenats Mac). Els PC són l'evolució dels primers ordinadors personals desenvolupats per IBM, tot i que actualment són fabricats per un nombre molt gran d'empreses. Els Mac, a hores d'ara també són ordinadors de tipus PC, però antigament eren ordinadors tipus PowerPC fabricats exclusivament per l'empresa Apple. Els Mac fan servir un sistema operatiu propi (MacOS) i tenen una quota de mercat més reduïda entre el públic general però més gran en determinades aplicacions especialitzades (per exemple, el disseny gràfic).

2.6 Configuració típica d'un ordinador personal

La configuració clàssica d'un ordinador personal de sobretaula model 2015 sol ser més o menys com segueix:

- La unitat base (la "caixa" o la "torre") conté:

[13] Aquests dispositius han desplaçat els antics *handhelds* o dispositius de mà, que eren una evolució de les antigues agendes electròniques i se solien anomenar *PDA* per *Personal Digital Assistant*, 'assistent digital personal'.

  - Un processador compost de quatre nuclis o processadors individuals (*quad-core*) o més, com ara un *Intel Core i5* o més o un processador equivalent de la marca AMD (vegeu el glossari, apartat 2.7) a 3 GHz.
  - La memòria RAM (per exemple, de 8 GB).
  - Una bona targeta gràfica, amb la seua pròpia unitat independent de processament gràfic o GPU (*graphics processing units*),
  - Un disc fix amb una capacitat de l'ordre d'1 TB
  - Una unitat enregistradora de DVD i de CD-ROM[14]
  - Una targeta de so amb altaveus i micròfon.
  - Una càmera (de vegades anomenada *webcam*)
  - Una o més targetes de comunicacions incorporades (amb fils o sense fils, vegeu el glossari, apartat 2.7).

- Un monitor o pantalla, normalment una pantalla LCD[15] de 17 o més polzades,
- Un teclat separat i un ratolí.
- Una impressora (d'injecció o de raig de tinta —la més típica—, o làser[16]).

Les especificacions dels portàtils (memòria, processador) solen ser similars, normalment una miqueta més reduïdes. Els telèfons mòbils intel·ligents o *smartphones* i les tauletes no solen tenir disc, sinó una memòria flaix no volàtil, per exemple, de 8 GB, i una memòria RAM de l'ordre d'1 GB.

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.

adaptador de vídeo (també anomenada targeta gràfica o controlador de vídeo): Dispositiu (targeta independent, o integrada en la placa base) 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

[14] 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 tinga una velocitat punta de lectura i d'escriptura de 52× o més. De tota manera, les velocitats *mitjanes* de tot un procés de lectura i escriptura solen ser més baixes.

[15] *liquid-crystal display* o pantalla de cristall líquid

[16] Les impressores *matricials* o *d'agulles* només s'usen per a aplicacions molt específiques.

nombre de punts, elements d'imatge (*píxels*) que caben en una imatge, per exemple 1024×768 (horitzontal $\times$ vertical), la *profunditat de color* (en bits: per exemple 24 bits permeten $2^{24} = 16\ 777\ 216$ colors diferents) i altres paràmetres com la *freqüència de refrescament* (que es mesura en hertzs o cicles per segon; vegeu "megahertz"). Actualment no és estrany tenir en ordinadors de taula o portàtils resolucions com l'anomenada *HD 1080* (1920 $\times$ 1080) o fins i tot més grans.

ADSL (de l'anglés *asymmetric digital subscriber line*, línia d'abonat digital asimètrica): Versió asimètrica de DSL (vegeu DSL). L'asimetria fa referència al fet que la velocitat de transmissió de dades de la central cap a l'abonat és superior que la velocitat de transmissió de dades de l'abonat cap a la central (per exemple, 8 Mb/s cap a l'abonat i 512 kb/s cap a la central).

cache o *memòria cau*: Memòria RAM intermèdia, d'accés més ràpid per part del processador, on es copia de tant en tant 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 *memòria cau*, 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 la *memòria cau* 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.

DSL (de l'anglés *digital subscriber line*, línia d'abonat digital), tecnologia de connexió que permet aprofitar les línies telefòniques i elèctriques per a fer connexions d'alta velocitat (fins a uns 10 Mb/s). En el cas d'usar les línies elèctriques, la tecnologia rep també el nom de PLC (*power line communications* o comunicacions a través de les línies de força), però a Espanya no s'usa per a proveir serveis d'Internet a les llars.

fibra òptica: tecnologia que transporta les dades usant una llum làser que es propaga a través d'un fil molt fi de material transparent. En el moment d'escriure aquestes línies, els proveïdors d'Internet han començat a oferir un servei domèstic de connexió que permeten connexions de l'ordre de centenars de Mb/s.

GHz: vegeu gigahertz.

gigahertz: un gigahertz són 1000 megahertzs (vegeu *megahertz* en aquest glossari).

GNU/Linux: un sistema operatiu multitasca i multiusuari gratuït, de l'estil de l'Unix que es podia trobar en els anomenats *miniordinadors* dels

anys 70 i 80, desenvolupat de manera col·laborativa per milers de voluntaris independents i per empreses arreu del món i que és *programari lliure* (vegeu l'entrada en aquest glossari): es pot copiar lliurement si es compleixen certes condicions. Es pot instal·lar GNU/Linux (que es presenta en moltes *distribucions* diferents com ara *Ubuntu*, *Mint*, *Fedora*, etc.) en un PC amb processador de la família x86 (vegeu *Pentium*) 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. Aquests ordinadors, llançats al mercat el 1984, popularitzaren la interfície gràfica d'usuari, tota una revolució per a l'època. Fa uns anys hi havia diferències significatives entre els PC i els *Mac* de manera que no eren compatibles, és a dir, que els programes d'un no funcionaven en l'altre, s'havien d'adaptar a les característiques particulars de cadascun. Aquestes diferències eren degudes al fet que el processador del *Mac* no era de la família x86 (vegeu *Pentium*), sinó d'una altra (antigament la família 68000 de Motorola, i després l'anomenat PowerPC). En l'actualitat aquesta diferència no existeix i tant uns com altres empren processadors de la família x86. En el cas dels *Mac* des de l'any 2006 incorporen processadors Intel, així que podem instal·lar-hi Microsoft Windows o GNU/Linux amb tots els seus programes, encara que també podem usar el sistema operatiu propi dels Mac, anomenat *MacOS*.

megahertz: Un megahertz (MHz) és un milió d'hertzs (Hz), és a dir, un milió de cicles per segon. La velocitat de les unitats centrals dels ordinadors es mesura en MHz i més recentment en GHz, és a dir, en milions o milers de 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. Els models actuals poden executar, en determinades circumstàncies, més d'una instrucció al mateix temps, el que fa que de vegades s'execute una instrucció per cicle de rellotge o fins i tot més d'una. Una velocitat típica en l'actualitat és 3 GHz, és a dir, 3000 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). Altres components, com ara la memòria RAM, també funcionen a una determinada velocitat, independent de la del processador, que es mesura també en MHz.

MHz: vegeu megahertz.

mòdem: abreviatura de modulador-desmodulador. En el cas del mòdem

més comú a finals del segle passat, el *mòdem telefònic*, es tractava d'un dispositiu (normalment una placa interna, encara que també pot ser extern) que permetia usar la línia telefònica (senyals analògics) per a comunicacions informàtiques (digitals) entre dos ordinadors, establint una telefonada; era aleshores la manera estàndard d'accedir a Internet des de casa. Un dels paràmetres més interessants d'un mòdem és la *velocitat* de transmissió de dades, que es mesura en b/s (bits per segon). Una velocitat clàssica en mòdems domèstics era 33.600 b/s (més recentment, 57.600 b/s; les línies telefòniques actuals poden admetre potser velocitats al voltant dels 100.000 b/s). Això permetia enviar una carta d'una pàgina en unes dècimes de segon però no seria suficient per a la major part dels usos actuals d'Internet.

La paraula *mòdem* es pot usar també per a altres tipus de mòdems, normalment més ràpids: els *mòdems de cable*, que permeten connectar l'ordinador a Internet a través dels cables d'empreses especialitzares que ofereixen televisió, telèfon i Internet, els *mòdems ADSL* (vegeu *ADSL* en aquest glossari), els *mòdems de fibra òptica* (vegeu *fibra òptica* en aquest glossari), etc.

Programari lliure: (*free software*, també anomenat *programari de codi font obert* o *open-source software*) és el programari que es distribueix amb llicències que donen una sèrie de llibertats a qui rep el programari: la llibertat d'usar-lo per a qualsevol propòsit sense restricció, la llibertat d'examinar-lo per veure com funciona i modificar-lo per adaptar-lo a un nou ús, i la llibertat de distribuir còpies —originals o modificades— lliurement a qui desitgem. Per a poder modificar el programari, no hi ha prou amb tenir accés a la versió executable en l'ordinador: hem de tenir accés a l'anomenat *codi font*, és a dir, a la versió del programari que escriuen i modifiquen les persones que programen (d'ací el nom *de codi font obert*) i que després es converteix automàticament en la versió executable. Exemples de programari lliure són: el sistema operatiu *GNU/Linux*, el navegador *Firefox*, o el processador de textos *LibreOffice*. No s'ha de confondre *programari lliure* amb *programari gratuït* (o *freeware*). Hi ha programari gratuït que no és lliure perquè no atorga totes les llibertats (per exemple, el lector de PDF *Adobe Acrobat*, o el programa de telefonia per Internet *Skype*: per exemple, tot i tenir el programari executable, no tenim accés al seu codi font).

Pentium: nom genèric d'una família actual de processadors centrals de la companyia Intel, els més recents de la sèrie "x86" de processadors que començà amb el 8086 a principis dels anys 80, passant pel 80286, el (80)386 i el (80)486.[17]. Els nous processadors tenien jocs d'instruccions

[17]El nom *Pentium* es va triar perquè Intel no podia registrar "586" com a marca.

més complexos i eren capaços d'executar els programes que executaven els anteriors (per exemple, un Pentium pot executar qualsevol programa escrit per a un 386) però introduïen millores que permeti-en ordinadors més ràpids, amb capacitat més gran de càlcul, capaços de processar més dades en cada instrucció (8, 16, 32 —a partir del 386—, i actualment 64 bits) i de gestionar més memòria. Els Pentium més recents tenen més d'un *nucli* o sub-processador, i poden, per tant, executar instruccions de programa en paral·lel.

placa de so: En els ordinadors més antics, s'havia de comprar a banda una placa (o targeta) de so si es volia usar l'ordinador per a processar, enregistrar, reproduir, i manipular sons digitalitzats. En l'actualitat tots els ordinadors porten aquestes capacitats incorporades.

targeta de xarxa: En els ordinadors més antics, per a connectar ordinadors i formar una xarxa (normalment local) per a compartir recursos, calia dotar a cada ordinador d'una placa o targeta de xarxa. Hi ha diversos estàndards de connexió en xarxa; els més anomenats són Ethernet (per a connexions amb fils) i *Wi-Fi* (vegeu *Wi-Fi* en aquest glossari).

USB (de l'anglés *universal serial bus*, bus sèrie universal): Estàndard o norma de connexió de dispositius perifèrics (impressores, mòdems, reproductors digitals de música, càmeres digitals, unitats de memòria) que transmet les dades en sèrie (és a dir, un bit darrere de l'altre) a velocitats que en les versions més modernes de l'estàndard poden arribar als Gb/s, i que permet la connexió i desconnexió de dispositius de moltes classes "en calent", és a dir, sense haver d'apagar l'ordinador.

Wi-Fi (probablement de l'anglés *wireless fidelity*, fidelitat sense fils): tecnologia de connexió sense fils (via ràdio), principalment per a formar xarxes locals, i que en l'actualitat (estàndard IEEE 802.11ac, gener de 2014) permet velocitats de transmissió de fins a 6.77 Gb/s.

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 una memòria USB té 6 *gigabytes* (gigaoctets), una pàgina de text (europeu occidental) típica té 50 línies de 60 caràcters (contant els blancs)

i cada caràcter ocupa 1 *byte* (octet), 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 300 kilooctets (*kilobytes*) per segon. Una d'aquestes tres *no* pot ser la velocitat del seu servei d'ADSL:

   (a) 1 Mb/s

   (b) 6 Mb/s

   (c) 4 Mb/s

4. Quina d'aquestes afirmacions és incorrecta?

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

   (b) Les velocitats típiques de connexió a Internet via ADSL són d'uns quants Mb/s.

   (c) El servei ADSL aprofita les línies de telefonia convencional per a oferir connexió a Internet.

5. Es podria enregistrar (guardar) en un CD-ROM tota la informació continguda en un instant determinat en la memòria RAM d'un ordinador vell que en té 512 MB?

   (a) Sí.

   (b) No, perquè no hi cap.

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

6. Quina d'aquestes afirmacions es certa?

   (a) En qualsevol dispositiu d'emmagatzemament (disc dur, CD-ROM, memòria USB) sempre hi ha un directori principal o arrel.

   (b) Un disc no pot contenir més de dos nivells de jerarquia de carpetes.

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

7. 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.

8. Quants valors possibles pot prendre un octet o *byte*?

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

9. Quin dels tres mitjans d'emmagatzemament següents no és òptic:

   (a) Un CD-ROM
   (b) Un DVD
   (c) Un disc fix.

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

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

11. 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.

12. Quines són les característiques de la memòria RAM d'un ordinador de taula?

    (a) és lenta, volàtil i d'accés aleatori.
    (b) és ràpida, volàtil i d'accés aleatori.
    (c) és ràpida, permanent i d'accés seqüencial.

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

    (a) Sí, si els ordinadors estan connectats formant una xarxa local.
    (b) Només si la impressora està connectada a Internet.
    (c) Sí, instal·lant-li un mòdem ADSL a la impressora.

14. Cada punt d'una pantalla pot tenir 256 colors: quants octets (*bytes*) de memòria ocupa cada punt?

(a) 1

(b) 256

(c) 8

15. Quan el processador central està executant un programa, on espera trobar la següent instrucció?

    (a) En el CD-ROM.

    (b) En el disc dur.

    (c) En la memòria RAM.

16. És possible posar un fitxer de text en el directori (carpeta) arrel?

    (a) Sí, com en qualsevol directori.

    (b) Només si és un fitxer propi del sistema operatiu.

    (c) No, primer s'hi ha de crear una carpeta (un directori).

17. En la Universitat d'Alacant hi ha al voltant de 35.000 alumnes. Si assignem un número a cada alumne, quants octets (*bytes*) fan falta per a guardar el número de cada alumne?

    (a) 15

    (b) 2

    (c) 3

18. Pràcticament tots els programes necessiten fer operacions bàsiques com ara obrir i tancar arxius o gestionar el ratolí i la pantalla. Vol dir això que tant un navegador com un processador de textos com una memòria de traducció contenen instruccions de programa per a executar aquestes operacions bàsiques?

    (a) No, només instruccions per a invocar els corresponents programes del sistema operatiu.

    (b) Sí, perquè formen part del processador central.

    (c) Sí, perquè, si no, no les podrien executar.

19. Dins d'una carpeta (directori) podem posar carpetes i documents (fitxers) mesclats?

    (a) No. Si una carpeta està dividida en subcarpetes, no pot contenir documents; els documents haurien d'anar dins de les subcarpetes

    (b) Només en la carpeta arrel.

    (c) Sí.

20. Quanta memòria ocupa una imatge de 1024×1024 punts on cada punt pot tenir 8 colors?

    (a) 1 megaoctet (*megabyte*)

    (b) 384 kilooctets (*kilobytes*)

    (c) 8 megaoctets (*megabytes*)

21. Alguns ordinadors portàtils estan dissenyats de manera que, quan les bateries estan a punt d'esgotar-se (o l'ordinador no s'està usant), copien *tota* la memòria RAM al disc dur i s'apaguen. Si tornem a carregar les bateries i encenem l'ordinador, fan l'operació inversa. Podem esperar que l'execució dels programes continue en el mateix punt on es trobava quan les bateries van fallar?

    (a) No, perquè la memòria RAM s'esborra quan falta l'alimentació elèctrica.

    (b) No, perquè només s'hi ha guardat el sistema operatiu.

    (c) Sí, perquè els programes en execució i les seues dades estaven tots en la memòria RAM (si no eren ja al disc).

22. Pot un fitxer contenir les instruccions d'un programa executable?

    (a) No.

    (b) Només si està escrit en un llenguatge de programació d'alt nivell, perquè només així serà un text i podrà guardar-se en un fitxer.

    (c) Sí.

23. Si les imatges enviades per una vella càmera digital sense colors (blanc i negre) tenen 100×100 píxels, quantes d'aquestes imatges podríem emmagatzemar en una memòria USB d'1 GB?

    (a) Depenent de la codificació escollida per als caràcters de la imatge, entre 400 000 i 800 000.

    (b) Unes 10 000.

    (c) Unes 800 000.

24. Quina característica és comuna a tots els tipus de programari?

    (a) Que comencen a executar-se en connectar l'ordinador.

    (b) Que consisteixen en una llista d'instruccions executables.

    (c) Que s'encarreguen de la gestió de tots els recursos del maquinari de l'ordinador on s'executen.

25. És possible que un fitxer de text i la carpeta en què està inclòs tinguen el mateix nom?

    (a) Només si el fitxer ha estat creat pel sistema operatiu.

    (b) Només si es tracta del directori arrel.

    (c) Sí, no importa el nom de la carpeta.

26. Quants bits necessitem per codificar un número de telèfon de 9 xifres suposant que codifiquem els dígits un a un?

    (a) 27

    (b) 36

    (c) 9

27. La targeta Compact Flash on s'emmagatzemen les fotografies d'una càmera digital té 2 GB. Si suposem que hem triat una resolució i un format d'imatge que fa que cada fotografia necessite un espai de 2.048 kB, quantes targetes d'aquestes hem de comprar si volem fer 2.500 fotos al llarg d'un viatge?

    (a) 1

    (b) 3

    (c) 4

28. El *sistema operatiu* d'un ordinador és...

    (a) ... maquinari (*hardware*).

    (b) ... programari (*software*).

    (c) ... una manera d'especificar el format dels textos.

29. Si reduïm de 3.000 MHz a 1.500 MHz la freqüència del rellotge d'un ordinador i encara funciona...

    (a) ... executarà els programes a la mateixa velocitat.

    (b) ... executarà els programes més lentament.

    (c) ... tardarà menys a executar els programes.

30. Windows usa les *extensions* dels noms de fitxers per a...

    (a) ... associar-los el programa que els obrirà quan fem doble clic sobre la icona del fitxer.

    (b) ... estalviar espai quan es guarden els fitxers.

    (c) ... saber si estan buits o contenen text.

31. Un mòdem és un dispositiu que...

(a) ... converteix la informació digital en senyals analògics.

(b) ... converteix senyals analògics en informació digital.

(c) ... fa les dues coses.

32. Indiqueu quina de les afirmacions següents és certa:

(a) La memòria RAM emmagatzema programes i dades mentre s'executen els programes.

(b) La memòria RAM és permanent i més ràpida que la memòria secundària.

(c) Les altres dues afirmacions són falses.

33. Els programes d'aplicació ...

(a) ...sempre interactuen directament amb el maquinari de l'ordinador.

(b) ...els posa en execució el sistema operatiu.

(c) ...no poden comunicar-se amb altres aplicacions.

34. Indiqueu quina de les afirmacions següents és falsa:

(a) Un *gigabyte* o gigaoctet equival a 1.024 kilooctets o *kilobytes*.

(b) Un octet o *byte* equival a *8 bits*.

(c) Un kilooctet o *kilobyte* equival a 1024 octets o *bytes*.

35. La informació s'emmagatzema en la memòria de l'ordinador en forma binària. Què vol dir això?

(a) Que cada dada és una seqüència d'*octets* i cada octet pot adoptar 512 valors.

(b) Que cada dada és una seqüència de nombres codificats en ASCII.

(c) Que cada dada és una seqüència de *bits*, cadascun dels quals només pot adoptar dos valors.

36. Quants *bits* fan falta per representar els 12 mesos de l'any?

(a) 4 *bits* i sobren 4 combinacions.

(b) 12 *bits*, un per cada mes de l'any.

(c) 6 *bits*, un per cada dos mesos.

37. Indiqueu quina de les afirmacions següents és falsa. Els programes d'aplicació ...

(a) ...solen estar escrits per ser executats sobre un sistema operatiu concret.

(b) ... accedeixen als recursos i dispositius connectats a l'ordinador a través del sistema operatiu.

(c) ... mai necessiten del sistema operatiu una vegada que han començat a executar-se.

2.9 Solucions

1. (c): Un gigaoctet té 1.024 megaoctets, i un megaoctet, 1.024 kilooctets: $1.024 \times 1.024 = 1.048.576$.

2. (c): Una memòria USB de 6 GB (gigaoctets o gigabytes) conté aproximadament 6.000.000.000 octets. Un carácter, en les codificacions usades comunament en Europa occidental, 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 300 kilooctets per segon equival a uns $300 \times 8 = 2.400$ kilobits per segon; al voltant 2,3 Mb/s.

4. (a): Els mòdems modulen (converteixen senyals digitals a analògics) i desmodulen (converteixen senyals analògics en digitals) per a enviar i rebre dades a través d'un determinat mitjà. Les línies ADSL domèstiques actuals admeten connexions via mòdem telefònic d'uns quants megabits per segon (vegeu la secció 2.7).

5. (a): Un CD-ROM pot emmagatzemar com a mínim 650 MB.

6. (a)

7. (c)

8. (b): $2^8 = 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 256$.

9. (c): Els discos fixos són generalment magnètics.

10. (b): Almenys la porció del programa que s'està executant ha de residir en la RAM.

11. (a)

12. (b)

13. (a)

14. (a): Cada punt pot prendre un de 256 colors. Per a poder emmagatzemar el color cal un nombre de bits suficient per a fer 256 combinacions. Amb 8 bits podem fer $2^8 = 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 256$ combinacions. Per tant, es necessiten 8 bits, és a dir, un octet.

15. (c)

16. (a)

17. (b): El nombre de bits necessari per a poder generar 35.000 combinacions és el nombre de vegades que cal multiplicar $2 \times 2 \times \ldots$ just fins al punt en què se supera 35.000. Cal fer-ho 16 vegades; per tant, necessitem 16 bits. Com que cada octet són 8 bits, són necessaris 2 octets.

18. (a)

19. (c)

20. (b): Cada punt pot presentar-se en 8 colors diferents. Amb 3 bits podem emmagatzemar $2 \times 2 \times 2 = 8$ colors. Per tant, la imatge ocupa $1.024 \times 1024 \times 3 = 3.145.728 bits$, que són $3.145.728/8 = 393.216$ octets, que són $393.216/1024 = 384$ kilooctets.

21. (c)

22. (c)

23. (c): Cada imatge ocupa $100 \times 100 = 10.000$ bits, que són $10.000/8 = 1.250$ octets. 1 GB = $1.024 \times 1.024 \times 1.204 = 1.073.741.824$ octets. $1.073.741.824/1.250 = 858.993$. Podem emmagatzemar més de 800.000 imatges.

24. (b)

25. (c)

26. (b): Cada dígit decimal pot prendre, per separat, 10 valors diferents (de zero al nou). Tres bits per dígit decimal no són suficients (permeten només 8 combinacions); quatre, sí. Així, cada dígit decimal ocupa 4 bits; si n'hi ha 9, necessitem 36 bits.

27. (b): 2GB són $2 \times 1.024 \times 1.024 = 2.097.152$ kB. En la targeta Compact Flash hi caben $2.097.152/2.048 = 1.024$ imatges. Amb 3 targetes puc emmagatzemar $3 * 1024 = 3.072$ imatges.

28. (b)

29. (b)

30. (a)

31. (c)

32. (a)

33. (b)

34. (a)

35. (c)

36. (a): Tenim 12 valors diferents (de gener a desembre). Tres bits per mes no són suficients (permeten només 8 combinacions); quatre sí (permeten 16 combinacions).

37. (c)

Capítol 3

Internet

Una de les eines informàtiques bàsiques que es troben a l'abast de les persones que es dediquen professionalment a la traducció és Internet. Internet permet bàsicament tres tipus d'ús:

Com a mitjà de comunicació: Internet permet la comunicació i l'intercanvi d'arxius (vegeu l'apartat 2.4) amb clients o proveïdors, la participació en fòrums de professionals, la realització de consultes, etc.

Com a font de documentació: A més de contenir textos de moltes classes que poden servir d'exemple o inspiració a l'hora de fer traduccions, s'hi poden trobar enciclopèdies, diccionaris, glossaris, memòries de traducció (vegeu el capítol 10), i moltes altres fonts de documentació.

Com a rebost de programari d'assistència a la traducció: Molts dels programes específics d'assistència a la traducció estan disponibles a Internet, com ara els sistemes de traducció automàtica (vegeu el capítol 8) en línia o els programes de concordances bilingües.[1] L'accés pot ser a través d'un navegador, o a través d'altres programes que tinguem instal·lats localment en el nostre ordinador.[2]

3.1 Què és Internet?

Anomenem *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

[1] Programes de concordances bilingües disponibles en Internet: Linguee (`http://www.linguee.es/`); Reverso Context (`http://context.reverso.net/`).

[2] Usant protocols ben especificats, normalment a través d'API, *Application Program Interfaces* o *interfícies de programació d'aplicacions*.

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 ni per als programes d'aplicació que fan ús d'aquestes connexions. Altres noms que s'usen en comptes d'*Internet* són *World Wide Web* o *WWW* ("teranyina d'abast mundial") o simplement *web* ("teranyina"), masculí en català (*el web*).

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).

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`); aquest ordre és l'invers al dels números IP (en això els noms s'assemblen a les adreces postals: primer es dóna el més concret i al final el país).

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; antigament 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).

Per saber més sobre servidors de noms

Podríem fer una analogia entre la relació entre els noms i els números IP dels ordinadors d'Internet i els noms i els números de telèfon de l'agenda del nostre mòbil. Quan telefonem a una persona, normalment ho fem buscant el seu nom en l'agenda, i

INDICATIU	PAÍS
.es	Espanya
.fr	França
.pt	Portugal
.it	Itàlia
.uk	Regne Unit
.ru	Rússia
.za	Sud-àfrica
.ie	Irlanda
.tv	Tuvalu
.to	Tonga
.nu	Niue
.fm	Estats Federats de Micronèsia

Taula 3.1: Indicatius d'Internet d'alguns països. Fixeu-vos que alguns indicatius (`.tv`, `.fm`, etc.) s'usen per a aplicacions no estrictament relacionades amb aquestos països.

INDICATIU	TIPUS
.gov	governamental
.mil	militar
.com	comercial
.org	organització no lucrativa
.edu	institució educativa
.info	webs informatives
.cat	cultura i llengua catalanes (patrocinat per la fundació puntCat)
.eus	cultura i llengua basques (patrocinat per la fundació PuntuEus)
.museum	museus (patrocinat per MuseDoma)

Taula 3.2: Alguns indicatius Internet usats originalment als Estats Units d'Amèrica i més recentment arreu del món, alguns d'ells patrocinats per determinades institucions.

poques vegades ho fem pel número, però per fer la telefonada és necessari el número. Quan accedim a ordinadors d'Internet, ho fem de manera similar: accedim pel nom i no pel número IP, el qual és indispensable per a fer la connexió. Però, en contrast amb l'agenda del nostre mòbil, és impracticable tenir tots els noms i els números IP corresponents a tots els ordinadors del món en el nostre ordinador. Per això s'usen *servidors de noms*: ordinadors als quals el nostre ordinador es connecta pel número IP i als quals pot preguntar pel número IP corresponent a un nom. Els *servidors de noms* s'organitzen de manera que es distribueixen la informació de manera jeràrquica fent ús del *sistema de noms de domini* (en anglés, *domain name system* o DNS).

Per exemple, quan volem connectar a `cercador.dlsi.ua.es`, el servidor de noms del nostre proveïdor d'accés a Internet veu que el nom acaba en `.es` i pregunta al servidor de noms que s'encarrega d'aquest *domini*; aquest servidor veu que l'element anterior és `.ua` i pregunta al servidor de noms de la Universitat d'Alacant (UA), i aquest, al seu torn, pregunta al servidor de noms del Departament de Llenguatges i Sistemes Informàtics, ja que l'element anterior és `.dlsi`. Aquest últim, finalment, entrega el número IP de l'ordinador anomenat `cercador` al servidor de la UA, i aquest al servidor del domini geogràfic `.es`, que l'entrega al nostre proveïdor de serveis d'Internet i aquest, al seu torn, al nostre ordinador per a fer la connexió. Per això, quan naveguem, la primera connexió tarda més: s'està *resolent* el nom que hem teclejat. Una vegada resolt, el nostre ordinador es guarda l'IP durant un temps per evitar preguntar de nou. A més, per reduir el tràfic a Internet els proveïdors de serveis d'Internet i els servidors de noms consultats també es guarden temporalment aquesta informació de manera que no sempre es desencadena el procés complet de consultes descrit.

3.4 Identificadors de recursos

Els serveis i els documents concrets presents en un ordinador (un servidor d'Internet) que els fa disponibles es poden designar mitjançant el seu *identificador uniforme de recursos* o, més comunament, *URI* (de l'anglés *uniform resource identifier*).[3] L'URI és, per tant, una expressió que identifica o localitza uniformement un servei o document (un recurs) de qualsevol dels que s'ofereixen en Internet.

Un URI té generalment tres parts, encara que s'hi donen algunes variacions i una d'elles no és obligatòria:

esquema: indica la classe de recurs i com l'ha d'usar l'ordinador sol·licitant (o *client*).

autoritat: identifica pel seu nom o número IP l'ordinador (*servidor*) on és el recurs.

trajectòria: (opcional) dóna informació sobre la localització del servei o document dins de l'ordinador servidor (moltes vegades similar a les *trajectòries* dels fitxers, p. 12).

Per exemple, l'URI

[3] La denominació més usual era URL, *uniform resource locator* o localitzador uniforme de recursos, que encara s'usa profusament, encara que no tots els URI son URL.

`http://www.canalcuina.tv/concurs/sms/index.html`

es refereix a un document d'*hipertext* —un document de text que conté enllaços que permeten accedir directament a altres hipertextos relacionats— compatible amb l'esquema `http` (*hypertext transfer protocol* o protocol de transferència d'hipertextos) situat en l'ordinador `www.canalcuina.tv` (de l'empresa fictícia Canal Cuina, possiblement pertanyent al món de la televisió[4]), i, dins d'aquest, en el directori `concurs`, 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 (vegeu l'apartat 4.1).

L'esquema `https://` és similar a l'esquema `http://` però incorpora, a més, mecanismes per a transmetre amb seguretat informació encriptada (xifrada). Molts dels servidors d'Internet encarregats de manipular informació privada usen aquest esquema.

Els URI no només serveixen per enllaçar hipertextos: l'URI `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 esquemes són `rtsp://`, *real-time streaming protocol*, per a enllaçar contingut com ara vídeo, àudio, etc. en temps real, o `ftp://`, *file transfer protocol*, usat, cada vegada menys, per a descarregar (transferir) fitxers per a guardar-los en el nostre ordinador.

3.5 Navegadors

Els programes navegadors es coneixen també per altres noms: `browsers` (fullejadors), *exploradors*, etc. (vegeu també la pàg. 8). 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 o recurs enllaçat; a més, els navegadors poden *llançar* automàticament altres programes d'aplicació per a poder obrir el recurs corresponent si no és un hipertext.

Els navegadors més usats són *Firefox* (un programa lliure i de codi font obert desenvolupat per centenars de col·laboradors arreu del món), *Chrome* (el navegador de la companyia Google, el qual té una versió lliure i de codi font obert anomenada *Chromium*), *Microsoft Internet Explorer* (incorporat en el sistema operatiu Windows), *Safari* (el qual forma part del sistema operatiu MacOS), i d'altres com *Opera*, etc.

[4]Encara que, com es mostra en la taula 3.1, l'indicatiu designa un estat del Pacífic anomenat Tuvalu

3.6 Buscadors

Un dels recursos d'Internet més útils són els *buscadors* o *cercadors*. Es tracta de pàgines *web* que permeten buscar documents d'Internet; s'hi ha de teclejar una o més paraules, a més d'altres *condicions de recerca* opcionals —com ara que els documents hagen d'estar en una llengua determinada o en un servidor determinat— i entreguen els URI dels documents que compleixen aquestes condicions, enllaços a aquests documents i un petit resum o retall (anomenat *snippet*) del contingut de les pàgines desitjades.[5]

Exemples de recerques:

- `megamòndria`: documents que continguen el mot *megamòndria* i potser formes del mateix com ara el plural *megamòndries*, el compost *mega-mòndria*, o la forma sense accent *megamondria*.
- `megamòndria síngula`: documents que continguen aquests dos mots o variants.
- `megamòndria síngula site:ua.es`: documents que continguen aquests dos mots o variants i que estiguen en documents amb URI que acaben en *ua.es*.
- `megamòndria síngula filetype:pdf`: documents PDF que continguen aquests dos mots.

Ha de quedar clar que els buscadors realment *no busquen* documents en Internet sinó que consulten *índexs* que han anat construint a partir dels documents que van visitant. Per tant, pot haver-hi documents que els buscadors no troben perquè mai no els han visitat. Per la mateixa raó, també pot passar que els buscadors entreguen resultats corresponents a pàgines que ja no existeixen.

Un dels buscadors més populars a l'hora d'escriure aquestes línies és *Google* (`http://www.google.com`); però també hi ha altres com *Duckduckgo!*, *StartPage*, etc. La major part d'aquests buscadors tenen interfícies d'ús en moltes llengües.

3.7 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)

[5] Alguns buscadors, com ara *Google*, modifiquen els enllaços que porten als resultats, de manera que no hi porten directament sinó passant primerament pel servidor del buscador, per tal de conéixer les preferències de la gent i millorar així la rellevància dels resultats, però fins i tot poden establir un perfil de cada persona usuària. Això fa que algunes persones es plantegen l'ús de buscadors que no facen aquest *seguiment* o *tracking*.

a usuaris d'altres ordinadors. Els missatges poden contenir, a més del text mateix del missatge, fitxers *annexos* (o *adjunts*), en anglés *attachments*) com ara imatges, documents, missatges reenviats, etc.

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 (o *part local*) és freqüentment l'identificador d'una persona i la segona (la *part de domini*) sol prendre la forma de nom d'un ordinador (o d'un grup d'ordinadors que comparteixen un nom). Per exemple, una adreça electrònica vàlida podria ser

`marty.mcfly@backtothefuture.info`

A vegades, podem usar la nostra adreça de correu electrònic per a identificarnos a l'hora d'accedir a alguns dels serveix que s'ofereixen en Internet, com ara els serveix de xarxes socials (vegeu l'apartat 3.9).

Una adreça electrònica pot també identificar 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.[6]

Per llegir, escriure o enviar els missatges de correu electrònic, s'usen *programes gestors de correu electrònic*, com ara Thunderbird, Outlook, etc. També és molt freqüent, accedir al correu electrònic des de qualsevol lloc usant un navegador, a través d'un servei anomenat *webmail*; són comuns els *webmails* gratuïts (GMail, Yahoo Mail, MicroSoft Outlook); cada alumna o alumne de la Universitat d'Alacant té, per ser-ho, una adreça de correu de la forma *xxxx*`@alu.ua.es`, accessible a través de la secció *webmail* de la pàgina web de la Universitat.

3.8 Missatgeria instantània i xat

La missatgeria instantània i el xat (en anglés *chat*) permeten —com en el cas del correu electrònic, a través de programes especialitzats o *webs* accessibles amb un navegador— una comunicació escrita molt ràpida ("en temps real"), consistent en missatges normalment curts —opcionalment amb annexos com ara fotografies, contactes—, de manera que el resultat és similar al

[6] Per exemple, la llista de distribució sobre traducció automàtica MT-List, mantinguda 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'URI `http://lists.eamt.org/mailman/listinfo/mt-list`. Si s'envia un missatge a `mt-list@eamt.org` el reben tots els subscrits. Una altra llista d'interés, Tradumàtica, sobre tecnologies de la traducció, permet subscripció a través de `https://listserv.rediris.es/cgi-bin/wa?A0=TRADUMATICA`.

d'una conversa, però per escrit,[7] cosa que permet un registre de comunicació molt informal que, de fet, ha donat lloc a una llengua molt diferenciada tant de l'oral com de l'escrita.

Amb la generalització de l'ús dels mòbils intel·ligents o *smartphones*, han aparegut moltes aplicacions d'aquest tipus, com ara Telegram, Whatsapp, Line, etc., que usen com a identificador el número de telèfon.

Les converses poden ser entre dues persones, o entre un *grup* de persones, a voltes reunits en una *sala*, amb interessos comuns. Les persones que participen en un *xat* d'aquests últims, poden de vegades elegir un àlies o malnom i "entrar" a la sala, o estar sempre connectades al grup, i "conversar" per escrit públicament. Des del grup o la sala es poden establir "converses a part" (a banda) quan cal amb alguna persona concreta.

Entre els serveis de missatgeria instantània comercial més populars estan els associats a xarxes socials com ara Facebook o Tuenti, o altres associats a altres aplicacions com ara Google Hangouts o Skype.

3.9 Serveis de xarxa social

Una de les aplicacions més freqüents d'Internet són en l'actualitat els *serveis de xarxa social*, comunament anomenats simplement *xarxes socials*. Són plataformes informàtiques que usen Internet (a través d'un navegador i freqüentment també a través de programes d'aplicació específics, molt populars per a telèfons mòbils) per a construir xarxes de persones que comparteixen interessos o objectius. Alguns exemples:

Facebook permet a cada persona publicar informacions sobre el seu *estat*, incloent fotos, i enviar-se missatges; l'estat pot ser visible per a tothom o per a persones *amigues*.

Google+ es pot veure com la resposta de Google a Facebook; en Google+ el concepte bàsic és el del *cercle*.

Twitter és una xarxa social que es basa en missatges de menys de 140 caràcters que poden dur adjuntes fotos, enllaços, etc.

Instagram fa èmfasi en la possibilitat de compartir fotografies i vídeos.

LinkedIn serveix per a construir xarxes relacionades amb l'activitat professional.

Hi ha molts altres d'abast mundial (Pinterest, Reddit, Tumblr, etc.) i alguns particulars de determinades àrees geogràfiques (com ara VK en els països on es parla rus).

[7] Encara que s'està popularitzant l'ús de *notes de veu*, arxius d'àudio enregistrats i que s'envien com a annexos.

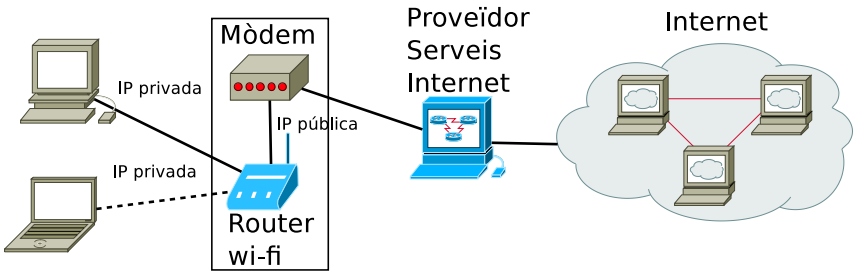


Figura 3.1: Esquema d'accés domèstic a Internet des de diversos dispositius connectats a un mòdem-encaminador.

3.10 L'accés a Internet

3.10.1 Accés domèstic

Per a accedir a Internet des de casa, cal, d'una banda, donar-se d'alta amb algun proveïdor de serveis d'Internet (ISP, *Internet service provider*) —alguns proveïdors ofereixen, a més de l'accés a Internet, televisió digital i telefonia convencional— i d'altra un *mòdem* adequat al tipus de connexió (mòdem ADSL, mòdem de cable, etc.). Normalment els mòdems que ens venen els proveïdors de serveis d'Internet són a l'hora mòdem i encaminador (en anglés *router*) per a permetre que hi connectem més d'un dispositiu, normalment a través d'una connexió sense fils Wi-Fi, formant una xarxa local (en anglés *local area network*, LAN). La figura 3.1 mostra un esquema de l'accés domèstic a Internet des de diversos dispositius.

El nostre proveïdor de serveis d'Internet assigna un número IP públic al nostre encaminador (*router*), de manera que forme part d'Internet i puga facilitar l'accés a tots els dispositius de la nostra xarxa local (clients) a tots els serveis i documents disponibles en qualsevol màquina (servidora) d'Internet. Als dispositius de la nostra xarxa local l'encaminador els assigna un número IP privat al que només tenen accés els ordinadors que formen part d'aquesta xarxa local; tots els dispositius d'aquesta xarsa local es connecten a Internet usant el mateix número IP, el número IP públic assignat al nostre encaminador.

En la majoria dels casos, el número IP públic que el nostre proveïdor d'accés a Internet assigna al nostre encaminador és temporal i va canviant, per això els ordinadors que tenim a casa no poden actuar com a servidors.

En l'actualitat, en els domicilis del País Valencià hi ha bàsicament tres modalitats d'accés a Internet, totes per preus que oscil·len al voltant dels 30–40 euros/mes:

ADSL: (vegeu el glossari de la secció 2.7) el mòdem fa la connexió a través

dels fils de telefonia convencional ja instal·lats a les cases, i en l'actualitat s'aconsegueixen velocitats de baixada d'alguns Mb/s i de pujada normalment inferiors a 1 Mb/s.

Cable: el mòdem fa la connexió a través d'un cable coaxial, tecnologia que s'usava tradicionalment per a televisió, amb velocitats i preus similars a l'ADSL.

Fibra: el mòdem fa la connexió a través d'un feix de fibra òptica (làser); aquesta tecnologia és la més recent i permet velocitats de baixada de desenes o fins i tot centenars de Mb/s i velocitats de pujada superiors al Mb/s.

Per saber més sobre els antics mòdems telefònics

Fins als primers anys del decenni del 2000, la major part dels domicilis particulars i les petites empreses es connectaven a Internet mitjançant la línia telefònica, però usant una tecnologia molt més rudimentària, que requeria fer una telefonada al número del proveïdor d'accés a Internet, que havia de durar el temps de connexió, independentment de la quantitat de dades que s'hi transferiren. La telefonada es podia pagar per minuts o amb plans que permetien connexions il·limitades en horari no comercial que s'anomenaven *tarifa plana*. El mòdem modulava i desmodulava senyals similars als que s'envien quan es fa una telefonada de veu. Durant la connexió, la línia quedava ocupada i no es podien fer ni rebre telefonades. Les velocitats eren molt baixes, de desenes de kb/s.

3.10.2 Accés mòbil

Com més va, menys ens connectem a Internet des de casa, i més des dels nostres dispositius mòbils com ara els *smartphones*. Si som a l'abast d'una xarxa Wi-Fi a la qual tenim accés (com la de la nostra casa o la que ens ofereix la Universitat), el nostre dispositiu mòbil es connectarà a Internet a través de Wi-Fi. Si no, haurem d'usar les *dades mòbils* que comercialitza el nostre proveïdor de telefonia mòbil a través de la seua xarxa cel·lular. En l'actualitat s'està fent la transició de la tecnologia anomenada de tercera generació o 3G (que es mostra de vegades també com una "H" en la pantalla), que permet velocitats de connexió d'uns pocs Mb/s a la de quarta generació o 4G, que permet velocitats molt més ràpides (de vegades, quan la cobertura no és bona, el nostre mòbil recorre a tecnologies més antigues i més lentes, com ara EDGE[8] —que permet centenars de kb/s i es mostra com una "E" en la pantalla— o GPRS[9] —que permet desenes de kb/s

[8] *Enhanced Data Rates for GSM Evolution* "Taxes de dades millorades per a l'evolució del GSM"

[9] *General Packet Radio Service* "Servei general de paquets [de dades] per ràdio"

i es mostra com una "G" en la pantalla). En l'actualitat, podem comprar paquets de dades i pagar uns 5–7 euros per GB, o una quota mensual de desenes d'euros i tenir dades il·limitades.

3.11 Questions i exercicis

1. La primera part d'un URI (identificador uniforme de recursos) especifica ...
   (a) ...l'esquema d'accés.
   (b) ...el nom del servidor.
   (c) ...el directori on es troba el servei.
2. Després de l'esquema d'accés, un URI (identificador uniforme de recursos) especifica
   (a) la velocitat de transferència.
   (b) el nom del servidor.
   (c) el directori on es troba el servei.
3. Què és "`http://www.tharaka.org.ke/nkoru`"?
   (a) Un URI.
   (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 un número IP?
   (a) 16
   (b) 32
   (c) 4
5. Quan en un navegador no s'indica l'esquema d'un URI, quin esquema se sobreentén?
   (a) `http://`
   (b) `mailto:`
   (c) L'esquema d'Internet
6. Què es pot dir dels números IP de dues màquines que es troben en la mateixa subxarxa?
   (a) No se'n pot dir res: els números IP poden no tenir res a veure.

(b) Que tenen en comú els primers octets.

(c) Que tenen en comú els últims octets.

7. Sakurako es connecta a Internet per via telefònica amb un ordinador que té un processador de 1000 MHz, 128 megaoctets (*megabytes*) de RAM i un mòdem de 57600 bits per segon. Ahmed té un ordinador amb un processador de 500 MHz i 128 megaoctets de RAM i ha contractat un mòdem de cable de 128 kilobits per segon per a connectar-se a Internet. Sakurako vol convéncer Ahmed que ella es descarrega els fitxers MP3 més ràpid que ell, però Ahmed li diu que en les mateixes condicions ell tarda menys a baixar-se els fitxers, de vegades la meitat de temps. Qui té raó?

   (a) Ahmed

   (b) Els dos es baixen els arxius en el mateix temps perquè els dos ordinadors tenen la mateixa RAM.

   (c) Sakurako

8. Si la màquina `fictici.deconya.ua.es` té el número IP 232.111.22.33 quin dels tres número IP següents és més probable que corresponga a la màquina `fals.deconya.ua.es`?

   (a) 230.111.22.33

   (b) 232.111.22.13

   (c) 67.15.22.99

9. Com s'indica en Internet on és un recurs concret?

   (a) Mitjançant un URI.

   (b) Mitjançant un hiperenllaç.

   (c) Mitjançant una etiqueta HTML.

10. Quin dels següents és un número IP vàlid?

    (a) 64.128.64

    (b) 255.256.111.1

    (c) 111.255.111.111

11. En un número IP, quina part és igual per a dos ordinadors connectats en la mateixa xarxa local?

    (a) La part inicial.

    (b) La part final.

    (c) Pot no coincidir-hi res perquè els números IP s'assignen aleatòriament.

12. A la nostra casa hem tingut telèfon tota la vida i ara estem pensant a connectar-nos a Internet mitjançant ADSL. Hem de fer cap instal·lació addicional a casa?

    (a) No, però ens quedarem sense telèfon i només hi tindrem Internet.

    (b) Sí, de segur que els tècnics vindran a passar cables arreu la casa.

    (c) No. I gaudirem de telèfon i Internet alhora.

13. Quan realitzem una recerca en Internet, el cercador ...

    (a) ... visita en aquell moment els diferents documents d'Internet i recopila aquells que satisfan el criteri de cerca.

    (b) ... consulta un índex que ha construït prèviament en visitar els diferents documents d'Internet.

    (c) ... consulta un índex que ha construït prèviament i també visita en aquell moment els documents d'Internet per si hi haguera algun de nou que no existia quan va crear l'índex.

14. Si en cercar un recurs en Internet el cercador ens diu que no hi ha documents que satisfacen el criteri de cerca ...

    (a) ... podem estar segurs que no existeix en tot Internet cap document que el satisfaça.

    (b) ... podria donar-se el cas que un document de recent creació satisfaça el criteri de cerca però no haja estat visitat pel cercador.

    (c) Les altres dues respostes són errònies.

15. Donada la URI `http://edu.gob.es/educacion/universidades.html`, indica quina de les següents afirmacions és falsa:

    (a) El recurs `universidades.html` està allotjat en un ordinador el nom del qual és `edu.gob.es`.

    (b) La ruta d'accés al recurs és `educacion/universidades.html`.

    (c) L'ordinador on s'allotja el recurs es diu `http://edu.gob.es`.

3.12 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: $8 \times 4 = 32$.
5. (a)
6. (b)
7. (a): 57.600 b/s són 57.600 / 1024 = 56,25 kb/s.
8. (b)
9. (a)
10. (c)
11. (a)
12. (c)
13. (b)
14. (b)
15. (c)

Capítol 4

Textos i formats

El tipus de fitxer bàsic amb què treballen els professionals de la traducció sol ser un fitxer amb text, és a dir, un text informatitzat, també anomenat *document de text*. Aquest fitxer pot contenir, a més del text mateix, informació sobre la presentació (el format dels paràgrafs i de les pàgines, els tipus i les grandàries de lletra que s'usen amb cada mot, etc.) o sobre l'organització del contingut (indicacions que una determinada part del text és un títol de capítol, el títol d'una secció o una nota a peu de pàgina, 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 el capítol 5).
- El podem haver rebut annex a un missatge de correu electrònic (vegeu l'apartat 3.7) o per missatgeria instantània (vegeu l'apartat 3.8).
- El podem haver descarregat (copiat) d'algun servidor d'Internet (vegeu la pàg. 33).
- El podem haver generat, potser a partir d'un altre text, usant un *processador de textos* (vegeu l'apartat 4.6).
- El pot haver generat un *sistema de reconeixement de la parla* 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.

4.1 Formats de text

Un *text informatitzat* és, com qualsevol porció de dades informatitzada, una *seqüència de bits*, és a dir, d'*uns* i *zeros* de l'estil de la següent:

```
01000001010011010100 1001...
```

Com ja hem vist en l'apartat 2.3, els *bits* s'agrupen en grups de vuit (*bytes* o *octets*):

```
01000001 01001101 01001001...
```

Hi ha moltes maneres d'organitzar aquests octets per a emmagatzemar els textos; molts dels problemes que apareixen quan es tracten textos amb l'ordinador provenen de discrepàncies quant a la manera de fer-ho. En les seccions següents estudiarem dos aspectes importants que anomenarem *codificació* i *format* pròpiament dit. La *codificació* és l'assignació d'una seqüència concreta, d'un o més *octets*, a cada caràcter possible d'un text. El *format* d'un document és la part no textual d'aquest i que serveix per a codificar informació estructural (sobre l'organització del contingut del document) o presentacional (sobre l'aparença que tindrà el document quan es presente).

4.2 Codificació de caràcters

La codificació dels caràcters d'un text consta de dues fases:

1. S'assigna a cada caràcter un nombre enter positiu anomenat *punt de codi* o simplement *codi*; per exemple: "`a`" $\rightarrow$ 97; "`?`" $\rightarrow$ 63.
2. els codis numèrics es converteixen en octets assignant-los una determinada seqüència de bits; per exemple: 97 $\rightarrow$ `01100001`; 63 $\rightarrow$ `00111111`).

4.2.1 ASCII

Com ja s'ha comentat en la pàgina 11, per a emmagatzemar textos s'ha usat històricament l'estàndard ASCII (*American Standard Code for Information Interchange*). Aquest estàndard assigna un número del 0 al 127 a cada caràcter de l'alfabet llatí usat en anglés i fa servir 7 bits per a codificar-lo,[1] de manera que permet emmagatzemar un caràcter per octet i encara en sobra un bit.[2] 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.

L'estàndard ASCII té la limitació que no permet escriure caràcters propis de moltes llengües europees, com ara lletres amb signes diacrítics (*á*, *ò*, *ç*, *ñ*, *ü*, etc.) o lletres especials com *ß*.

[1] Amb 7 bits es poden fer $2^7 = 128$ combinacions. Per això els codis assignats per ASCII van del 0 al 127.

[2] Inicialment aquest bit es feia servir com a bit de control per a detectar errades en la transmissió dels textos.

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.

4.2.2 Extensions d'ASCII

Amb l'arribada dels microordinadors[3] en els anys vuitanta es va decidir ampliar l'estàndard ASCII, de 7 bits i per tant amb $2^7 = 128$ caràcters diferents, a un codi de 8 bits amb $2^8 = 256$ caràcters diferents. El vuité bit o "bit 7"[4]—el primer per l'esquerra— és 1 per als nous caràcters (numerats del 128 al 255) i zero per als caràcters estàndards d'ASCII.

Hi ha diverses extensions d'ASCII, cadascuna adreçada a un conjunt de llengües concret que fan servir (quasi) el mateix alfabet. En la nostra àrea geogràfica s'usa normalment la codificació ISO-8859-1 o *Latin-1* (vegeu la taula 4.2); aquesta codificació serveix per a les llengües següents: *afrikaans* (llengua germànica parlada en la República de Sud-àfrica), alemany, anglés, basc, català, danés, escocés, espanyol, feroés, finés, francés, gallec, irlandés, islandés, italià, neerlandés, noruec, portugués i suec.[5] El sistema operatiu Windows de Microsoft usa la codificació de 8 bits anomenada CP-1252, també anomenada *WinLatin-1*, que és més àmplia que ISO-8859-1, ja que usa alguns dels codis 128–159 per a caràcters (per exemple, usa el codi 128 per al símbol de l'euro).

Hi ha altres codificacions en la família ISO-8859. Per exemple, l'albanés, el bosni, el croat, el txec, l'hongarés i el romanés usen una codificació anomenada ISO-8859-2 o *Latin-2*; el letó, el lituà i l'estonià usen l'ISO-8859-4 o *Latin 4*; el rus usa l'ISO-8859-5 que conté l'alfabet ciríl·lic a més de l'alfabet llatí bàsic. Per això, és molt important conéixer quin esquema de codificació de caràcters s'ha usat en un document de text determinat per a poder-lo llegir correctament; alguns formats de text inclouen aquesta informació dins del mateix document.

El fet que hi haja diverses maneres 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 (els caràcters d'ASCII es veuen normalment bé: els que fallen són els nous). Fixeu-vos que si en un document ISO-8859-1 s'escriu la frase *Què és això?*, on els caràcters "è", "é" i "ò" tenen codis per damunt de 127 (233, 232 i 242 respectivament), i provem a llegir-lo com si fóra un document ISO-8859-2, llegirem *Quč és aixň?*, perquè dos d'aquests codis (233 i 242) tenen una altra interpretació en aquesta codificació ("č" i "ň" respectivament). Es per això que no podem mesclar en un mateix document textos en llengües que fan servir extensions d'ASCII diferents.

[3] Els primers ordinadors que tenien una grandària que permetia tindre'n un a casa.

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

[5] Hi ha una modificació anomenada ISO-8859-15, que inclou, entre altres, el símbol de l'euro i resol alguns problemes referents al francés i al finés.

CODI BINARI	CODI DECIMAL	CARÀCTER
10100000	160	(espai no trencable)
10100001	161	¡
...	...	...
10110101	181	μ
10110110	182	¶
10110111	183	·
10111000	184	¸
...	...	...
11000000	192	À
11000001	193	Á
11000010	194	Â
11000011	195	Ã
11000100	196	Ä
11000101	197	Å
11000110	198	Æ
11000111	199	Ç
11001000	200	È
11001001	201	É
11001010	202	Ê
11001011	203	Ë
11001100	204	Ì
11001101	205	Í
11001110	206	Î
11001111	207	Ï
...	...	...
11100000	224	à
11100001	225	á
11100010	226	â
11100011	227	ã
11100100	228	ä
11100101	229	å
11100110	230	æ
...	...	...
11111111	255	ÿ

Taula 4.2: Alguns exemples d'ISO-8859-1 (*Latin-1*). Els codis del 0 al 127 són com els d'ASCII. Els codis del 128 al 159 no estan assignats.

4.2.3 Unicode

Els codis de 8 bits com ISO-8859-1 (*Latin-1*) 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. Per a aquestes llengües 256 combinacions no són suficients i s'hi han proposat diverses solucions. *Unicode*[6] (ISO 10646) és un nou estàndard per a codificar pràcticament els caràcters de totes les llengües del món i fins i tot mesclar diversos alfabets en un mateix fitxer.

Unicode fa servir 31 bits; és a dir, permet $2^{31} = 2.147.483.648$ caràcters diferents. La versió més comunament usada d'Unicode (BMP, *Basic Multilingual Plane*) té 65.534 caràcters; això comportaria l'ús de 2 octets (16 bits) en comptes d'un ($2^{16} = 65.536$); això faria que un text Unicode senzill fóra el doble de gran que el text ASCII corresponent. Per tal d'estalviar espai, hi ha mètodes de serialització 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). En concret, UTF-8 usa:

- per als codis del 0 al 127, 1 octet (compatible amb ASCII);
- per als codis del 128 al 2047, 2 octets;
- per als codis del 2048 al 65535, 3 octets, i així successivament.

4.2.4 Limitacions

Tot i que ampliem l'ASCII a ISO-8859 o Unicode, encara és molt limitat. Per exemple, si volem que un text tinga un cert format, només podrem usar caràcters de control com ara l'espai en blanc, el tabulador, el salt de línia, etc. Per exemple, no podrem canviar fàcilment de tipus o de grandària de lletra, o indicar que una determinada part del text és el títol d'una secció o el text d'una nota a peu de pàgina. De qualsevol manera, les extensions d'ASCII (ISO-8859-X) i Unicode encara 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 normalment en fitxers amb noms que tenen l'extensió `.txt`. Aquests textos es poden produir i llegir amb qualsevol *editor de textos* (vegeu l'apartat 4.6).

[6] `http://www.unicode.org`

4.3 Format pròpiament dit

Els documents de text són en general més rics que simples seqüències de caràcters; els textos, a més de caràcters, contenen informació de *format*. Per això, és necessària l'assignació de *codis* (que també es convertiran en octets) per a regular altres característiques del text com:

- l'aparença *visual* que tindrà el document quan es presente (per exemple, "inici cursives", "final negretes", "lletra de 16 punts"), o
- l'*estructura*, és a dir, l'organització del contingut del document (per exemple, "títol de secció", "llista numerada", "nota a peu de pàgina", "fila d'una taula", etc.)

Per a guardar aquesta informació, s'usen:

- D'una banda, codificacions o formats basats en text (ISO-8859-X, Unicode, etc.). Tal és el cas del format SGML (*standardized generalized markup language*), la seua versió simplificada (i molt més estesa) XML (*extensible markup language*), el format HTML (*hypertext markup language*; basat en SGML), el format RTF (*rich text format*; proposat per Microsoft i sense relació amb SGML o XML), o el llenguatge per a impressores anomenat Postscript. Tots aquests formats usen combinacions especials[7] de caràcters de text per a indicar aquestes característiques d'estructuració o de presentació.[8]
- D'altra banda, codificacions o formats basats en codis binaris no interpretables com a caràcters. Tal és el cas dels formats particulars dels processadors de textos comercials com ara Corel WordPerfect o Microsoft Word.[9]

Com ja s'ha dit, l'ús de formats de text més avançats no només serveix per a determinar-ne 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

[7] Combinacions de caràcters poc freqüents en textos usuals.

[8] Aquests caràcters que indiquen el format no són normalment visibles per a la persona usuària mentre redacta o veu el document, excepte si demana explícitament que els vol veure.

[9] Hi ha una tendència a considerar el format de document de Microsoft Word, amb extensió `.doc`, com la manera estàndard d'enviar documents de text annexos a un missatge electrònic, sense considerar el fet que aquest format és privat i està associat a l'ús d'un determinat producte no lliure i de codi font tancat. El format `.docx` també anomenat Office Open XML o OOXML, està millor documentat i estandarditzat i pot ser processat més satisfactòriament amb processadors lliures i de codi font obert.

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, com veurem més endavant. De fet, recentment, amb l'aparició de XML (vegeu l'apartat 4.4.2), 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 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 (vegeu la secció 4.7).

4.4 SGML i XML

4.4.1 SGML

SGML (*standardized generalized markup language*), el llenguatge estàndard generalitzat de marques, havia tingut un èxit relatiu fins a mitjans dels noranta; però l'aparició cap a finals dels noranta d'una versió restringida i simplificada de SGML anomenada XML (*extensible markup language*) 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; per això, ens centrarem en aquest últim format. De tota manera, encara hi ha formats molt importants que es basen en SGML, com el llenguatge de marques per a hipertextos HTML (vegeu l'apartat 4.4.3) (excepte pel més recent, HTML5, estandarditzat el 2014, que ja no és una aplicació SGML). Hi ha també versions estàndards d'HTML conegudes com XHTML, que es basen directament en XML (vegeu l'apartat 4.4.2).[10]

4.4.2 XML

Marques

Un document XML és un document de text on, a més del text pròpiament dit, 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 missatge de correu electrònic 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 (*Latin-1*; vegeu l'apartat 4.2.2). Com s'hi pot veure, les etiquetes que apareixen entre parèntesis angulars indiquen les diverses parts del document, anomenades *elements*. Típicament, s'obrin amb <*nom*> i es tanquen amb </*nom*>. En l'exemple, es pot veure

[10] Noteu que HTML5 també té una versió *serialitzada en XML*, XHTML5.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE EMAIL SYSTEM "http://www.dlsi.ua.es/%7Efsanchez/tt/email.dtd">
<EMAIL>
  <DESTINATAR>
    <NOM>Mikel L. Forcada</NOM>
    <ADREÇA>mlf@dlsi.ua.es</ADREÇA>
  </DESTINATARI>
  <REMITENT>
    <NOM>Felipe Sánchez Martínez</NOM>
    <ADREÇA>fsanchez@dlsi.ua.es</ADREÇA>
  </REMITENT>
  <DATA>9 de novembre de 2015</DATA>
  <ASSUMPTE>Capítol 4 del llibre de TT</ASSUMPTE>
  <TEXT>
    <P>Mikel, estic acabant de fer modificacions al capítol
    dedicat a textos i formats. Quan acabe t'avise.</P>
    <P>Per favor, envia'm els apunts que vam preparar
    sobre traducció automàtica que no els trobe. Gràcies!</P>
  </TEXT>
</EMAIL>
```

Figura 4.1: El text d'un missatge de correu electrònic en XML.

que un missatge de correu (`<EMAIL>...</EMAIL>`) té un destinatari, un remitent, una data, un títol i un text; és a dir, els elements poden contenir altres elements; les marques funcionen com a parèntesis. Seguint amb la jerarquia d'inclusió d'elements en altres, tant el destinatari com el remitent tenen nom i adreça, i el text es compon de paràgrafs (`<P>...</P>`).

Documents XML ben formats

Aquestes són algunes de les característiques que fan que un document XML estiga *ben format*, és a dir, siga un document XML i no una altra cosa:

- Cada etiqueta d'inici d'element de la forma *<nom>*, *<nom atribut="valor">*, *<nom atribut1="valor" atribut2="valor">*, etc. (amb zero o més assignacions de valors a atributs) ha d'estar emparellat amb una etiqueta de final d'element de la forma *</nom>*, sense atributs però amb el mateix nom.[11] Si l'element és buit, *<nom...></nom>*, també es pot

[11] En SGML es permet que alguns elements es tanquen *implícitament*, sense necessitat d'una etiqueta de final d'element.

escriure <*nom*.../>

- Un element pot contenir qualsevol nombre d'elements.
- Els elements no es poden solapar o creuar: no és possible escriure, per exemple, `<a>text<b>més text</a>més text encara </b>`.
- El document conté un únic element *arrel* que conté tots els elements del text.
- El document pot contenir comentaris entre `<!--` i `-->` o instruccions de processament del tipus `<?`*nom*...`?>` en qualsevol lloc excepte dins de les etiquetes.
- Els valors dels atributs han d'anar entre cometes dobles (`"`*valor*`"`) o simples (`'`*valor*`'`).
- Un element no pot tenir dos atributs amb el mateix nom.
- Els caràcters `<` i `&` no poden aparéixer en el text dels elements ni dels atributs. Això és perquè `<` indica el començament d'una etiqueta i `&` el començament d'una *entitat* com ara `&copy;` que es pot usar per a representar el càracter ©: si es necessiten aquests caràcters, s'han d'escriure les entitats `&lt;` i `&amp;`, respectivament.

Com es pot veure, aquestes regles que defineixen un document XML ben format no diuen quines etiquetes són vàlides i quines no, o quins atributs pot tenir un determinat element, o quins elements poden anar dins d'un determinat element, en quin ordre o en quina quantitat.

Tipus de documents

Per a especificar, per a un tipus determinat de document XML, quines etiquetes són vàlides, quins atributs pot tenir cada element, o quins elements poden anar dins d'un determinat element, en quin ordre o en quina quantitat, es pot usar una DTD (*document type definition* o definició del tipus de document).[12]

La segona línia del missatge de correu de la figura 4.1 especifica el tipus del document tot indicant d'una banda l'etiqueta arrel o principal del document (`EMAIL`) i l'URI (`SYSTEM`) on es troba la DTD. Aquesta DTD es veu en la figura 4.2; examinem ara 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 (*Latin-1*):

[12]Les DTD no són l'única manera d'especificar famílies de documents XML; una altra manera més potent són els anomenats *esquemes XML* (en anglés *XML schema*).

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

Figura 4.2: La DTD que defineix missatges de correu electrònic com el de la figura 4.1.

```
<?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 EMAIL -->
   ```

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

   ```
   <!ELEMENT EMAIL (DESTINATARI+, REMITENT?, DATA, ASSUMPTE, TEXT)>
   ```

   defineix l'element arrel o principal, `EMAIL`, i especifica que es compon (en l'ordre especificat) d'un o més `DESTINATARI`s (el símbol + indica que pot haver-hi un o més), d'un `REMITENT` opcional (indicat amb `?`), d'una `DATA`, d'un `ASSUMPTE` i d'un `TEXT`.

4. Un `DESTINATARI` del missatge de correu té dues parts: el `NOM` (opcional) i l'adreça de correu (`ADREÇA`):

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

5. El remitent es defineix igual:

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

6. El `NOM`, l'`ADREÇA` la `DATA` i l'`ASSUMPTE` contenen text sense marques (indicat amb `#PCDATA`):

```
<!ELEMENT NOM (#PCDATA)>
<!ELEMENT ADREÇA (#PCDATA)>
<!ELEMENT DATA (#PCDATA)>
<!ELEMENT ASSUMPTE (#PCDATA)>
```

7. El `TEXT` es compon d'un o més (`+`) paràgrafs (`P`). Si volguérem que el text estiguera compost per *zero* o més paràgrafs usaríem "`*`" en comptes de "`+`":

```
<!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 aquest últim és vàlid, és a dir, si segueix l'especificació donada en la DTD. Perquè un document siga vàlid con respecte a una DTD qualsevol, primer ha d'estar ben format, és a dir, ha de complir amb les regles bàsiques d'escriptura de documents XML esmentades més amunt.

Tot i que una DTD serveix per a la validació automàtica de documents XML del tipus que la DTD defineix, el *significat* de les etiquetes (és a dir, quines 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, aquest significat pot estar associat, per exemple, a la manera (*estil*, vegeu la p. 63) de presentar el document quan s'imprimeix (per exemple, els destinataris del missatge de correu poden anar en negretes), però també podria servir per a facilitar el processament de la informació (per exemple, buscar tots els missatges que tenen un determinat destinatari, o, en llibres codificats en XML, decidir quines parts han de ser traduïdes automàticament de l'espanyol a l'anglés i quines no perquè són cites literàries.[13] 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.

Una altra aplicació de les DTD és fer-les servir per a facilitar l'edició de documents XML vàlids: un editor de documents XML pot consultar la DTD per suggerir a la persona usuària l'element o elements correctes en el context actual, o per emetre un missatge d'error tan aviat com el document perda la seua validesa.

[13] Hi ha un estàndard anomenat TEI, de l'anglés *text encoding initiative*, "iniciativa de codificació de textos" (`http://www.tei-c.org`) que usa famílies de DTD per a definir diferents tipus d'obres (literàries i no literàries). De fet, existeixen, d'una banda, les antigues DTD per a SGML, i, d'altra, les DTD TEI per a XML.

4.4.3 (X)HTML

El format XHTML (*extensible hypertext markup language* o *llenguatge extensible de marques per a hipertextos* és un dels tipus de document que es poden definir amb XML i es correspon amb la versió XML del llenguatge HTML (*hypertext markup language*), aquest últim basat en SGML (format precursor d'XML). Ambdós llenguatges s'usen per a escriure els hipertextos d'Internet (vegeu el capítol 3) i són el que interpreten els navegadors d'Internet (vegeu l'apartat 3.5).

Tant en XHTML com en HTML, les marques tenen un significat determinat. Per exemple, (X)HTML indica el començament d'un segment de text destacat (emfatitzat) amb la marca "`<em>`" (4 caràcters ASCII) i el final amb la marca "`</em>`" (5 caràcters). (X)HTML serveix per a codificar hipertextos: els enllaços (hiperreferències) a altres documents (que al seu torn poden també ser hipertextos) comencen amb "`<a href="` *URI* `">`" —on *URI* és l'identificador del document enllaçat— i acaben amb "`</a>`", etc. Els documents (X)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>`").

La principal diferència entre HTML i XHTML és que com aquest últim està basat en XML, el document ha de ser XML ben format i per tant no pot haver-hi elements que s'obrin però no es tanquen; això sí que és valid en HTML quan es tracta d'elements buits com `img` o `meta`. Altra diferència notable és que en XHTML els noms dels elements van sempre en minúscula, mentre que en HTML poden anar en majúscules.

El document XHTML que es mostra en la figura 4.3 es mostraria en un navegador aproximadament com en la figura 4.4. Com es pot veure, la primera línia, que comença amb "`<!DOCTYPE`" declara que el document és un tipus de document XHTML estàndard segons la versió 1.0 *estricta* de XHTML (hi ha diverses versions). En la tercera línia, l'etiqueta "`<html>`" indica el començament del document XHTML, i l'etiqueta "`</html>`" del final indica el final del document. Dins de l'element `html` trobem dos elements: `head` (l'*encapçalament*) i `body` (el *cos* del document). Dins de l'encapçalament, un element `meta` que no té contingut (fixeu-vos com s'obri i es tanca al mateix temps) indica, a través de dues assignacions del tipus *atribut*="*valor*", que el *joc de caràcters* que usa el document és l'ISO-8859-1, el més comú a Europa occidental.[14] Dins de `head` també trobem l'element `title`, el qual conté un títol que es presentarà, quan obriu el document amb un navegador, en la barra del navegador, *però no com a part del text del document*. Dins de `body` veiem encapçalaments de nivell 1 (`h1`), encapçalaments de nivell 2 (`h2`), paràgrafs (`p`), parts del text destacades (`em`), i enllaços (`a`). La taula 4.3 descriu algunes de les etiquetes més impor-

[14] Per a escriure documents en *txec* o en *coreà*, caldria canviar part del valor de l'atribut `content` perquè la codificació ISO-8859-1 no permet escriure en aquests idiomes.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
      "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
  <head>
    <meta http-equiv="Content-Type"
          content="text/html; charset=iso-8859-1"/>
    <title>Títol del document</title>
  </head>

  <body>
    <h1>Encapçalament de nivell 1</h1>

    <h2>Encapçalament de nivell 2</h2>

    <p>Aquest és el <em>primer</em> 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. </p>

    <h2>Un altre encapçalament de nivell 2</h2>

    <p>Aquest és l'<em>últim</em> paràgraf
    d'aquest document XHTML. Els documents XHTML poden contenir
    <a href="http://www.apertium.org">enllaços</a>
    a altres documents (X)HTML, locals o remots. </p>
  </body>
</html>
```

Figura 4.3: Un document XHTML, tal com el presentaria un editor de textos normal o usant l'opció "view HTML source" (veure font HTML) del navegador.

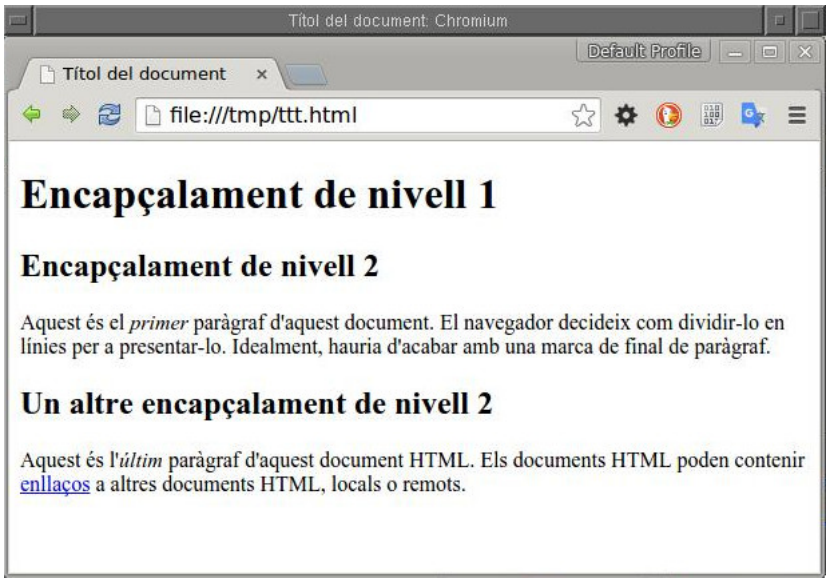


Figura 4.4: El document XHTML de la figura 4.3, vist a través d'un navegador d'Internet.

tants que s'usen en XHTML.

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").

4.4.4 Altres formats basats en XML

Hui dia s'han popularitzat els formats de documents basats en XML. Alguns dels formats basats en XML que interessen als traductors són: el format TMX (*translation memory exchange*) que s'usa per a l'intercanvi de memòries de traducció (fitxers amb extensió `.tmx`, vegeu el capítol 10), el format TBX (*termbase exchange*) per a l'intercanvi de bases de dades terminològiques (fitxers amb extensió `.tbx`, vegeu el capítol 5) i el format XLIFF (*XML localization interchange file format*). Aquest últim és un format creat per a la estandardització del format empleat per les diverses ferramentes que s'utilitzen durant el procés de *localització* d'un producte (vegeu el capítol 10).[15]

[15] La *localització* es pot definir com el procés d'adaptació d'un producte als usos d'una regió específica del món.

ELEMENT	DESCRIPCIÓ	MÉS INFORMACIÓ
`<html>...</html>`	Conté tot el document	
`<head>...</head>`	Encapçalament	
`<body>...</body>`	Cos	
`<meta.../>`	Informació sobre el document	L'element és buit
`<title>...</title>`	Conté el títol del document	
` `	Salt de línia forçat	L'element és buit.
`<h1>...</h1>`	Encapçalament de nivell 1	
`<h2>...</h2>`	Encapçalament de nivell 2	
...	...	...
`<h6>...</h6>`	Encapçalament de nivell 6	
`<p>...</p>`	Paràgraf	
`<ul>...</ul>`	Llista sense numerar	Conté elements `li`
`<ol>...</ol>`	Llista numerada	Conté elements `li`
`<li>...</li>`	Element de llista	Pot contenir una altra llista en el seu interior.
`<em>...</em>`	Èmfasi	
`<strong>...</strong>`	Èmfasi fort	
`<code>...</code>`	Exemple de codi	
`<a...>...</a>`	"Àncora"	Si porta un atribut `href="URI"`, el text entre "`<a...>`" i "`</a>`" funciona com un enllaç al document que hi ha en l'URI.
`<img.../>`	Imatge	L'atribut `src="URI"` indica l'adreça on és la imatge. L'atribut `alt="`*text*`"` descriu la imatge amb paraules. L'element és buit.

Taula 4.3: Alguns elements bàsics d'XHTML, la versió XML de HTML.

A més dels formats descrits en el paràgraf anterior hi ha dos formats usats pels processadors de textos més moderns que estan basats en XML. Aquests formats són OpenDocument i Office Open XML.

OpenDocument

OpenDocument és un format d'arxius obert i estàndard per a emmagatzemar, entre altes, textos (fitxers amb extensió `.odt`) fulls de càlcul (fitxers amb extensió `.ods`) i presentacions (fitxers amb extensió `.odp`). Aquest format es el emprat per defecte per les aplicacions ofimàtiques LibreOffice i OpenOffice.org i consisteix en diversos documents XML —per al contingut, els estils usats en el document, etc.— comprimits amb ZIP.[16]

Office Open XML

Office Open XML, també conegut com OOXML o OpenXML, és un altre format estàndard —impulsat per Microsoft— bastat en XML que s'utilitza per a emmagatzemar textos (fitxers amb extensió `.docx`), fulls de càlcul (fitxers amb extensió `.xlsx`) i presentacions (fitxers amb extensió `.pptx`). Igual que OpenDocument, un arxiu OpenXML consisteix en diversos documents XML comprimits amb ZIP.

4.5 Altres formats

4.5.1 RTF

RTF (*rich text format*, és a dir, *format de text ric*) va ser un format impulsat per l'empresa Microsoft per a facilitar l'intercanvi de documents entre processadors de textos mantenint-ne el format, i que encara s'usa de vegades. 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 alguns caràcters.

4.5.2 PDF

PDF (de l'anglés *portable document format*, format portable de document) és un altre format desenvolupat per a capturar completament les carac-

[16] ZIP és un format per a l'emmagatzemament de fitxers comprimits; els fitxers d'aquest tipus solen tenir l'extensió `.zip`.

```
{\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.

terístiques presentacionals dels documents. En PDF, el document es mostra exactament amb la mateixa aparença independentment de l'ordinador, sistema operatiu o aplicació que user per a veure'l. Els documents PDF poden emmagatzemar, a banda del text, tipus de lletra, gràfics, sons, etc. Aquest format va ser impulsat en els anys noranta per l'empresa Adobe, que ofereix en l'actualitat un programa gratuït[17] anomenat Adobe Acrobat Reader DC— per visualitzar els documents;[18] per crear-los podem usar programes especialitzats o qualsevol processador de textos que permeta *exportar* (realment *imprimir*) el nostre document a PDF.

4.6 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 plans (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 d'entrada per a un altre programa, o textos molt senzills on el format no és crucial, com un missatge electrònic senzill.

[17] però no lliure ni de codi font obert

[18] Hi ha alternatives lliures i de codi font obert com ara Sumatra PDF, Evince, Okular, etc. Fins i tot, els mateixos navegadors vénen ja amb visors de PDF.

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 processador 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"); la secció 4.7.1 descriu alguns problemes derivats d'aquesta tendència.

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 sobreescriure. 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.[19] 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.

[19] Aquestes tecles i combinacions de tecles que permeten un accés ràpid a operacions rutinàries se solen anomenar en anglés *hotkeys*; per exemple, en Windows, la combinació control–X retalla el text seleccionat, la combinació control–V insereix un text prèviament retallat, etc.

Sobre la cerca de paraules. Alguns processadors de textos permeten buscar usant les anomenades *expressions regulars*, les quals permeten, mitjançant caràcters especials anomenats *jòquers* (anglés *wildcards*), buscar tots els mots i totes les porcions de text que segueixen un patró determinat. Per exemple, una recerca amb l'expressió regular `pres*a` trobaria els mots *prea*, *presa*, *pressa*, *presssa*, etc., o l'expressió regular `<[^>]+>` que trobaria totes les etiquetes de l'estil de XML, ja que comencen per <, tenen un o més (+) caràcters que *no* (^) són >, i acaben amb >. Per saber més sobre expressions regular podeu consultar la pàgina de la Viquipèdia `https://ca.wikipedia.org/wiki/Expressi%C3%B3_regular`.

4.7 Contingut, estructura i presentació dels documents

4.7.1 El problema *wysiwyg*

La majoria dels processadors de textos actuals són *wysiwyg* en el sentit explicat més amunt: el text que s'edita es presenta gràficament en la finestra pràcticament igual com es veurà en el paper quan l'enviem a la impressora; això ha facilitat enormement l'accés de tothom als processadors de textos. Però l'esquema *wysiwyg*, completament generalitzat des de meitat dels vuitanta, té també, com veurem, els seus inconvenients. La persona escriptora tendeix a centrar-se en els atributs *visuals* del text (tipus i grandàries de lletra, marges, etc.), ja que confia que una bona *presentació* transmetrà a les persones lectores l'estructura *lògica* que la persona escriptora té al cap. Amb el document, per tant, només es guardarà aquesta informació de presentació, pràcticament sense cap indicació de l'estructura lògica dels continguts. Imaginem les següents situacions problemàtiques:

> Vladimir ha decidit que els títols de secció de l'informe anual que li han encarregat estaran en Helvetica de 14 punts, negreta i els de subsecció en Arial de 12 punts, negreta cursiva. A la seua directora no li agraden així i li'ls ha fet canviar a Lucida Sans de 14, negreta i Lucida de 12, negreta sense cursives. Com que l'informe ha d'estar acabat per a demà de matí, Vladimir es queda a l'oficina fins a les 11 de la nit, canviant un per un els tipus de lletra dels títols de seccions i subseccions. L'endemà, de matí, Marina, la directora, li passa un document amb una secció més que s'ha d'inserir entre la 4 i la 5. Vladimir no pot anar a esmorzar: ha de canviar els números de seccions i subseccions a partir de la 5 i repassar si s'ha de canviar alguna referència que es faça des d'una part del text a una secció pel seu número.
>
> Ens han encarregat traduir un text informatitzat. En la llengua d'origen és costum posar en *cursives* tant els mots estrangers

> ("*Sprachgefühl*") com els termes quan es defineixen per primera volta ("Un *octet* és..."), se sagna la primera línia de tots els paràgrafs, i els números de secció porten un punt al final ("1.1. Introducció"), però en la llengua d'arribada els termes nous van en negretes ("Un **octet** és ..."), se sagna la primera línia de tots els paràgrafs excepte la del primer paràgraf d'una secció, i els números de secció no porten punt al final ("1.1 Introducció"). El text ha estat traduït mantenint les convencions de la llengua d'origen: per a fer-lo adequat a la llengua d'arribada, ens toca, d'una banda, anar mirant un per un els segments de text en cursives, decidir si són definicions, i canviar-los a negretes si cal; d'altra banda, ens toca anar llevant el puntet final de tots els números de secció.

En aquests dos casos, si la persona que va escriure els textos només va codificar informació relativa a la presentació visual no podrem evitar fer els treballs tediosos descrits. Es podria dir que si qui escriu es deixa portar per la filosofia "what you see is what you get" acaba amb "what you see is *all* you get", és a dir, només té el que veu. Però la majoria dels processadors de textos *wysiwyg* actuals permeten un cert nivell de codificació de l'*estructura*, a través dels anomenats *estils*:[20] hi ha *estils de paràgraf* (paràgraf del cos de text, encapçalaments de diversos nivells, etc.) i *estils de caràcter* (definicions, èmfasi, èmfasi forta, text d'ordinador, etc.). A cada estil se li assignen unes determinades característiques de presentació: per exemple, els encapçalaments de nivell 2 van en Helvetica de 14 punts negreta i numerats automàticament amb el número de la secció de nivell 1, un punt i el número de la secció de nivell 2; les definicions van en negreta i l'èmfasi en cursiva, etc. El processador de textos aplica automàticament les mateixes característiques de presentació *a tots els segments del document* que tenen aquell mateix estil. Això resoldria les situacions problemàtiques explicades més amunt.

En el primer problema, si s'hagueren usat els estils com s'indica, la numeració i l'estil de les seccions es determinaria automàticament i només caldria indicar (només una vegada) quin tipus de lletra correspon als títols de secció; a més és renumerarien automàticament totes les seccions. Si les referències d'unes seccions a altres s'hagueren fet usant referències creuades simbòliques (molts processadors de textos les permeten), també s'actualitzarien automàticament.

En el segon problema, si el document haguera contingut informació sobre quins termes són definicions i quins són mots estrangers, només caldria canviar l'estil de les definicions i totes quedarien en negretes. D'altra banda

[20] Aquesta és la denominació usada per *Word* i pels processadors lliures i de codi obert *OpenOffice.org* i *LibreOffice*.

només caldria indicar que no és necessari l'últim punt en els números de secció i tots passarien automàticament al format desitjat.

Aquests exemples il·lustren la conveniència que els autors dels documents se centren més en l'estructuració lògica del contingut del document que escriuen. Després, només cal indicar al processador quina ha de ser la presentació de cada element d'aquesta estructura lògica i obtindrem la presentació desitjada.

4.7.2 Fulls d'estil

En XML i HTML, aquesta separació entre l'estructura del contingut i la presentació d'un document s'executa a través d'especificacions anomenades *fulls d'estil*. Un dels tipus més senzills de fulls d'estil són els anomenats fulls d'estil en cascada[21] (CSS, *cascaded style sheets*) que s'usen sobretot amb navegadors i HTML, encara que també es poden usar per a presentar XML directament en els navegadors.

Els fulls d'estil CSS assignen característiques de presentació a cada element del document. Per exemple, l'ordre CSS

```
h2 {   display : block ;
       font-size : large ;
       font-family : sans-serif ;
       text-align : left ;
       margin-top: 0.2cm ;
       margin-bottom : 0.2cm ; }
```

indica que tots els encapçalaments de segon nivell (`h2`) de (X)HTML es visualitzen (`display`) com a blocs de text separat (`block`), amb una grandària de lletra (`font-size`) gran (`large`) de la família *sans serif*, alineat (`text-align`) a l'esquerra, i amb màrgens superior i inferior de 0,2 cm.

Els fulls d'estil es poden usar també per a visualitzar documents XML directament en els navegadors més recents. Per exemple, podem fer que la presentació visual del missatge de correu electrònic de la figura 4.1 tinga un encapçalament amb el text "*Missatge de correu*" centrat, gran i en negretes amb aquesta ordre CSS (amb comentaris entre `/*` i `*/`):

```
EMAIL:before {                      /*Abans del EMAIL*/
   content : "Missatge de correu" ; /*El text desitjat*/
   display : block ;                /*com un bloc*/
   font-weight : bold ;             /*en negreta*/
   text-align : center ;            /*centrat */
   font-size : x-large ;            /*i amb lletra extra-gran*/
}
```

[21] Més informació en `http://www.w3c.org/Style/CSS/`.

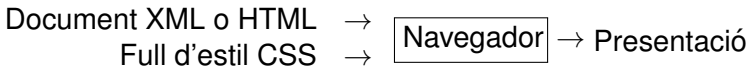


Figura 4.6: Presentació de documents XML i HTML amb fulls d'estil CSS.

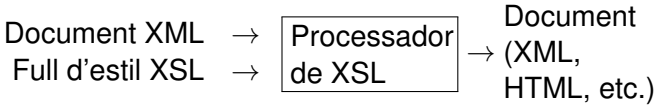


Figura 4.7: Transformació de documents XML amb fulls d'estil XSL.

En el cas dels fulls d'estil CSS, l'esquema d'ús és el que s'indica en la figura 4.6: el navegador llig el document HTML o XML, hi aplica els estils del full CSS, i genera una presentació. El full d'estil CSS pot estar en el mateix fitxer que el document HTML o XML, o en un fitxer extern.

Per a la presentació de documents XML, existeix un llenguatge de programació de fulls d'estil molt més potent que CSS anomenat XSL[22] (*extended stylesheet language*) que permet *transformar* un document XML (amb etiquetes que n'indiquen l'estructura del contingut) en un altre document XML, HTML o de qualsevol altre format (Postscript, PDF, RTF, etc.), per exemple, per a presentar-lo visualment (veure figura 4.7). Els navegadors més recents ja són capaços d'aplicar fulls d'estil XSL a pàgines *web* escrites en XML i presentar-les com si estigueren escrites originalment en HTML.

Aquest tipus de presentació visual no és l'única transformació possible que podem obtindre amb fulls d'estil: com es mostra en la figura 4.8, amb el mateix document XML podem generar un conjunt de *vistes* del seu contingut en diferents mitjans (*media*) només usant el full d'estil adequat.

4.7.3 Accessibilitat

En quasi tota la discussió anterior hem parlat de presentació referint-nos sempre a un mitjà visual, de manera que hem exclòs, per exemple, les persones que tenen discapacitats o limitacions relacionades amb el sentit de la vista (poden no veure gens o veure molt malament, o patir atacs epilèptics quan veuen una imatge que canvia ràpidament de color). Quan presentem un document visualment, intentem que la representació visual de les diverses parts del document comuniquen l'estructura lògica del contingut a la persona lectora, però, com presentem l'estructura lògica d'un document a una persona cega? Mecanismes com els tipus o grandàries de lletra o

[22] Més informació en `http://www.w3c.org/Style/XSL/`.

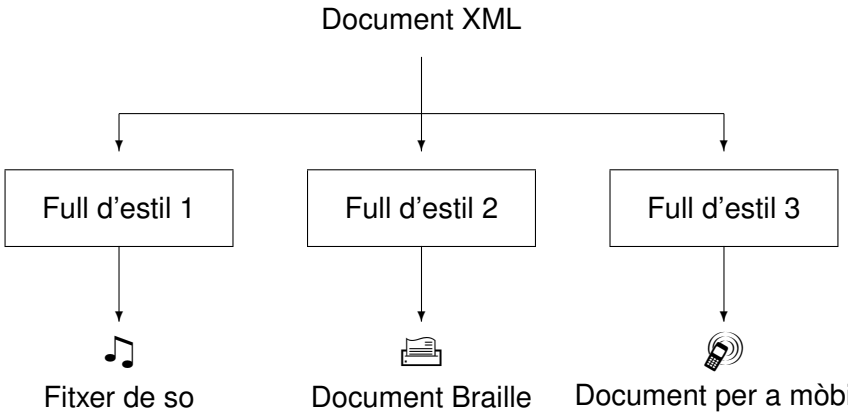


Figura 4.8: Obtenció de tres presentacions diferents d'un únic document XML mitjançant fulls d'estil.

la forma o l'alineament visual dels paràgrafs, llistes o taules no li serveixen; aquesta persona potser vol accedir als documents mitjançant un tauler Braille (una espècie de pantalla tàctil on es formen els signes de l'alfabet dels invidents, vegeu la figura 4.8) o mitjançant un sistema de síntesi de veu que llig la pàgina en veu alta.

Si qui ha escrit el document només ha codificat l'estructura lògica que tenia en el seu cap mitjançant indicadors visuals de format (negretes o cursives per a l'èmfasi, paràgrafs d'una línia en lletra més grossa per a títols, etc.) serà difícil transformar aquest format per a una altra presentació.

Però no només les persones amb discapacitats visuals poden tenir problemes; qui llig un document ho pot estar fent a través d'una pantalla de text (no gràfica) o menuda (com la d'un telèfon mòbil), o a través d'una connexió molt lenta a la xarxa, o pot estar en una situació en la qual els seus ulls estiguen ocupats (per exemple, quan condueix un vehicle). La presentació de documents en aquests mitjans té problemes similars.

Si l'èmfasi en el document ha estat emmagatzemat com a èmfasi i no amb lletra negreta o cursiva, si els títols de secció estan indicats com a tals i no perquè són paràgrafs d'una única línia en negretes grosses, serà molt més fàcil transformar-lo per a presentar-lo en un mitjà no visual o en una pantalla reduïda o limitada, per exemple, usant *fulls d'estil* especialment concebuts per a la presentació *aural* (sonora), *tàctil* (Braille), etc.

La separació de l'estructuració del contingut, d'una banda, i dels mecanismes de presentació del document, d'una altra banda, facilita l'*accessibilitat* al document a través de diversos mitjans a persones discapacitades o en situacions especials.

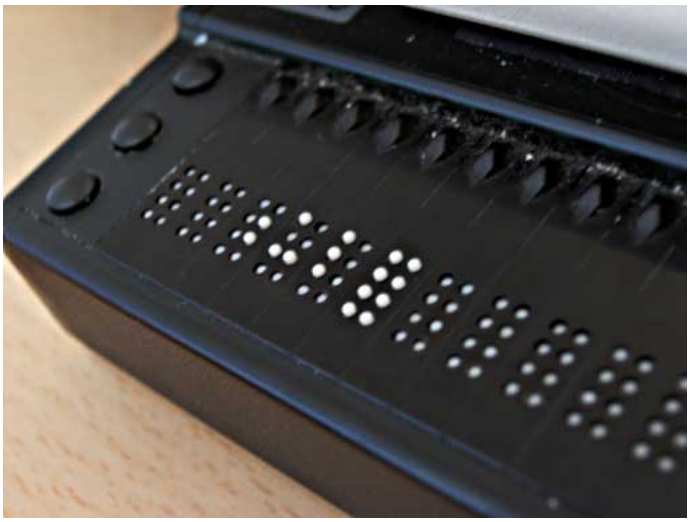


Figura 4.9: Tauler braille (imatge presa de l'entrada *Refreshable Braille Display* de la Wikipedia en anglés)

Per saber més sobre tecnologies auxiliars per a la generació de textos

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, encara que recentment s'ha incorporat bastant satisfactòriament en dispositius com ara telèfons mòbils (per exemple, els dispositius amb sistema operatiu Android permeten fer recerques per veu). En canvi, el RAP per a un propòsit específic (per exemple, la consulta telefònica d'horaris de trens o de les condicions del trànsit) 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 un dispositiu de captura 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 últims normalment han de ser *entrenats* per la persona abans del seu ús). El RAP és especialment difícil per la gran variabilitat acústica que presenten els fonemes:

- 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 òrgans 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 que ens vol dir *Per què no em passes l'altre joc de claus?*, o si sentim dir en veu alta "mu han dim moltis baltes" és molt probable que entenguem clarament "m'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). 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*; un altre doblet el formen les expressions *sax and violins on TV* i la més versemblant *sex and violence on TV*. 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.

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 (escanejat o escandit) 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 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 seqüències 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.

4.8 Qüestions i exercicis

1. Per a validar un document XML necessitem ...

   (a) ... un altre document XML, aquest últim amb les marques sense contingut.

(b) ... un full d'estil CSS.

(c) ... una definició de tipus de document (DTD).

2. Com s'indica en una DTD que l'element `teixit` conté opcionalment els elements `grandaria` i `color` en aquest ordre?

   (a) `<!MARK teixit grandaria, color #OPTIONAL>`

   (b) `<!MARK teixit (grandaria?,color?)>`

   (c) `<!ELEMENT teixit (grandaria?,color?)>`

3. Un document XML és *vàlid* ...

   (a) ... si només usa els noms d'elements definits a la DTD; la resta de les directrius de la DTD només serveixen per a fer documents *ben formats*.

   (b) ... si només usa les marques vàlides dels documents HTML.

   (c) ... si segueix les regles de la DTD quan inclou un element dins d'un altre i, a més, no inclou cap element no definit a la DTD.

4. Un text informatitzat es caracteritza principalment ...

   (a) ... pel seu format, d'una banda, i pel joc de caràcters amb què està codificat, d'altra.

   (b) ... per la versió del sistema operatiu i el processador de textos amb què ha estat escrit.

   (c) ... pel full d'estil que indica els aspectes estètics de la seua presentació.

5. Què fa que el següent fragment de XML estiga *mal format*?

```
<tit int=hi>Zjuknim agarnow</tit>
```

   (a) Entre `tit` i > no pot haver-hi res.

   (b) L'etiqueta `tit` no és vàlida en XML; hauria de ser `title`.

   (c) Si hi ha algun atribut, el valor ha d'anar entre cometes.

6. Si en una DTD trobem les regles

```
<!ELEMENT taula (capçalera?,fila+)>
<!ELEMENT fila (casella*)>
<!ELEMENT casella (#PCDATA|taula)*>
```

   quina de les tres situacions següents és vàlida d'acord amb aquesta DTD?

(a) `<taula></taula>`

(b) `<taula><fila><casella>zz<taula><fila></fila></taula>zz </casella><fila></taula>`

(c) `<taula><fila><casella>zz</casella><casella>ww</casella> </fila></taula>`

7. Què indica el fragment `encoding="...&quot;` en la primera línia (`<?xml...?>`) d'un document XML?

   (a) La versió de XML.

   (b) On és la DTD necessària per a validar-lo.

   (c) Quin és el joc de caràcters que usa el document XML.

8. Quants octets (*bytes*) ocupa el segment de XML següent:

   `<qq>ww</qq>`

   (a) 11 com a mínim, depenent de la codificació.

   (b) 11, independentment de la codificació.

   (c) 4 exactament.

9. Quan les marques de format només especifiquen el *contingut* d'un document (identificant les parts i l'estructura de cada una), com s'assigna una *presentació* determinada al document?

   (a) Amb un o més fulls d'estil.

   (b) Amb una codificació de caràcters (p.e., Unicode o ISO-8859-1).

   (c) No s'hi pot assignar presentació.

10. Què es conserva d'ASCII en els sistemes de codificació de caràcters més avançats com Unicode UTF-8, ISO-8859-1 (*Latin-1*), etc.?

    (a) Els caràcters i els seus números de codi.

    (b) Els caràcters, però amb números de codi diferents.

    (c) No en queda res. S'ha reorganitzat tota la codificació.

11. Som a Eslovàquia, on s'usa la codificació de caràcters ISO-8859-2 (*Latin-2*). Des d'Alacant, ens envien un document de text pla, escrit en codificació ISO-8859-1 (*Latin-1*) i l'obrim com si fóra ISO-8859-2 (*Latin-2*). Què passa?

    (a) No veiem bé cap lletra: tot són símbols estranys i inintel·ligibles.

    (b) Veiem bé totes les lletres excepte les accentuades, les que porten dièresi, la *ñ* o la *ç*: en el seu lloc apareixen altres símbols o lletres típiques de les llengües d'Europa de l'Est.

(c) Veiem bé totes les lletres excepte les accentuades, les que porten dièresi, la *ñ* o la *ç*: en el seu lloc apareixen les versions sense accent, la *n* o la *c*.

12. Què és RTF?

    (a) Un esquema avançat de codificació de caràcters.

    (b) Un format obert d'intercanvi de memòries de traducció.

    (c) Un format obert per a intercanviar documents de text entre processadors de textos.

13. Un document HTML té un enllaç amb el text "Més informació" i amb URI de destinació `http://www.detalls-e.com/mes.html`. Com és aquest enllaç en HTML?

    (a) `<a href="http://www.detalls-e.com/mes.html">Més informació</a>`

    (b) `<a href="Més informació">http://www.detalls-e.com/mes.html</a>`

    (c) `<a htxt="Més informació" href="http://www.detalls-e.com/mes.html">`

14. On va el títol d'un document HTML (el que es mostra en la *barra* del navegador)?

    (a) En un element `title` dins de `head`.

    (b) En un element `title` dins de `body`.

    (c) En un element `h1` dins de `head`.

15. Si els caràcters d'un text estan codificats usant el joc de caràcters ISO-8859-1 (*Latin-1*), quins codis tenen les lletres de la `A` a la `Z`?

    (a) Depén del format del text (HTML, etc.).

    (b) Els mateixos que en la codificació ASCII.

    (c) Els que tenien en la codificació ASCII més 128.

16. Quants octets (*bytes*) ocupa com a mínim el següent fitxer HTML?

```
<html><body><p>Text.</p></body></html>
```

    (a) 11

    (b) 38

    (c) 76

17. En la codificació de caràcters ISO-8859-1 (*Latin-1*), tots els caràcters accentuats de l'espanyol o del català tenen codis ...

(a) ... entre 0 i 127.

(b) ... entre 128 i 255.

(c) ... més grans que 256.

18. Un text codificat en ISO-8859-1 (*Latin-1*) té 1000 caràcters justos (comptant els espais en blanc i els salts de línia). Quants octets (*bytes*) ocupa?

    (a) 1000 exactament, 1 per caràcter.

    (b) 2000 exactament, 2 per caràcter.

    (c) entre 1000 i 2000, entre 1 octet i 2 octets per caràcter.

19. En XML, si s'obri un element amb la marca `<frase>`, amb quina marca es tanca?

    (a) Amb `</frase>`.

    (b) Amb `<frase>`.

    (c) Automàticament quan s'obri qualsevol altre element.

20. Si en un document XML trobem la situació

    ```
    <rec><id>Zork</id><addr>Zmeggs</addr></rec>
    ```

    i una DTD defineix l'element `rec` amb la regla

    ```
    <!ELEMENT rec (id,up?,addr*)>
    ```

    Pot ser que el document siga vàlid segons la DTD?

    (a) Depén de com siga d'estricte el programa validador.

    (b) No, perquè aquesta situació no és vàlida.

    (c) Sí, si la resta del document és vàlida.

21. Què veiem si obrim un text HTML amb un editor de textos senzill com el *Bloc de notes*, *Llibreta* o *Notepad* de Windows?

    (a) El text HTML però sense les marques entre "<" i "/>".

    (b) El text HTML tal com està fet per dins, amb les marques entre "<" i ">" i tot.

    (c) Una pantalla en blanc.

22. En què es diferencien dues extensions d'ASCII diferents?

    (a) En els caràcters assignats als 256 codis.

    (b) En els caràcters assignats als codis del 0 al 127.

(c) En els caràcters assignats als codis del 128 al 255.

23. Si en una DTD trobem la regla

```
<!ELEMENT cv (nom,any?,ob+)>
```

quina de les tres situacions següents no és vàlida d'acord amb aquesta DTD?

(a) `<cv><nom>Pere</nom><any>1992</any></cv>`

(b) `<cv><nom>Pere</nom><ob>Escrits</ob></cv>`

(c) `<cv><nom>Pere</nom><any>1992</any><ob>Crits</ob><ob>Plors</ob></cv>`

24. En un document HTML volem que la frase *aquest document* siga un enllaç al document que té l'URI `http://www.uc.za/t.html`: quina de les següents porcions de HTML és la correcta?

(a) `<a url="http://www.uc.za/t.html">aquest document</a>`

(b) `<a href="http://www.uc.za/t.html">aquest document</a>`

(c) `<link url="http://www.uc.za/t.html">aquest document</link>`

25. El fragment de document HTML "`<strong><em>link</strong></em>`" té una errada. Quina n'és la causa?

(a) El nom de les marques no és vàlid, perquè no n'indica cap informació sobre el contingut.

(b) L'ordre de les marques d'obertura i clausura no és correcte.

(c) No s'ha indicat el valor de l'atribut `href` de l'element `em`.

26. La longitud mitjana d'un mot en gondavés és de 5,5 caràcters i l'edició electrònica de *Gundhawól Vlâj* ("La Veu de Gondàvia"), té uns 100.000 mots diaris com a mitjana. Si el gondavés s'escriu usant la codificació ISO-8859-1 (*Latin-1*), quants exemplars del diari es poden guardar en un CD-ROM?

(a) Més de dos anys.

(b) Un exemplar només.

(c) Un mes aproximadament.

27. Com s'indica en una DTD que l'element `<fitxa>` conté obligatòriament els camps `<nom>` i `<tel>` i, opcionalment, el camp `<email>`?

(a) `<!ELEMENT fitxa (nom,tel,email?)>`

(b) `<!ELEMENT fitxa (nom,tel,email+)>`

(c) `<!ATTLIST fitxa nom CDATA #required tel CDATA #required email CDATA>`

28. En XML, què vol dir `<mang/>`?

    (a) No vol dir res, perquè no hi ha cap element que es diga `mang`.

    (b) El mateix que `<mang></mang>`.

    (c) No vol dir res, perquè hauria de ser `</mang>`.

29. Quin d'aquests tres elements XHTML (o HTML) no pot anar dins de l'element `body`?

    (a) `meta`.

    (b) `img`.

    (c) `h1`.

30. Tenim un document XML que és vàlid d'acord amb una determinada DTD. Esborrem un element complet (per exemple `<element>text</element>`), el qual no és l'element arrel del document. Quina d'aquestes tres situacions no és possible?

    (a) Que el document XML resultant no siga vàlid respecte de la DTD.

    (b) Que el document XML siga vàlid respecte de la DTD.

    (c) Que el document XML no siga un document XML ben format.

31. En HTML, podem posar un enllaç `<a href="...">...</a>` dins d'un element de llista `<li>...</li>`?

    (a) Sí.

    (b) No, perquè el document estaria mal format.

    (c) No, perquè el document no seria vàlid.

32. Un fitxer de text escrit en anglés conté només caràcters ASCII. L'obrim amb un editor i el guardem en format Unicode UTF-8. Ara ocupa ...

    (a) ... el doble d'espai.

    (b) ... exactament el mateix espai.

    (c) ... la meitat d'espai.

33. Assenyala el fragment d'HTML que generarà el text més gran:

    (a) `<h1>Títol</h1>`

    (b) `<h2>Títol</h2>`

    (c) `<h3>Títol</h3>`

34. Com es diu el joc de caràcters estàndard universal, el que assigna un número de codi diferent i únic a cada un dels caràcters de cada una de les llengües del món?

    (a) Unicode.

    (b) ISO-8859-1 (*Latin-1*).

    (c) XML.

35. Quan és preferible utilitzar el joc de caràcters Unicode en lloc de l'ISO-8859-1 (*Latin-1*)?

    (a) Quan anem a mesclar text en diferents idiomes.

    (b) Quan un text en espanyol té molts accents.

    (c) Quan el text només s'usarà en una situació d'assimilació.

36. Podem emmagatzemar correctament el text és espanyol "*La España de charanga y pandereta, cerrado y sacristía, devota de Frascuelo y de María, de espíritu burlón y de alma quieta*" usant el joc de caràcters ASCII?

    (a) Sí, sense problemes.

    (b) Sí, si abans l'hem convertit a UTF-8.

    (c) No.

37. Quant ocupa un fitxer de text que conté 2000 caràcters pertanyents a l'alfabet espanyol?

    (a) Si s'ha codificat en ISO-8859-1 (*Latin-1*), entre 2000 i 4000 octets.

    (b) Si s'ha codificat en ASCII, 2000 octets.

    (c) Si s'ha codificat en UTF-8, entre 2000 i 4000 octets.

38. Si en visualitzar el document de text `noticia.txt` en el navegador hi apareixen paraules com `camiÃ³n` e `InformÃ¡tica`, quina en pot ser la causa?

    (a) Un error per part de la persona que ha escrit el document de text.

    (b) Que el navegador està usant per llegir el document un joc de caràcters diferent del que es va usar en escriure'l.

    (c) Que el document no és un document HTML.

4.9 Solucions

1. (c)

2. (c)

3. (c)
4. (a)
5. (c)
6. (c)
7. (c)
8. (a)
9. (a)
10. (a)
11. (b)
12. (c)
13. (a)
14. (a)
15. (b)
16. (b)
17. (b)
18. (a)
19. (a)
20. (c)
21. (b)
22. (c)
23. (a)
24. (b)
25. (b)
26. (a)
27. (a)
28. (b)
29. (a)

30. (c)
31. (a)
32. (b)
33. (a)
34. (a)
35. (a)
36. (c)
37. (c)
38. (b)

Capítol 5

Bases de dades

Els gestors de terminologia són uns dels programes més usats pels traductors per a la traducció humana assistida per una màquina (MAHT; vegeu l'apartat 6.4). Com que es tracta d'un cas especial d'allò que s'anomena en informàtica *bases de dades*, en aquest capítol introduirem primer aquest concepte i llavors n'estudiarem l'aplicació a la gestió terminològica a través de les bases de dades terminològiques.

5.1 Què és una base de dades?

Com s'explica en la pàg. 12, un *fitxer* és un conjunt 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 en 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* (BD); les fitxes s'anomenen *registres* i cada element d'informació de la fitxa s'anomena *camp*.

Més generalment, Una BD es pot veure (en l'anomenat model *pla* o *de taules*) com un conjunt de taules en les quals les columnes o *camps* guarden valors del mateix tipus i on els elements d'una fila o *registre* estan relacionats entre ells. Així, una taula és un conjunt de registres (fitxes) cada un dels quals té la mateixa estructura de camps (informacions), emmagatzemades en un fitxer informàtic.

Així, els registres d'una base de dades bibliogràfica contenen informació en camps: un per als autors, altre per al títol, etc. D'altra banda, els registres de les bases de dades usades per a la gestió d'un videoclub conte-

nen camps per a emmagatzemar la informació referent als socis (com ara, el nom, el telèfon o el domicili), a les pel·lícules (el títol, la persona que l'ha dirigida o el nombre de còpies disponibles) i als préstecs (les dades d'inici i termini del préstec o el preu de lloguer). Els camps poden ser de diversos *tipus*, segons la naturalesa de les dades que s'hi guarden (cadenes de caràcters, valors numèrics enters o amb decimals, dates de calendari, etc.)

5.2 Operacions amb bases de dades

Les operacions més comunes que es realitzen sobre una base de dades també són similars a aquelles que podem fer amb un fitxer de fitxes de cartolina, però la gestió és més senzilla i són possibles molts més usos:

- *Creació de l'estructura de la base de dades*: definir cada taula, definint l'estructura de les fitxes i el tipus de dades de cada camp.
- *Altes o addicions:* afegir un nou registre a la base de dades, tot fent que els seus camps prenguen els valors corresponents.
- *Baixes o esborrats:* eliminar un o més registres.
- *Modificacions:* canviar el valor d'un o més camps d'un o més registres de la base de dades.
- *Recerques o consultes:* buscar un o més registres que compleixen un determinat criteri de recerca. 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 d'un determinat camp de totes les fitxes amb un cert valor o patró (per exemple, els autors que comencen per *Ant*) i llistar el contingut d'un altre camp (per exemple, el títol) per a cada fitxa coincident.

  La combinació de diverses estratègies de recerca permet l'*explotació* de les dades emmagatzemades en una base de dades. Per exemple, es pot generar, a partir d'un fitxer bibliogràfic, les referències bibliogràfiques citades en un text ordenades alfabèticament i en el format requerit per una determinada revista, o generar, a partir d'una base de dades de clients, una carta de recordatori per als morosos, amb detalls sobre els seus deutes.

El programa que permet fer, entre altres, aquestes operacions de manera senzilla o fins i tot automàtica (aspecte molt important quan la base de dades conté milers de registres) és un programa *gestor de bases de dades*.[1]

[1] De vegades, quan es parla desacuradament, s'anomena per metonímia *base de dades* al programa gestor.

Normalment, els usuaris reals no executen un programa gestor de bases de dades universal o genèric, 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; també poden fer ús de programes que internament inclouen un gestor de bases de dades o l'invoquen. En particular, és possible organitzar les bases de dades de manera que estiguen instal·lades en un o més servidors i que es puguen consultar i explotar des d'un altre ordinador a través d'Internet.

5.2.1 Recerques

Quan volem buscar una determinada informació en un fitxer de fitxes de cartolina (per exemple, quins autors han usat el mot *arbre* en el títol de les seues obres), i aquest fitxer no està ordenat d'acord amb 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*.

En 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, perquè després de cada consulta es redueix a la meitat la grandària del fitxer que cal explorar. Per exemple, si el fitxer té 1234 fitxes, hi ha prou amb $n = 11$ recerques perquè $2^{10} = 1024$ i $2^{11} = 2048$.

Per descomptat, és possible calcular el nombre màxim de consultes de manera més pedestre, dividint el nombre de fitxes per 2 i posant-nos en el cas pitjor. En l'exemple de les 1234 fitxes:

- Després de la 1a consulta, si la fitxa central no és la que busquem, ens queden dues meitats: una de 616 fitxes, i una altra de 617. Imaginem que anem al pitjor cas: 617 fitxes.
- Després de la 2a consulta, si la fitxa central no és la que busquem, ens queden dues meitats de 308 fitxes.
- 3a consulta: o la trobem, o hem de buscar en 154 fitxes.
- 4a consulta: o la trobem, o hem de buscar en 77 fitxes.
- 5a consulta: o la trobem, o hem de buscar en 38 fitxes.
- 6a consulta: o la trobem, o hem de buscar en 19 fitxes.

- 7a consulta: o la trobem, o hem de buscar en 9 fitxes.
- 8a consulta: o la trobem, o hem de buscar en 4 fitxes.
- 9a consulta: o la trobem, o hem de buscar en 2 fitxes.
- 10a consulta: o la trobem, o hem de buscar en 1 fitxa.
- 11a consulta: o és la que busquem, o no hi és.

Total, 11 consultes com a molt.

Un altre exemple: la taula 5.1 il·lustra gràficament el procés de recerca dicotòmica sobre una llista de cognoms ordenats alfabèticament. Per a cada pas de la cerca, s'hi mostra una taula on l'element consultat en cada moment està en negretes i la part de la llista descartada està ombrejada. Imaginem, en primer lloc, que volem buscar l'element "Garrido". Inicialment, mirem l'element d'enmig (el 13) de la llista sencera (la llista que inclou els elements de l'1 al 25) i hi trobem "Larrañaga"; com que ens hem passat, ens quedem amb la meitat baixa de la llista (que inclou els elements de l'1 al 12) oblidant-nos de l'altra meitat. Ara repetim el procés i mirem l'element d'enmig (el 6) de la nova subllista i hi trobem l'element "Esteve"; com que és menor alfabèticament que l'element que estem buscant, ens quedem amb la meitat alta (que inclou els elements del 7 al 12) de la subllista actual. Tornem a repetir el procés, aquesta vegada considerant només la subllista d'elements entre el 7 i el 12; tenim sort i en mirar l'element central (el 9) ens adonem que coincideix amb l'element buscat: la recerca, doncs, acaba amb èxit. Si la recerca haguera estat de l'element "González", els passos anteriors haurien estat els mateixos, però a més hauríem mirat l'element 11, l'element 10 i, finalment, hauríem conclòs que l'element no es troba a la llista. En aquest cas, el nombre de consultes hauria estat de 5 i, doncs, es compleix l'afirmació anterior: la recerca dicotòmica mira com a molt n fitxes si el fitxer té entre 2^{n-1} i 2^n fitxes; ací el nombre d'elements total (25) està entre $2^4 = 16$ i $2^5 = 32$, i el nombre d'elements consultats ha estat $n = 5$.

La recerca dicotòmica és només una de les moltes tècniques que es poden fer servir per a accelerar les consultes que es refereixen a un camp ordenat; el programa gestor de bases de dades pot usar una altra tècnica de recerca depenent del seu disseny o del tipus de camp.

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; amb un sistema de bases de dades no cal fer aquesta duplicació: només cal marcar els camps pels quals buscarem més freqüentment (els quals de vegades se solen anomenar *índexs*), i el programa gestor de bases de dades *indexarà* la base de dades per a permetre recerques ràpides per aquests camps.

(a) Estat abans de començar la recerca.

1 Beléndez	2 Canals	3 Carrasco	4 Escolano	5 Espí
6 Esteve	7 Forcada	8 Garcia	9 Garrido	10 Gómez
11 Guardiola	12 Iturraspe	13 Larrañaga	14 Marco	15 Micó
16 Montserrat	17 Muñoz	18 Nogueroles	19 Odriozola	20 Oncina
21 Ortiz	22 Pastor	23 Pérez	24 Sempere	25 Zubizarreta

(b) Primer pas: mirem l'element 13 i ens quedem amb la meitat baixa de la llista.

1 Beléndez	2 Canals	3 Carrasco	4 Escolano	5 Espí
6 Esteve	7 Forcada	8 Garcia	9 Garrido	10 Gómez
11 Guardiola	12 Iturraspe	**13 Larrañaga**	14 Marco	15 Micó
16 Montserrat	17 Muñoz	18 Nogueroles	19 Odriozola	20 Oncina
21 Ortiz	22 Pastor	23 Pérez	24 Sempere	25 Zubizarreta

(c) Segon pas: mirem l'element 6 i ens quedem amb la meitat alta de la subllista anterior.

1 Beléndez	2 Canals	3 Carrasco	4 Escolano	5 Espí
6 Esteve	7 Forcada	8 Garcia	9 Garrido	10 Gómez
11 Guardiola	12 Iturraspe	13 Larrañaga	14 Marco	15 Micó
16 Montserrat	17 Muñoz	18 Nogueroles	19 Odriozola	20 Oncina
21 Ortiz	22 Pastor	23 Pérez	24 Sempere	25 Zubizarreta

(d) Tercer i últim pas: mirem l'element 9 i trobem l'element buscat.

1 Beléndez	2 Canals	3 Carrasco	4 Escolano	5 Espí
6 Esteve	7 Forcada	8 Garcia	**9 Garrido**	10 Gómez
11 Guardiola	12 Iturraspe	13 Larrañaga	14 Marco	15 Micó
16 Montserrat	17 Muñoz	18 Nogueroles	19 Odriozola	20 Oncina
21 Ortiz	22 Pastor	23 Pérez	24 Sempere	25 Zubizarreta

Taula 5.1: Exemple de recerca dicotòmica sobre una llista de cognoms ordenada alfabèticament. L'element a cercar és "Garrido". Els elements ombrejats han sigut descartat durant la recerca, és a dir, sabem que l'element buscat no hi serà entre ells.

Per saber més sobre la indexació de bases de dades

Si les consultes més freqüents d'una base de dades es refereixen a un camp —o a una combinació de camps, com ara el dia, el mes i l'any que formen una data— 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 (Tot i que la duplicació d'una base de dades menuda pot no semblar en principi problemàtica, les coses canvien si considerem una base de dades amb milers de registres i amb un gran nombre de camps, tots plens d'informació. És evident, doncs, que cal establir un sistema alternatiu per poder fer recerques ràpides per més d'un camp). 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 en 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*. Els índexs s'han de refer parcialment quan es fan altes, baixes o modificacions de registres 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; el gestor de la base de dades crearà automàticament els índexs corresponents.

Considereu la base de dades amb 10 registres de la taula següent:

NÚM	BASC	SERBO-CROAT	CATALÀ
1	bat	jedin	un
2	bi	dva	dos
3	hiru	tri	tres
4	lau	četiri	quatre
5	bost	pet	cinc
6	sei	šest	sis
7	zazpi	sedam	set
8	zortzi	osam	vuit
9	bederatzi	devet	nou
10	hamar	deset	deu

Com hi podeu veure, la base de dades conté 10 registres amb 4 camps; els registres estan ordenats pel camp "`NÚM`", que indica la posició de cada registre. Si volem fer consultes ràpides pels camps "`BASC`" i "`SERBO-CROAT`" sense haver de visitar (en el cas pitjor) tots els registres, el gestor de bases de dades ha de definir un índex per a cada un d'aquests camps com els de les taules següents, que mostren els índexs corresponents als camps "`SERBO-CROAT`" (esquerra) i "`BASC`" (dreta) de la base de dades de la taula anterior:

SERBO-CROAT	NÚM
četiri	4
deset	10
devet	9
dva	2
jedin	1
osam	8
pet	5
sedam	7
šest	6
tri	3

BASC	NÚM
bat	1
bederatzi	9
bi	2
bost	5
hamar	10
hiru	3
lau	4
sei	6
zazpi	7
zortzi	8

Hi podeu comprovar com cada fitxa de l'índex conté només el camp indexat i una referència a la seua posició en la base de dades. Si penseu en el cas que la base de dades de la taula anterior tinguera, diguem-ne, 20 camps més (amb els equivalents en 20 llengües més), us podeu fer una idea de l'estalvi que s'aconsegueix respecte a la duplicació sencera.

5.3 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 professionals de la traducció gestionen, amb aquests gestors, bases de dades lèxiques que els ajuden a traduir consistentment els termes i, potser, les locucions i frases d'una determinada àrea de coneixement (terminologia).

Els registres o les fitxes d'una base de dades lèxica multilingüe estableixen correspondències entre els termes usats en diverses llengües en un camp determinat de coneixement. En aquestes bases de dades, *cada fitxa representa un concepte* i conté un camp índex per a emmagatzemar el terme corresponent en cada llengua. Aquesta organització (una fitxa per concepte) és coherent amb la definició de terminologia com la disciplina l'objecte de la qual és l'estudi sistemàtic de l'etiquetatge o designació de *conceptes* particulars d'una o més àrees temàtiques o d'un o més àmbits de l'activitat humana amb el propòsit de documentar i promoure l'ús correcte;[2] a més, és especialment adequada quan la base de dades terminològica és multilingüe.

Les bases de dades lèxiques o terminològiques 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).

[2] `https://ca.wikipedia.org/wiki/Terminologia`

- 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 data de creació i de modificació de la fitxa.
- La definició del terme en una o més llengües. La major part dels gestor permeten que els termes usats en la definició d'un determinat terme es marquen *remissions*, és a dir, enllaços actius a les fitxes on es defineixen.
- El camp temàtic de la fitxa.
- Altres termes relacionats.
- Informació sobre la morfologia o la flexió del terme en cada una de les llengües.
- Variants ortogràfiques o geogràfiques; sinònims; antònims; etimologia, etc.

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, molts programes de traducció assistida per ordinador (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.

També hi ha bases de dades terminològiques que es poden consultar en línia:

- L'Institut d'Estudis Catalans manté el TERMCAT (`http://www.termcat.cat`). La intenció de TERMCAT és principalment normativa: s'hi associa a cada concepte el terme preferit en català (i també els termes usuals en espanyol, anglés, francés, etc.).
- IATE (*Inter-Active Terminology for Europe*, `http://iate.europa.eu/`) és la base de dades terminològica de referència de la Unió Europea i proporciona termes per a cada concepte en les 24 llengües oficials de la Unió Europea i en llatí. La base de dades, en format TBX (vegeu l'apartat 5.3.1) es pot descarregar totalment o parcialment per a usar-la fora de línia.

5.3.1 L'intercanvi de bases de dades terminològiques

La creació i el manteniment d'una bona base de dades terminològica requereix un gran esforç i moltes hores de treball. Això la converteix en un recurs valuós i sovint els traductors en fan intercanvi. Però la informació emmagatzemada sobre cada terme pot ser molt diferent d'una base de dades a una altra i, per tant, fer servir una base de dades terminològica en diferents sistemes alhora no és una tasca fàcil. Afortunadament, en els darrers anys s'ha desenvolupat un conjunt de formats estàndards per facilitar aquest intercanvi; un dels més coneguts es TBX[3], (per *TermBase exchange*, "intercanvi de bases de dades terminològiques"), encara que hi ha altres com OLIF (*Open Lexicon Interchange Format*). Amb el temps, els programes que inclouen una base de dades terminològica van incorporant la capacitat de llegir i escriure documents en format TBX.

Per saber més sobre el format TBX

El format TBX segueix les especificacions XML (vegeu l'apartat 4.4); és a dir, els documents TBX són un tipus de document XML definit per una DTD concreta. A més, també segueix les directrius de l'estàndard ISO 12620, que defineix un conjunt de camps, i els seus possibles valors, per a la informació terminològica.

Heus ací un exemple de document TBX:

```
<?xml version='1.0'?>
<!DOCTYPE martif SYSTEM  "./TBXcoreStructureDTD-v-1-0.DTD">
<martif type='TBX' xml:lang='en' >
 <martifHeader>
  <fileDesc>
   <sourceDesc>
    <p>from an Oracle corporation termBase</p>
   </sourceDesc>
  </fileDesc>
  <encodingDesc>
   <p type='DCSName'>TBXdefaultXCS-v-1-0.XML</p>
  </encodingDesc>
 </martifHeader>
 <text>
  <body>
    <termEntry id='eid-Oracle-67'>
      <descrip type='subjectField'>
        manufacturing
      </descrip>
      <descrip type='definition'>
        A value between 0 and 1 used in ...
      </descrip>
      <langSet xml:lang='en'>
       <tig>
```

[3]L'URI per a més informació és `http://www.lisa.org/tbx/`.

```
        <term tid='tid-Oracle-67-en1'>
          alpha smoothing factor
        </term>
        <termNote type='termType'>
          fullForm
        </termNote>
       </tig>
      </langSet>
      <langSet xml:lang='hu'>
        <tig>
         <term tid='tid-Oracle-67-hu1'>
           Alfa sim&#x00ED;t&#x00E1;si t&#x00E9;nyez&#x0151;
         </term>
        </tig>
      </langSet>
    </termEntry>
  </body>
 </text>
</martif>
```

La informació de cada terme (en l'exemple només un) s'inclou dins de l'element `termEntry`, que conté descripcions (`descrip`) sobre el domini d'ús i la definició del terme, a més d'una secció `langSet` per a cada idioma (ací, anglés i hongarés) on s'especifica el terme (`term`) i informació sobre ell (`termNote`). Abans del cos (`body`), l'element arrel `martif` conté una capçalera (`martifHeader`) amb informació sobre l'origen de la base de dades. El terme hongarés és "Alfa simítási tényező"; en el document d'exemple, els caràcters especials s'indiquen amb el seu codi Unicode (per exemple, `&#x0151;` és el caràcter "ő")

5.4 Qüestions i exercicis

1. Tenim una base de dades amb fitxes de 100 alumnes ordenada pel NIF. Si busquem una alumna pel NIF, quantes consultes fa, com a màxim, el programa gestor de la base de dades fins a arribar a la fitxa desitjada?

   (a) 100.

   (b) La meitat, 50.

   (c) 7, perquè $2^6 = 64 < 100 < 2^7 = 128$.

2. Volem tenir una base de dades ordenada simultàniament per dos camps per a optimitzar les recerques. És això possible? Com?

   (a) No és possible.

   (b) Sí, però és necessari duplicar totes les fitxes en memòria.

   (c) Sí. S'han de construir dos índexs, un per a cada camp.

3. És necessari tenir dues còpies de totes les fitxes d'una base de dades per a poder-la tenir ordenada per més d'un camp?

(a) Depén; si els camps són numèrics, no és necessari.

(b) No hi ha cap altre remei: cal duplicar la base de dades.

(c) No; s'hi pot crear més d'un *índex*.

4. Quan usem un programa gestor de bases de dades terminològiques per a buscar un terme en una base de dades...

   (a) ... alguns gestors ens permeten buscar-lo sense conéixer-ne la forma exacta.

   (b) ... hem de conéixer com s'escriu exactament el terme per a poder trobar-lo.

   (c) ...sempre és necessari conéixer-ne la categoria lèxica i indicar-la al gestor

5. Si busquem en una base de dades de 240 fitxes d'alumnes una fitxa (un registre) pel número de telèfon, un camp pel qual no està ordenada, quantes consultes haurà de fer, com a màxim, el programa gestor de la base de dades fins arribar a la fitxa desitjada?

   (a) 240

   (b) 120

   (c) 9

6. Quan volem tenir els registres (les fitxes) d'una base de dades ordenada simultàniament per dos camps per a optimitzar les recerques construïm dos índexs. Els índexs contenen...

   (a) Entrades compostes pel valor del camp índex i el número de la fitxa (del registre).

   (b) Només els números de les fitxes ordenats segons el valor del camp.

   (c) Una còpia de les fitxes ordenades pel camp índex corresponent.

7. Què tenen en comú tots els camps d'una fitxa terminològica?

   (a) Es refereixen al mateix concepte.

   (b) Es refereixen al mateix terme.

   (c) Es troben en el mateix índex.

8. En què es diferencia un camp clau o camp índex de la resta dels camps d'una fitxa?...

   (a) Es guarda en un tipus de memòria RAM més ràpida anomenada *cache* o *memòria cau*.

(b) La manera d'emmagatzemar el camp és diferent (els camps índex o clau s'emmagatzemen de manera comprimida i els altres no).

(c) Les recerques de fitxes per aquest camp són molt més ràpides que les que es facen per camps que no són clau o índex.

9. Quan es duplica el nombre de fitxes d'una taula determinada d'una base de dades, què succeeix amb el nombre de consultes que realitza una recerca dicotòmica?

   (a) Es duplica.

   (b) Es queda exactament com està.

   (c) S'incrementa en 1 per terme mitjà.

10. En una base de dades lèxica o terminològica amb 2.000 fitxes, quan demanem al programa gestor que busque un determinat terme en una determinada llengua, quantes consultes ha de fer en el pitjor cas per a entregar-nos la fitxa?

    (a) Ha de fer, necessàriament, 2.000 consultes perquè hi ha 2.000 fitxes.

    (b) No ha de fer cap consulta, va directament a la fitxa.

    (c) Depén. Si està indexada pel terme en aquella llengua, farà com a molt 11 consultes, i si no, com a molt 2.000 consultes.

11. En una base de dades terminològica, cada concepte ...

    (a) ...es correspon amb un camp d'un registre general.

    (b) ...es correspon amb múltiples registres, un per cadascun dels termes usats en cada llengua per a representar aquest concepte.

    (c) ...es correspon amb un registre en el qual es troben els termes usats en cada llengua per a representar aquest concepte.

12. Hem cronometrat el temps que tarda un programa gestor de bases de dades en trobar la fitxa que té un determinat valor per un camp determinat, quan augmenta el nombre de fitxes. Els resultats són:

NOMBRE DE FITXES	TEMPS
1.000.000	4,9 s
2.000.000	5,1 s
3.000.000	5,3 s
4.000.000	5,4 s
6.000.000	5,5 s

    Què podem dir de la base de dades?

(a) Que no està ordenada pel camp pel qual estem buscant.

(b) Que usa XML per a obtenir una velocitat acceptable.

(c) Que està ordenada pel camp pel qual estem buscant.

13. El programa gestor d'una base de dades lèxica feia en el pitjor cas 480 consultes abans de trobar la fitxa corresponent a un terme anglés concret, abans d'ordenar-la pel terme en anglés. Després d'ordenar-la fa

(a) 240, la meitat.

(b) moltes menys consultes: 9.

(c) 480 consultes igualment.

5.5 Solucions

1. (c)
2. (c)
3. (c)
4. (a)
5. (a)
6. (a)
7. (a)
8. (c)
9. (c)
10. (c)
11. (c)
12. (c)
13. (b)

Capítol 6

La traducció automàtica i les seues aplicacions

Una de les aplicacions més importants de la informàtica a la traducció és la *traducció automàtica* (TA). Però abans de considerar l'automatització de la traducció i les seues aplicacions fóra bo que ens paràrem un poc per a discutir què vol dir exactament el mot *traducció*. Més endavant, en la secció 9.3, discutirem sobre la relació entre traducció humana i traducció automàtica.

6.1 Què és la traducció?

Per començar, s'ha de tenir en compte que el mot *traducció* és ambigu[1] perquè es pot referir al *procés* de traduir o al *producte* (resultat) d'aquest procés.

Sager (1993)[2] 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, etc.) de la comunitat lingüística de la LM, basat en un subllenguatge paral·lel de la LO;

[1] Com molts altres substantius acabats en *-ció*

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

- els documents i les traduccions corresponents es poden classificar en tipus[3] i aquesta tipologia afecta la traducció;
- 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*[4] 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).

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

[4] Editorial Enciclopèdia Catalana, 7a ed., 1987-

- Alcaraz Varó i Martinez Linares (1997) defineixen traducció com "expresión de un enunciado en la lengua de llegada [lengua meta] que sea 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 el context d'aquest llibre:

- 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

La traducció automàtica (TA) es pot definir com el procés (o el producte) de traduir un text informatitzat[5] 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 resultat de la traducció automàtica és normalment un producte bastant diferent a la traducció professional, i en la majoria dels casos no es pot usar en el seu lloc tal com està; per això, el capítol 6 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 4.4 i 4.5.1) 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

[5] Anomenarem *text informatitzat* un fitxer que conté un text codificat en un format conegut (vegeu el capítol 4), i que pot ser editat amb un editor o amb un processador de textos adequat.

capaços d'usar la informació de les etiquetes com a context per a elegir una traducció on hi ha més d'una alternativa.

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.

Sobre el nom en altres llengües. 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 es parla, com en català o en espanyol, de *traduction automatique*. Altres llengües, com ara el neerlandés, usen un mot compost amb la paraula *ordinador*: *Computervertaling*.

Per saber més sobre la història de la traducció automàtica

La major part del discutit d'aquest apartat està pres dels treballs de John Hutchins, especialment de Hutchins (1995) i Hutchins (2001). El Dr. John Hutchins és considerat l'historiador de la traducció automàtica, i fins fa poc mantenia activament l'arxiu `www.mt-archive.info`, on es poden trobar reproduccions facsímils de molts articles dels inicis d'aquesta disciplina.

Els pioners, fins 1954: La traducció mitjançant màquines és una ambició humana des de fa segles que no es va fer realitat fins al segle 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 del 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 (de vegades erròniament anomenada *traducció directa*, i similar als sistemes de traducció indirecta per transferència morfològica avançada, vegeu l'apartat 8.3.1). 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 6.5.3) i en l'ajuda humana en tasques de preedició i postedició (vegeu 6.5); era prou clar que els sistemes reals no podrien produir més que traduccions de molt baixa qualitat. El 1952 es va celebrar als Estats Units el primer congrés sobre TA on es van definir les línies fonamentals a seguir.

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 usant un diccionari de només 250 mots i 6 regles gramaticals; aquestes llengües van ser les elegides per raons geopolítiques per als primers sistemes de TA. 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 gran nombre 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 8).

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 —amb excepcions comptades: Masterman (1967) va estudiar la utilitat de la traducció *mot per mot* com a *pidgin*, és a dir, com a llengua de contacte, en comparació amb una traducció *nativa*. Per què pensar-hi si molt prompte 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 a tasques menys ambicioses 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.

Des de l'informe ALPAC (1966) fins 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 són, 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.

Els primers vuitanta: Als 1980 apareixen nous sistemes de TA arreu del 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 (vegeu la p. 120) gràcies al control de la llengua origen (vegeu l'apartat 6.5.3); 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 llengües escandinaves).

Durant aquest decenni els esforços es dirigeixen vers la traducció indirecta amb representacions intermèdies o sense (com la interlingua; vegeu l'apartat 8.4) mitjançant anàlisis morfològiques i sintàctiques i, de vegades, coneixements semàntics. 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, finançat per la Comunitat Europea i el projecte japonés CICC amb participants a la Xina, Indonèsia i Tailàndia.

Eurotra és un dels projectes de traducció més coneguts del decenni dels 80. 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 s'esperava que la traducció resultant seria de gran qualitat, encara es preveia una gran quantitat de postedició. El projecte, que va ser abandonat el 1992, no va ser capaç d'entregar un sistema de TA funcional, però va estimular la investigació sobre tecnologies lingüístiques a tot Europa.

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 sempre a traductors (molts dels quals no no consideren açò com a part del seu treball), sinó a persones que es presumeix 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'intel·ligència artificial al processament del llenguatge humà (sistemes experts i sistemes basats en el coneixement dissenyats per entendre els textos).

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 prompte 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 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 mètodes basats en exemples (vegeu el capítol 10) s'aprofiten també de l'e-

xistència 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 adés, 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, recentment alliberat com a programari lliure/de codi font obert amb el nom OpenTM2, `http://www.opentm2.org`, o el Translator Workbench de Trados, ara anomenat SDL Trados Studio. La major part d'aquestes estacions de treball estan disponibles per a ordinadors personals.

Dels 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, com ara Google Translate, `http://translate.google.com` o Bing Translator, `http://translator.bing.com`, d'ús molt comú per part del públic en general per a l'*assimilació* (vegeu l'apartat 6.3.1) de continguts web escrits en altres llengües i fins i tot per a la traducció de cartells i textos fotografiats amb la càmera del telèfon mòbil.

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 llengües del món; amb excepcions com ara el projecte Apertium (`http://www.apertium.org`), que ofereix traducció automàtica per a llengües menys centrals, com ara el gallec, l'occità, el bretó.

En el moment d'escriure aquestes línies (desembre de 2015), la major part dels sistemes de traducció automàtica es basen en una evolució de la *traducció automàtica estadística* iniciada en IBM durant la dècada dels 1990; se sol parlar de *traducció automàtica estadística basada en frases*, en anglés *phrase-based statistical machine translation*. Aquesta hegemonia es deu en gran part a la disponibilitat de programari lliure/de codi font obert per a *entrenar* i aplicar sistemes de traducció automàtica, com ara Moses (`http://statmt.org/moses`). Fins i tot hi ha empreses com ara KantanMT (`http://kantanmt.com`) que construeixen sistemes a mida per als seus clients usant simplement un navegador.

En els últims anys s'està investigant una nova modalitat de traducció basada en l'anomenat *aprenentatge profund*, en anglés *deep learning*, que usa mètodes d'un camp de la intel·ligència artificial anomenat *xarxes neurals*, i els resultats comencen a ser, en proves de laboratori, comparables als millors disponibles.

6.3 Utilitat de la traducció automàtica

La traducció automàtica produeix resultats que normalment no poden substituir directament els produïts per professionals de la traducció (vegeu el capítol 9). 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, 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.

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 molt diferent de les traduccions fetes per professionals competents, pot ser una eina molt útil en aquests dos grups d'aplicacions.

6.3.1 Assimilació

En situacions d'*assimilació* de la informació no sembla necessària una traducció gramaticalment correcta i similar al text que produiria una persona nativa, sinó més aïna una traducció ràpida i raonablement intel·ligible. S'ha de tenir en compte que hi ha característiques dels textos natius que poden no ser necessàries per a la comprensió. Per exemple, un text pot ser intel·ligible encara que no concorden els adjectius amb els noms o fins i tot encara que se n'hagen eliminat els articles (*A amic meu li agraden xiques vell*), o l'ordre dels mots no siga gramatical (*La guerra evitar no podrem*).[6]

Una de les primeres aplicacions de la traducció automàtica als EUA va ser l'anomenat *screening* o exploració de documents per a decidir quins eren rellevants i mereixien una atenció més detallada: 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. Altres exemples d'ús de la traducció automàtica per a l'*assimilació* són:

[6] De fet, hom podria dissenyar els sistemes de traducció automàtica perquè no es *preocuparen* d'aquests assumptes menors.

- 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, com ara *Google Translate* o *Bing Translator*).
- La traducció automàtica de *converses electròniques* interactives (usant el teclat i la pantalla d'ordinadors connectats entre si; *xat*) entre persones que parlen dos idiomes diferents. Les mancances de la traducció automàtica 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*).
- La traducció de despatxos de premsa en altres idiomes.

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.

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 (vegeu també la pàg. 8.2. A tall d'exemple, considereu el següent text en tok pisin[7] (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:[8]

> *En temps passat, un illa, enorme porc - soler viure esmentat i ell - soler menjar més-d'un home. Ell - soler menjar més-d'un*

[7] 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.

[8] 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.

> *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? Una traducció més idiomàtica podria ser:

> *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 inintel·ligible 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:

> *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 bastantes pistes sobre el significat del text original.

En els últims anys, especialment des que s'ha generalitzat l'accés públic a Internet, s'observa una tendència a incorporar sistemes de traducció automàtica com un dels components de sistemes més grans de comunicació. Aquesta aplicació de la TA per a l'assimilació es pot veure en *xats* bilingües, o en els sistemes que tradueixen les pàgines *web* segons anem visitant-les seguint enllaços; en aquests sistemes, la TA no s'invoca explícitament, sinó implícitament quan usem el servei.

6.3.2 Disseminació

En situacions de *disseminació* de la informació cal revisar l'esborrany de traducció produït pel traductor automàtic i fer les modificacions oportunes per convertir-la en una traducció adequada al propòsit de les traduccions. Per tal de minimitzar les modificacions a fer a la traducció automàtica, 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.[9] 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.[10]

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 des del 1982 fins al 2001 produïa informes meteorològics simultanis en francés i en anglés al Canadà.

6.4 Traducció semiautomàtica

Moltes situacions de traducció automàtica es poden classificar com a situacions de traducció assistida per ordinador (en anglés *computer-aided translation*; CAT), també anomenada de vegades *traducció semiautomàtica*. El terme *computer-aided translation* s'usa normalment per a referir-se a l'entorn de programari que permet la traducció professional amb el suport de bases de dades lèxiques (vegeu l'epígraf 5.3), i dels suggeriments de traducció provinents de memòries de traducció (vegeu el capítol 10), i fins i tot, de la traducció automàtica.

Per precisar millor què volem dir amb això d'"assistida per ordinador", es fa necessari considerar les nocions de traducció humana assistida per una màquina (en anglés *machine-aided human translation*; MAHT), i traducció automàtica assistida per un humà (en anglés *human-aided machine translation*; HAMT), 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.

MAHT: L'usuari (un traductor competent o un professional independent) utilitza diccionaris bilingües, tesaurus o *thesauri*, conjugadors i declinadors,

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

[10] Vegeu l'apartat 6.5.3, on es discuteix un concepte molt relacionat, el de *llenguatge controlat*.

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 5.3), o les memòries de traducció (vegeu el capítol 10).

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* (vegeu l'apartat 6.5) 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.

6.5 Automatització del procés de traducció

A l'hora d'abordar l'automatització del procés de traducció cal fer una anàlisis dels costos de traducció per tal d'estimar l'estalvi en recursos (com ara temps i diners) que es produirà amb la introducció de la traducció automàtica. El capítol 9 es centra en la avaluació dels sistemes de traducció automàtica i l'anàlisi de costos de traducció; en aquest apartat discutirem les diferents tasques i opcions per automatitzar el procés de traducció.

6.5.1 Postedició

La *postedició* és la modificació *mínima* d'una traducció generada per ordinador per a *fer-la adequada a un* ***propòsit*** *ben definit*: el text meta produït pel sistema es refina o revisa (*postedita*) perquè siga gramaticalment correcte o estiga escrit d'acord amb un registre determinat.

A l'hora de posteditar hem d'evitar fer canvis *preferencials* (aquesta solució adequada "m'agrada més" que aquesta altra que també és adequada). Els canvis estilístics s'han de fer estrictament quan, si no es feren, la traducció resultant no compliria amb el propòsit per al qual va ser encarregada. Les modificacions poden ser: *esborrats* d'un mot que sobra, *substitu-*

cions d'un mot per un altre, o *insercions* d'un mot que falta. Han de ser les *mínimes* necessàries: si hi ha més d'una edició possible, cal elegir la que es faça amb el mínim de modificacions necessàries.

Hem de tenir en compte que la persona posteditora, a més de conéixer la llengua meta i ser capaç de convertir el text en brut a una forma genuïna en aquesta llengua (és a dir, a més de ser professional de la traducció), ha de ser una veritable especialista en postedició, que coneix el sistema de traducció automàtica i quins en són els errors més típics. Així, la tasca de postedició és molt més eficient, ja que en conéixer l'origen i la causa dels errors se'n fa més fàcil i ràpida la correcció.

En primera aproximació, la postedició serà convenient quan

$$\textbf{cost}\left(\begin{array}{c}\text{traducció automàtica}\\ + \\ \text{postedició}\end{array}\right) < \textbf{cost}(\text{traducció professional}).$$

Cal comprovar que la fórmula anterior es compleix, encara que siga a llarg termini, abans de triar una estratègia de traducció per a la disseminació basada en la postedició.[11]

6.5.2 Preedició

La *preedició* consisteix a preparar o adaptar (*preeditar*) el text origen per a facilitar la seua traducció i millorar el comportament del sistema de traducció automàtica, reduint-ne la necessitat de postedició de la traducció en brut. Això s'aconsegueix, per exemple, eliminant l'ambigüitat del text,[12], evitant l'ús de la veu passiva, reduint l'ús d'oracions subordinades o usant frases curtes i completes sintàcticament i semànticament.[13] La preedició del text origen es pot fer 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.

La preedició sol ser tant més convenient quant a més llengües es traduïsca el text preeditat perquè un canvi al text origen pot estalviar tantes postedicions com llengües d'arribada tinguem.

En resum, hi ha tres modalitats bàsiques d'interacció entre les persones i els programes de traducció automàtica:

- la preedició (preparació del text *abans* de la traducció automàtica),
- la postedició (correcció del text *després* de la traducció automàtica) i

[11] Per a una anàlisi de costos més detallada, vegeu l'apartat 9.2.1.

[12] Per exemple, en anglés tècnic, el mot *replace* presenta una *ambigüitat lèxica* (vegeu l'apartat 7.2.1), ja que pot voler dir *exchange* (reemplaçar) o *put back* (tornar a col·locar).

[13] Kohl (2008) ofereix indicacions per a escriure textos en anglés per a una audiència global, de manera que els textos siguen més fàcils d'entendre per als no natius i més fàcils de traduir manualment i automàticament.

- la interacció de la persona amb el sistema de traducció automàtica durant el procés de traducció.

6.5.3 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 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; O'Brien 2003) 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 llenguatge controlat té sovint associat un conjunt de programes de suport que ajuden a avaluar i escriure documents que en complisquen les restriccions. L'escriptor de llenguatges controlats usa normalment un editor de textos intel·ligent que fa les següents tasques:

- Comprovar el compliment de les restriccions:
  - terminològiques (com el cas del mot *replace* esmentat més amunt; per a això, pot ser útil accedir a una base de dades terminològica, com les esmentades en el capítol 5);
  - sintàctiques (per exemple, fent l'anàlisi sintàctica de les oracions i detectant les ambigüitats estructurals, vegeu l'apartat 7.2.2), i
  - d'estil (per exemple, especificant quin ha de ser el format de les dates o de les hores).
- Emetre un missatge d'error com més informatiu millor quan es detecte una violació de les especificacions del llenguatge.
- Proposar a la persona usuària formes alternatives vàlides al text erroni.

Com es pot veure, els desenvolupaments tècnics fets al voltant del disseny d'un llenguatge controlat es relacionen amb molts conceptes que es tracten en aquest llibre.

Un exemple històric de llenguatge controlat que va ser usat per a millorar els resultats de la traducció automàtica —en concret, els obtinguts amb un sistema també històric anomenat Weidner MicroCat— és PACE (*Perkins Approved Clear English*), el llenguatge controlat usat durant els anys vuitanta i part dels noranta per l'empresa d'enginyeria Perkins Engines

per a facilitar la traducció automàtica dels manuals que descriuen les característiques i el manteniment d'aquests motors (Newton 1992; Douglas i Hurst 1996). Un dels principis de PACE és "un mot, un significat", és a dir, s'hi estableixen restriccions lèxiques clares a través d'un diccionari, cosa que simplifica el disseny dels diccionaris del sistema de traducció automàtica. A més del lèxic, PACE també especifica la sintaxi (Arnold et al. 1994, secció 8.3). Altres exemples de llenguatges controlats són l'*ScaniaSwedish* usat per la firma de camions i autobusos Scania (Almqvist i Sågvall Hein 1996), o el *Caterpillar Technical English* de la companyia de maquinària d'excavació Caterpillar.

També hi ha llenguatges controlats no específicament dissenyats per a la traducció automàtica, com ara 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[14] per mot (normalment una)", amb "l'objectiu de produir textos breus i no ambigus" (AECMA 2007).

Alguns dels avantatges de l'ús de llenguatges controlats (Schwitten 2007) 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 disseny d'un llenguatge controlat no és gens trivial: cal estudiar amb profunditat corpus de textos pertanyents al domini i prendre decisions difícils;
- 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.

[14]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.

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.

Per últim, cal deixar clar que l'ús d'un llenguatge controlat és una alternativa a la preedició dels textos, però que no elimina per complet la necessitat de postedició o, almenys, de revisió de les traduccions en brut.

6.6 Qüestions i exercicis

1. (*) Elegiu un idioma qualsevol que conegueu bé, L. És ben segur que L té mots polisèmics que en una 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 llengües.

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) ... impossibilita l'ús del sistema per a tasques de disseminació d'informació.

3. Indica en quina d'aquestes situacions de traducció automàtica són menys crucials la gramaticalitat o naturalitat lingüística de la traducció.

   (a) Joan usa el Web Translator mentre navega per les pàgines d'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 a l'espanyol 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 durant la traducció.

(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.

9. Imagineu que un traductor professional cobra 0,05 euros per mot de text traduït i que un corrector de textos cobra 0,10 euros per mot de text corregit. Imagineu que tenim un sistema de traducció automàtica

que ens costa uns 0,03 euros per mot traduït i que produeix un 10% de mots incorrectes en les traduccions. Convé adoptar-lo i contractar el corrector o és millor contractar el traductor professional? (si no sabeu calcular-ho en general, feu els càlculs amb un text de, per exemple, 1000 mots).

10. La traducció automàtica instantània de pàgines *web* durant la navegació és un cas de traducció automàtica...

    (a) ... amb preedició.

    (b) ... per a la disseminació.

    (c) ... per a l'assimilació.

11. El "control" dels llenguatges controlats...

    (a) ... es refereix tant a la terminologia com a la sintaxi.

    (b) ... només pot referir-se a la sintaxi.

    (c) ... només pot referir-se a la terminologia.

12. Quan s'usen per a la traducció, els llenguatges controlats restringeixen directament...

    (a) ... la llengua meta.

    (b) ... la llengua origen.

    (c) ... tant la llengua origen com la llengua meta.

13. Si un angloparlant usa el traductor automàtic portugués–anglés de `babelfish.altavista.com` per a llegir en línia el diari brasiler *O Globo*, està usant la traducció automàtica per a un propòsit...

    (a) ...d'assimilació d'informació.

    (b) ...de disseminació.

    (c) ...per al qual no està pensada.

14. Quina és l'alternativa estàndard a la preedició en un entorn de producció massiva de documentació multilingüe?

    (a) L'ús d'un llenguatge controlat

    (b) L'ús d'un sistema d'interlingua.

    (c) La postedició sistemàtica

15. Fran consulta a través d'Internet la base de dades terminològica IATE (vegeu l'apartat 5.3) quan tradueix dossiers antiglobalització de l'anglés al neerlandés. En quina de les tres situacions següents es troba?

(a) Traducció automàtica assistida per la persona

(b) Traducció humana assistida per la màquina

(c) Usa un llenguatge controlat

16. Si enviem un document HTML a un servidor de traducció automàtica i després posteditem el resultat perquè siga una traducció acceptable de l'original abans de publicar-la, estem usant la traducció automàtica...

(a) ... amb memòria de traducció.

(b) ... per a la disseminació.

(c) ... per a l'assimilació.

17. L'adopció d'un llenguatge controlat en una situació de traducció de documents d'una llengua a moltes llengües per a la disseminació és, en el procés complet, una alternativa a...

(a) ...la postedició repetitiva dels documents meta.

(b) ...la preedició repetitiva dels documents origen.

(c) ...la traducció de fragments ja traduïts anteriorment.

18. Un sistema que suggereix millores a l'estil d'un document es pot considerar com ...

(a) ... HAMT.

(b) ... MAHT.

(c) ... un sistema de traducció automàtica ergonòmic.

19. Una inventora monolingüe consulta documents web traduïts a la seua llengua per tal de descobrir si el seu nou invent ha estat patentat abans. Si la traducció es fa mitjançant un sistema automàtic, quin ús n'està fent?

(a) Assimilació; més concretament per a allò que es diu *screening*.

(b) Disseminació.

(c) Postedició, ja que l'idioma del document canvia per què puga ser entés.

20. El programa de la Generalitat Valenciana SALT 4.0 tradueix textos de l'espanyol a la varietat valenciana del català i pregunta esporàdicament a la persona usuària quin equivalent és més adequat per a alguns mots ambigus difícils. Aquesta és una situació de...

(a) ... postedició.

(b) ... traducció automàtica assistida per la persona.

(c) ... traducció humana assistida per l'ordinador.

21. Volem posteditar un text traduït automàticament mirant tan poc com siga possible el text original. Ens ajuda conéixer quins són els mots homògrafs (vegeu la p. 120) més comuns de la llengua origen?

    (a) No, perquè els homògrafs del text origen no afecten el text meta en brut.

    (b) No, perquè només estem mirant el text meta.

    (c) Sí, perquè són una font molt important d'errors especialment difícils de corregir si no es coneix què ha passat.

22. Una persona està escrivint un document en llengua origen que després serà traduït automàticament a més d'una llengua meta i el sistema que usa per a escriure l'avisa quan tecleja un mot que donarà problemes de traducció —i li suggereix alternatives— o quan escriu una estructura que serà difícil de traduir. Aquesta és una situació...

    (a) ...de preedició.

    (b) ...d'aplicació d'un llenguatge controlat.

    (c) ...de postedició.

23. Quina de les següents situacions és absurda en traducció automàtica?

    (a) La postedició en una aplicació d'assimilació.

    (b) La postedició en una aplicació de disseminació.

    (c) L'ús d'un llenguatge controlat en una aplicació de disseminació.

24. Només una d'aquestes tres afirmacions és certa. Quina?

    (a) Els llenguatges controlats defineixen regles de postedició.

    (b) L'ús d'un llenguatge controlat elimina completament la necessitat de postedició.

    (c) Quan s'apliquen les regles d'un llenguatge controlat, el text resultant és gramaticalment acceptable però s'hi eviten construccions i mots que donen problemes.

25. Un sistema de traducció automàtica hipotètic del rus al català produeix text que és bàsicament correcte excepte pel fet que no genera ni articles determinats (*el*, *la*, *l'*, *els*, *les*) ni indeterminats (*un*, *una*, *uns*, *unes*). Què diríeu d'aquest sistema?

    (a) Que és especialment adequat per a l'assimilació, però no tant per a la disseminació en vista que els articles són més del 10% del text.

(b) Que és especialment adequat per a la disseminació, perquè els articles són paraules molt poc freqüents en el text i per tant no serà necessària molta postedició.

(c) Que no és útil ni per a l'assimilació ni per a la disseminació.

26. Es pot posteditar sense mirar el text original?

(a) Sí.

(b) En general, no. Qui postedita produeix una traducció. Per tant, ha d'estar segur que el resultat és traducció del text original.

(c) Si el text és tècnic, es pot fer sense mirar. En altre cas, cal sempre mirar mai el text original.

27. L'ús d'un llenguatge controlat fa que ...

(a) ...l'escriptura siga més ràpida.

(b) ...l'estil del document resultant siga més homogeni.

(c) ...el poder d'expressió de l'idioma siga més gran.

28. Quan posteditem un text trobem una paraula que el traductor automàtic no ha sabut traduir, i ens l'ofereix en la llengua origen. No obstant això, aquest error no ha afectat a la traducció de la resta de l'oració. Què hem de fer?

(a) Preeditar el text original complet substituint la paraula per un sinònim que sí que reconega el traductor automàtic i tornar a traduir tot el text.

(b) Provar a traduir tot el text amb un altre traductor automàtic.

(c) Corregir-la i seguir posteditant.

29. Si en una fàbrica de frigorífics s'usen sistemes de traducció automàtica per a traduir a moltes altres llengües els manuals dels nombrosos models que s'hi fabriquen (i que són molt similars entre ells), la solució més eficient per a evitar errors de traducció és...

(a) ...regular la manera en què els autors escriuen els manuals.

(b) ...posteditar totes les traduccions.

(c) ...preeditar els manuals abans de traduir-los.

30. A l'hora de preeditar un text per a traduir-lo automàticament convé ...

(a) ...usar frases curtes.

(b) ...usar la forma passiva.

(c) ... usar oracions subordinades.

31. La postedició de la traducció realitzada per un traductor automàtic és sempre necessària ...

    (a) ... per a usar-la amb finalitats de disseminació.

    (b) ... per a usar-la amb finalitats d'assimilació.

    (c) ... quan s'ha realitzat també preedició.

6.7 Solucions

1. (*) Per exemple, si L és l'espanyol, 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 espanyol:

   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", "inscripción" 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)

9. **Solució 1 (n mots):** El traductor professional tradueix un text de n mots per 0,05$\times n$ euros. El sistema de traducció automàtica el tradueix per 0,03$\times n$ i corregir-lo costa 0,10$\times(10/100)\times n$, és a dir,

0,01$\times n$ euros perquè només 10 de cada 100 mots són incorrectes. Per tant, el sistema semiautomàtic costa només $(0{,}03{+}0{,}01) \times n$ =0,04$\times n$ euros, davant dels 0,05$\times n$ euros del traductor professional.

Solució 2 (1000 mots): El traductor professional tradueix un text de 1000 mots per 0,05$\times$1000=50 euros. El sistema de traducció automàtica el tradueix per 0,03$\times$1000=30 euros, i corregir-lo costa 10 euros, perquè en 1000 paraules hi ha 1000$\times$10/100=100 mots incorrectes i corregir cada un costa 0,10 euros: 0,10$\times$100=10. Per tant, el sistema semiautomàtic costa només 40 euros, davant dels 50 del traductor professional.

10. (c)
11. (a)
12. (b)
13. (a)
14. (a)
15. (b)
16. (b)
17. (b)
18. (b)
19. (a)
20. (b)
21. (c)
22. (b)
23. (a)
24. (c)
25. (a)
26. (b)
27. (b)
28. (c)
29. (a)

30. (a)

31. (a)

Capítol 7

Per què és difícil la traducció automàtica? Ambigüitat

7.1 Els quatre problemes de la traducció automàtica

Per què és difícil per a un sistema informàtic traduir com un professional? Una classificació interessant dels problemes de la traducció automàtica la dóna Arnold (2003). Segons aquest autor, la traducció automàtica té quatre grans problemes:

1. *La forma no determina completament el contingut* (és a dir, la *interpretació*): no sempre és fàcil determinar la interpretació que es volia que tinguera el que s'ha escrit. Aquest és el *problema de l'anàlisi*, també anomenat *ambigüitat*. Exemples: *Portaven notícies de Grècia* (tema o procedència?), *Ha venut les taronges que ha comprat a Joan* (Joan ven taronges o les compra?), *Treballa en l'estudi que li han encarregat* (prepara un document o està dissenyant un espai de treball?), etc.

2. *El contingut no determina completament la forma*. És a dir, és difícil determinar com s'ha d'expressar una interpretació concreta perquè hi ha més d'una manera de dir el mateix en qualsevol llengua. Aquest és el *problema de la síntesi*. Exemples: com es diu en quin moment del dia ens trobem? Cada idioma ho fa diferentment: català: *Quina hora és?*; portugués: *Que horas são?* (*Quines hores són?*); alemany: *Wie spät ist es?* (*Com és de tard?*); alemany: *Wieviel Uhr ist es?* (*Quantes del rellotge són?*), etc.[1]

3. *Les llengües divergeixen.* És a dir, hi ha diferències irreductibles en la manera que el mateix contingut s'expressa en llengües diferents. Aquest és el *problema de la transferència*, perquè es manifesta típicament en els sistemes de traducció automàtica per transferència (vegeu 8.3).

[1] Vegeu l'exemple *m'agrada nadar* en la p. 167

Per exemple, l'ordre estàndard de les oracions en català és *subjecte–verb–objecte*, mentres que en basc o en turc és *subjecte–objecte–verb*, en irlandés és *verb–subjecte–objecte* i en malgaix és *verb–objecte–subjecte*. O per exemple, els idiomes difereixen en la manera en la qual expressen les relacions entre dos noms: on en català diem *president* ***de*** *Kazakhstan*, el rus diu *prezident Kazakhstan****a***, el basc diu *Kazakstan****go*** *presidente*, o el Kazakh diu *Qazaqstan prezident****i***.

4. Construir un sistema de traducció automàtica comporta la gestió d'una gran quantitat de coneixement, que s'ha d'aplegar, descriure, i representar en una forma útil per al processament per ordinador. Aquest és el *problema de la descripció*.

D'aquests quatre, dedicarem la resta del capítol a descriure amb més detall el més important per a la traducció automàtica (i en general, per a qualsevol programa que haja de processar textos en llenguatge natural): l'ambigüitat inherent al llenguatge humà.

7.2 Ambigüitat

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).[2] Per tant, un enunciat ambigu pot tenir més d'una traducció a un altre idioma, tot i que de vegades pot tenir només una traducció a un altre idioma perquè aquesta traducció conserva l'ambigüitat de la frase original;[3] d'això, se'n sol dir *free ride* ("passi gratuït") i és tant més freqüent com més propers siguen les llengües involucrades en la traducció. 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. 2009, cap. 23):

> *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

(7.1) *El pare escura plats*

és diferent de la de l'oració

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

[3] Per exemple, l'oració espanyola *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?

(7.2) *La mare llegeix llibres*

Les oracions (7.1) i (7.2) tenen la mateixa sintaxi però diferent interpretació perquè contenen mots diferents amb interpretacions diferents. També explica per què l'oració

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

no té la mateixa interpretació que la frase

(7.4) *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

(7.5) **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:

(7.6) *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

(7.7) *Té molts amics*

té un subjecte buit (també anomenat el·líptic). En aquests casos podem considerar que les categories buides són mots "de zero lletres" 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, i l'ambigüitat deguda a les dues causes alhora.

7.2.1 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*,[4] 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, quan s'escriuen de la mateixa manera, se sol dir que són *homògrafes*; a més, s'anomena simplement *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. Podem diferenciar tres tipus d'ambigüitat per homografia:

1. *Ambigüitat entre categories lèxiques diferents*. Per exemple, el mot *moc* té dues formes lèxiques possibles, cadascuna amb una categoria lèxica diferent: *moc*, substantiu, masculí, singular; *moure*, verb, 1a. persona del singular, present d'indicatiu.

2. *Ambigüitat dins de la mateixa categoria lèxica sense canvi de lema*. Per exemple, el mot *canta* té dues formes lèxiques amb el mateix lema, la mateixa categoria lèxica, però distinta informació de flexió: *cantar* verb, 3a. persona del singular, present d'indicatiu; *cantar*, verb, 2a. persona del singular, imperatiu.

3. *Ambigüitat dins de la mateixa categoria lèxica amb canvi de lema*. Per exemple, el mot *poden* té, entre d'altres, dues formes lèxiques amb la mateixa categoria lèxica, la mateixa informació de flexió, però distint lema: *poder*, verb, 3a. persona del plural, present d'indicatiu; *podar* verb, 3a. persona del plural, present d'indicatiu.

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* i aquest tipus d'ambigüitat, *polisèmia*. Per exemple, el

[4] Vegeu la nota al peu de la pàg. 107

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. La polisèmia afecta totes les formes flexionades d'un determinat mot de la mateixa manera (*estacions* té exactament la mateixa ambigüitat que *estació*), ja que és una propietat del *lema*.[5]

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

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

   (7.8) *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 pot causar en molts casos 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 espanyol *destino* es pot traduir al català com a *destí* (sort futura) o *destinació* (punt d'arribada).

   Un altre exemple: l'oració

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

   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 (espanyol *mandar a freír espárragos*) però podria ser també la literal en un taller de terrisseria.

[5] Hi ha casos que no són tan senzills. Per exemple, en anglés, el mot *case*, un substantiu singular, pot referir-se a un tipus de contenidor (*a case of wine*) o a un exemple o situació particular (*It does not apply in this case*). Cada un dels mots ve d'un mot llatí diferent: el primer del mot femení *capsa*, i el segon de *casus*, el participi de *cado* 'caure'. Els diccionaris anglesos, que solen agrupar els mots polisèmics en una entrada, típicament en fan dues entrades diferents, i, de fet, en lexicografia no és estrany referir-se a *case* com un homògraf.

2. L'oració conté 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 espanyol *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 espanyol *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 espanyol *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 possessiu, etc., la qual pot tenir, en principi, més d'una possible interpretació, però aquesta interpretació està determinada per la relació de *coreferència* 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 o *dixi*: quan la interpretació és per *coreferència* amb un antecedent que apareix anteriorment en el text s'anomena *anàfora*, i *catàfora* si l'antecedent és posterior. En la frase

   (7.10) *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, espanyol 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

   (7.11) *Anna va apunyalar Marta. Joan va veure com queia redolant*

   qui va caure redolant, Anna 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,

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

i al qual s'assigna interpretació mitjançant processos díctics[6] 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.

7.2.2 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

   (7.13) *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. 7.1). L'ambigüitat estructural associada a les conjuncions coordinatives se sol anomenar *d'origen coordinatiu*.

2. *Ambigüitat estructural d'adjunció* (en anglés *attachment ambiguity*): es tracta d'un cas típic d'ambigüitat estructural que es manifesta quan hi ha un *adjunt*[7] (típicament un sintagma preposicional) que es pot inserir de diverses maneres en l'arbre d'anàlisi sintàctica de la frase. Per exemple, la frase

   (7.14) *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 7.2). Més exemples:

[6] Relacionats amb la *dixi*.

[7] Un *adjunt* és un sintagma o constituent que, conjuntament amb un altre sintagma o constituent, forma un constituent del mateix tipus que aquest últim (per exemple, un sintagma nominal més un sintagma preposicional formen un sintagma nominal); en un cert sentit, l'adjunt no és necessari sinó opcional.

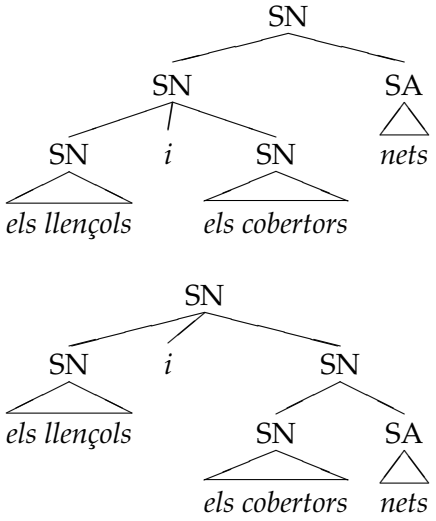


Figura 7.1: Dos arbres per a la frase "Posa els llençols i els cobertors nets a l'armari" (SN = sintagma nominal, SA = sintagma adjectival).

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

(7.16) *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.

Per saber més sobre ambigüitat estructural

Altres tipus d'ambigüitat estructural:

1. *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. (2009, p. 399):

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

   L'oració té dues interpretacions:

   (7.18)

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

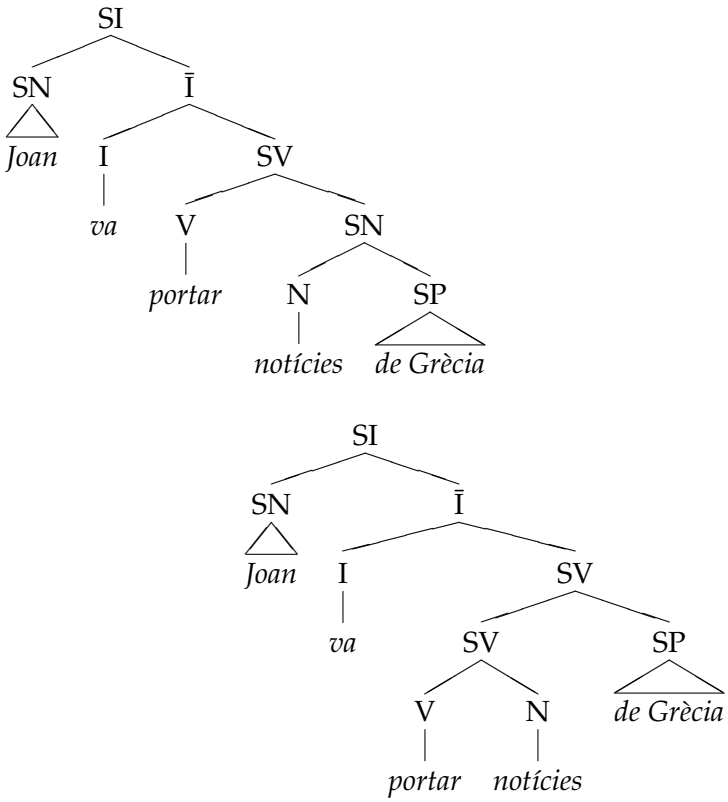


Figura 7.2: Dos arbres per a la frase "Joan va portar notícies de Grècia" (SI = sintagma inflexional, Ī = projecció intermèdia de la inflexió, I = inflexió, SV = sintagma verbal, V = verb, N = nom, SP = sintagma preposicional).

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

En aquests dos casos, l'ambigüitat és 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. 7.3).

2. *Ambigüitat estructural per moviment de Qu*. De vegades, l'anàlisi de la sintaxi d'una oració es complica per la presència de fenòmens de moviment de constituents. Considerem l'oració

(7.19) Qui diu que vindrà?

Aquesta oració té, fonamentalment, dues interpretacions. Una és

(7.20) Qui diu que PRO vindrà?

i l'altra

(7.21) * 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*) al principi de l'oració, el qual obligatori en molts idiomes —no en tots: el xinés o el turc no ho fan, per exemple— per als mots amb funció interrogativa. En aquest cas, com en l'exemple 7.17, 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

(7.22) *Com dius que Jordi 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:

(7.23)

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

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

(c) **Dius que Jordi 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*, és a dir, cap a l'arrel de l'arbre d'anàlisi sintàctica, des de les subordinades substantives completives amb verbs del tipus de *dir*, *explicar*, etc. En l'oració

(7.24) No m'agrada la manera com vas dir que vivia encara.

el primer *com* és un relatiu que pot modificar *dir* en l'oració *vas dir que vivia encara* ("no m'agrada la manera de dir-ho") o pot modificar *vivia* però ha estat mogut fora de la subordinada completiva *vivia encara*, que modifica *la manera* ("no m'agrada la manera com vivia, segons que vas dir").

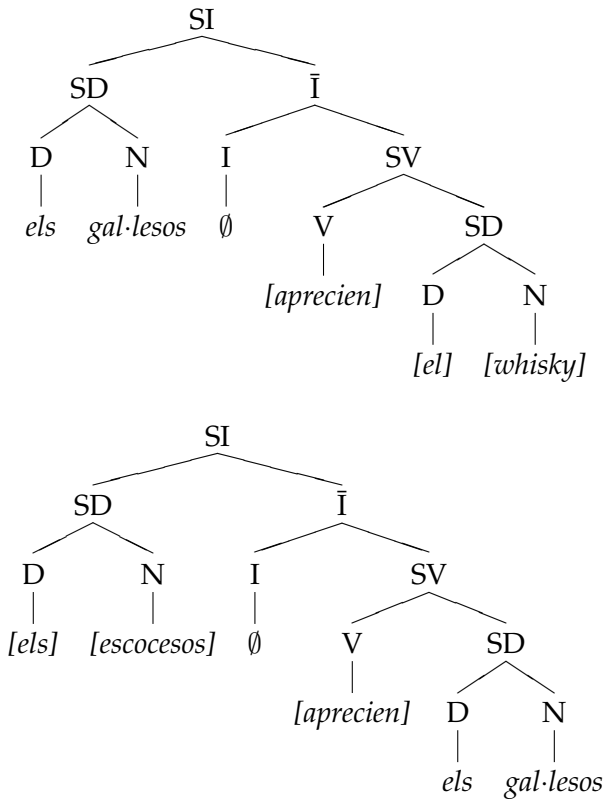


Figura 7.3: Dos arbres per a la segona part ("els gal·lesos") de la comparació "Els escocesos aprecien més el whisky que els gal·lesos" (SD =sintagma determinant, D = determinant).

[a] 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.

7.2.3 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 amb canvi de categoria (vegeu la pàg. 120). Per exemple, el mot *deu* pot voler dir "nou més un" (numeral) o "ha de donar o pagar" (verb). O el mot *cap*

que 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

(7.25) *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 (*cronometrar*) i substantiu (*temps*), *flies* pot ser verb (*vola*) i substantiu (*mosques*) i *like* pot ser verb (*agradar*) i preposició (*com*). 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 general en les llengües romàniques— 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

(7.26) *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

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

   té dues interpretacions:

   (7.28)

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

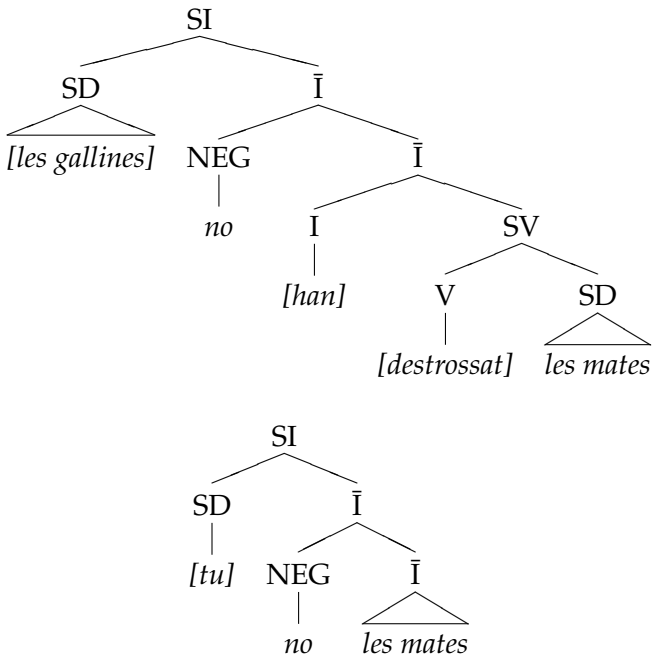


Figura 7.4: Dos arbres per a la segona oració coordinada en "Les gallines han destrossat el sembrat però no les mates" (NEG = negació). Els triangles s'usen per a no haver d'indicar tots els detalls d'un subarbre concret.

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

En (7.28a) *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 (7.28b) *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. 7.4).[8]

Per saber més sobre ambigüitats més complexes

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*.

[8] Dir que l'anàfora és només un procés lèxic és una simplificació. Hi ha involucrats aspectes sintàctics. Per exemple, en l'oració *Maria va parlar amb ella*, el pronom *ella* no es pot mai referir a *Maria*, però en l'oració *Maria va parlar amb una amiga d'ella*, sí que pot, i això és degut al fet que en l'estructura sintàctica de la primera oració hi ha una *barrera* a la coreferència que en la segona no existeix.

Els quantificadors són mots com *algun*, *tot*, *cada*. Quan l'abast d'un quantificador (és a dir, els mots que afecta) és imprecís, una oració pot tenir més d'una interpretació. Considerem l'exemple de Hutchins i Somers (1992)

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

Aquest exemple pot tenir dues interpretacions

(7.30)

(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. 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". La interpretació de les oracions se sol explicar de vegades en termes de *formes lògiques* (Radford et al. 2009, cap. 23); en el cas de les oracions amb quantificadors, aquestes formes lògiques contenen d'una banda, *variables* que poden referir-se a un rang d'objectes que cal considerar i, d'altra, operacions sobre aquestes variables. Doncs bé, en aquests casos, es pot assignar més d'una forma lògica a una oració.

7.2.4 Estratègies de resolució de l'ambigüitat

En general, els humans usem els nostres coneixements, les nostres expectatives i 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 els coneixements, les creences i les expectatives 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.

Les estratègies de resolució de l'ambigüitat solen basar-se en *restriccions* i *preferències*. Com veurem, les restriccions són normalment de naturalesa lingüística —per tant, requereixen un cert nivell d'anàlisi del text— i permeten descartar certes interpretacions, però no eliminen completament l'ambigüitat. Per a acabar de resoldre l'ambigüitat, s'usen preferències: s'assigna algun tipus de puntuació o valor a cada interpretació per elegir la millor. Les preferències se solen basar freqüentment en mètodes estadístics, basats en observacions obtingudes en grans quantitats de text.

Resolució de l'ambigüitat lèxica categorial

La resolució de l'ambigüitat lèxica categorial dels mots homògrafs, altrament coneguda com etiquetatge (dels mots) amb parts de l'oració (en anglés, *part-of-speech (PoS) tagging*) està molt ben estudiada. L'ambigüitat es redueix normalment usant restriccions basades en el coneixement lingüístic, i, com que normalment això no sol ser suficient, s'hi estableixen preferències basades en l'estudi estadístic de la freqüència d'aparició conjunta en els textos de determinades seqüències curtes de categories lèxiques.

En alguns casos, les restriccions poden ser suficients. Per exemple, el mot espanyol *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 determinant–verb personal no és permesa.

Però de vegades, les restriccions només redueixen l'ambigüitat sense eliminar-la completament. Per exemple, la paraula *sobre* pot ser un substantiu masculí singular, una preposició, i tres formes del verb *sobrar* (present de subjuntiu, 1a. i 3a. persona del singular, i un imperatiu cortés de 3a. persona del singular). En el context *Porta'm aquell sobre de la caixa*, una vegada feta l'anàlisi morfològica dels mots de l'oració, es podrien aplicar restriccions lingüístiques basades en seqüències de dues categories lèxiques per a reduir l'ambigüitat. Per exemple, una preposició no pot anar seguida d'una altra preposició. Com que *sobre* va seguit de *de*, que només pot ser preposició, podem descartar que *sobre* siga una preposició. Però no es poden descartar la resta de formes lèxiques: si *aquell* és un determinant, *sobre* pot ser un nom (com és el cas); si *aquell* és un pronom, *sobre* podria ser un verb (com en les oracions *Ara ens han sobrat dos cotxes, però pot ser que aquell sobre també més endavant.*).

Per tant, cal considerar l'ús de preferències. Per exemple, podem usar l'aproximació estadística. Si prenem un corpus (conjunt) suficientment gran de textos on un expert ha indicat la categoria lèxica de cada mot i comptem quantes voltes apareixen totes les seqüències possibles de dues 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.

Per saber més sobre estratègies de resolució de l'ambigüitat

Resolució de la polisèmia. La resolució de la polisèmia (en anglés *word sense disambiguation*) consisteix a assignar a un mot polisèmic, en un text o discurs, una interpretació concreta, possiblement 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 7.2.1).

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)). És possible aplicar restriccions per resoldre la polisèmia però s'han de basar en una anàlisi de naturalesa bastant profunda. Per exemple, podem decidir que el mot espanyol *gato* és un animal i no una ferramenta per alçar vehicles en la frase *El gato me miró desde debajo del coche*, perquè *gato* és el subjecte de *miró*, i el verb *mirar* requereix un subjecte animat, però com es pot veure, això requereix que s'hi haja fet una anàlisi sintàctica i semàntica.

Per això, en general s'usen mètodes basats en preferències. Heus ací dos exemples:

- L'ús de *xarxes semàntiques* on els conceptes se situen en els nodes (nusos) de la xarxa i s'agrupen jeràrquicament en superconceptes cada vegada més generals (per exemple, els conceptes *poma*, *pera*, *taronja* s'agruparien sota el concepte *fruita*): un exemple de xarxes semàntiques és *Wordnet*, `http://wordnet.princeton.edu`, que s'està generalitzant a altres llengües d'Europa (`http://www.illc.uva.nl/EuroWordNet/`). Quan tenim un mot polisèmic, li podem associar més d'un concepte o *sentit*. Per a elegir-ne un, podem, per exemple, prendre tots els possibles sentits del mot ambigu i assignar-los el sentit associat al concepte que està més prop dels conceptes representats pels mots veïns en el text. 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 espanyol–català l'aparició de *destino* prop d'*incierto* en espanyol 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 espanyol coincideix amb l'aparició *destinació* en català, podem elegir el sentit de "punt d'arribada". Aquesta informació podria servir per a traduir després de l'espanyol a l'anglés i elegir *destiny* o *destination* en cada cas amb molta probabilitat d'èxit.

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 també en restriccions i preferències.

Les *restriccions* es poden basar en informació morfològica, sintàctica, o fins i tot semàntica; tot depén del nivell d'anàlisi que estiga disponible:

- Un pronom masculí no pot tenir un antecedent femení (restricció morfològica): *Maria* no pot ser l'antecedent de *ell* en l'oració *Maria es va passar tot el dia parlant d'ell*.
- La informació sintàctica pot ser més rellevant que no ho sembla: si diem

  (7.31) *Marta la va veure*

  l'antecedent de *la* no pot ser *Marta*, per causa de les anomenades *barreres*, restriccions associades a determinades característiques de l'estructura sintàctica de l'oració. En canvi, si diem

(7.32) *Marta va parlar amb qui la va veure*

no es pot descartar completament que l'antecedent de *la* siga *Marta*.

- Hi ha vegades que només podem recórrer a una anàlisi semàntica; en l'exemple (ja discutit en la secció 6.3)

  (7.33) *[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*).

Les restriccions no solen ser suficients, i sol ser necessari l'establiment de *preferències*. Per exemple, es poden preferir

- els antecedents més recents,
- els antecedents que fan de subjecte als que fan d'objecte, o
- els antecedents que han estat introduïts explícitament com l'assumpte del discurs o de la conversa: *Doncs, pel que fa a Joan...*.

Açò se sol instrumentar a través d'un sistema que assigna *puntuacions* per cada una de les característiques: se sumen les puntuacions per a tots els antecedents possibles i s'elegeix el que obté la puntuació més alta (Lappin i Leass 1994).

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), quines són *acceptables* i, entre les acceptables, quina és la més versemblant i per tant preferida 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, l'oració 7.16) hi ha massa estructures a considerar. De fet, hi ha experiments psicolingüístics que indiquen que de vegades usem estratègies fonamentalment 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 (presumiblement 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 l'anglés *lead up the garden path*) 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, en anglés *eyetracking*) 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 del final d'aquesta secció:

(7.34) *Joan besà Maria i la seua germana...*[a]

(7.35) *Com que Joan sempre corre un parell de quilòmetres...*[b]

(7.36) *En el otro accidente murieron sesenta y cinco...*[c]

(7.37) *The horse raced by the barn...*[d]

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

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

podríem considerar la interpretació que s'hi parla de la durada de l'afaitat (el que *va aprendre* és *afaitar-se en dos minuts*) com a més probable que la que interpreta que s'hi parla de la durada de l'aprenentatge (*aprendre a afaitar-se* li va costar *dos minuts*), ja que en el segon cas potser hauria estat més natural dir

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

La regla que afavoreix que els adjunts s'associen a l'últim sintagma que els admeta —i que permet, per tant, anar construint l'arbre d'anàlisi sintàctica gradualment sense haver de fer-hi grans reorganitzacions— se sol anomenar regla de *clausura tardana* —en anglés *late closure*—; per exemple, aquesta regla afavoreix el primer dels arbres de la figura 7.2. Una altra regla que se sol usar és la d'*adjunció mínima* —en anglés *minimal attachment*— 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 (vegeu la secció 8.3), ja que no s'hi fa cap processament semàntic.

El punt de vista purament sintàctic és pot considerar una simplificació excessiva; moltes vegades, les persones resolem l'ambigüitat estructural usant restriccions semàntiques o fins i tot lèxico-semàntiques:

- 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 per causa de l'ambigüitat estructural, però si diem la frase estructuralment idèntica —i per tant idènticament ambigua— "Va presentar l'home que menjava taronges a Joan" no hi ha més que una interpretació possible.
- Considereu aquestes dues frases estructuralment idèntiques afectades per la mateixa ambigüitat estructural pura d'adjunció:

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

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

  En l'oració 7.40, 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 7.41 no considerem la primera interpretació (encara que siga la preferida sintàcticament segons la regla de clausura tardana), 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 que resol l'ambigüitat és capaç d'usar informació semàntica, podria elegir correctament en aquest cas.

---

[a] ... el va recriminar per haver-ho fet.
[b] ... li semblen poc.
[c] ... resultaron heridos.
[d] ... fell down.

7.3 Qüestions i exercicis

Per a poder respondre a les preguntes marcades amb (*) cal que us llegiu els quadres *Per saber més*.

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?)
   (f) *Us han dit que vaja?* (Qui ha d'anar?).
   (g) *Vale más que las comas* (1: "...que els signes de puntuació"; 2:"...que les menges")
   (h) *El mecánico revisó la suspensión del auto de Garzón* (1: "Aquest mecànic és un expert en legislació i s'ha llegit la resolució judicial sencera"; 2: "Els amortidors del cotxe de Garzón ja necessitaven una revisió").
   (i) *A pesar de haber sido soldado, salió despedido del avión* (1: "El sistema fotogràfic estava fortament fixat al fuselatge però es va soltar de l'aparell quan l'avió va girar en ple vol"; 2: "Malgrat el seu passat militar gloriós, el president el va destituir abans d'arribar a l'aeroport de destinació").
   (j) *Els lladres van ser atrapats a una fàbrica incendiada per un policia* (1: "Els lladres van ser capturats pel comissari a una fàbrica abandonada"; 2: "La fàbrica on van ser capturats va ser l'objectiu d'un agent piròman").
   (k) *Coto privado de caza* (1: "Aquesta àrea de caça no és pública"; 2: "Aquesta és una àrea sense caça").
   (l) *Vull ballar i cantar cançons de bressol* (1: "Ballarem durant una estona i després et cantaré perquè et dormes"; 2: "M'estime tant ballar cançons de bressol com cantar-ne").

(m) *El manifestant es refà de la pallissa que li van donar en l'hospital* (1: "La manifestació va ser una mica violenta i algunes persones han hagut de ser traslladades a l'hospital"; 2: "Quina pallissa va donar-li la infermera al quiròfan!").

(n) *Serviran polp a la gallega* (1: "Serviran polp a una senyora de Galícia"; 2: "Serviran polp preparat a l'estil gallec").

(o) *No puc veure bé la foto que m'has enviat per correu electrònic perquè no puc tancar totes les finestres* (1: "Encara hi entra sol i es reflecteix en la pantalla"; 2: "Tinc l'escriptori ple de documents oberts").

(p) *–Hem rebut notícies que diuen que, per causa de la humitat i la calor en l'interior del temple, els bancs i els altars de fusta han rebrotat i els han crescut branques i fulles. –I les creus?* (1: "Creus aquestes notícies?" 2: "Els crucifixos també han rebrotat?").

(q) *S'han de repassar les entrades i les despeses que s'hagen fet en euros* (1: "Les entrades s'han de repassar totes; les despeses només si s'han fet en euros"; 2: "De les entrades, se n'han de repassar només les fetes en euros").

(r) *Després que la venedora acabà la descripció dels avantatges de la urbanització projectada, el comentari unànime dels inversors va ser que la trobaven molt interessant*. (1: "Els inversors, la veritat, prestaven més atenció a la venedora que al producte"; 2: "Els va agradar l'estil de la descripció"; 3: "La venedora no parlava clar, la descripció era incompleta, però a pesar de tot, la urbanització era una inversió prometedora").

(s) *A la trapezista, últimament, no li eixien els números* (1: "Sempre acabava caent a la xarxa"; 2: "tenia més despeses que ingressos").

2. (*) Hi ha ambigüitats de tipus lèxic que poden ser sempre correctament resoltes després de fer una anàlisi morfològica. No obstant això, hi ha d'altres que només poden ser tractades si es fa una anàlisi sintàctica (tot i que l'ambigüitat siga de tipus lèxic) i fins i tot n'hi ha que requeririen una anàlisi semàntica per a resoldre-les.

   Elegiu una llengua origen i una llengua meta (francés, anglés, alemany, català o espanyol) i poseu un exemple d'oració per a cada un dels tres casos anteriors, on siga necessari un determinat nivell d'anàlisi per tal de resoldre una ambigüitat i produir-ne la traducció correcta. Expliqueu quina informació usa el sistema en cada cas per a prendre una decisió.

3. (*) Indiqueu breument quines estratègies 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*

4. Si una oració té només una ambigüitat lèxica pura...
   (a) ...té un únic arbre d'anàlisi sintàctica, però 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.

5. La frase *M'agrada més que la bata* pot tenir dues interpretacions; en la primera es parla d'una peça 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.

6. 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.

7. 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.

8. 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.

9. (*) Considereu l'homògraf espanyol *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.

10. (*) Moltes vegades, l'ambigüitat lèxica no és ni polisèmia (*estació*, *bomba*), ni ambigüitat lèxica amb canvi de categoria gramatical (*sobre* [preposició, substantiu i, en la varietat valenciana, verb], *riu* [substantiu i verb]) sinó que succeeix perquè dues formes *de la mateixa categoria lèxica* són homògrafes: *volem* pot ser una forma del verb *volar* i del verb *voler*; *podeu* pot ser una forma del verb *poder* i del verb *podar*; en espanyol, *creo* és una forma de *creer* o de *crear*, *fui* és una forma de *ir* o de *ser*, etc. Per a resoldre l'ambigüitat d'un mot polisèmic s'ha d'usar informació semàntica; per a resoldre l'ambigüitat lèxica categorial sol ser suficient usar informació sintàctica (per exemple, la categoria gramatical del mots anterior i posterior); però, és possible resoldre l'ambigüitat deguda a l'homografia de mots de la mateixa categoria usant només la sintaxi o és necessari l'ús d'informació semàntica?

11. Els sistemes de traducció mot per mot poden cometre, per exemple, errors deguts a l'elecció incorrecta de la categoria gramatical d'un mot lèxicament ambigu. Elegiu dues llengües, L_1 i L_2 i poseu dos exemples de traduccions errònies de L_1 a L_2, indicant la frase original, la frase mal traduïda i la frase correcta.

12. Si una forma superficial és ambigua però té només una forma lèxica...

   (a) ... es tracta d'un mot homògraf.
   (b) ... hi ha algun error en l'anàlisi morfològica.
   (c) ... podem dir que el lema és polisèmic.

13. Pot una oració tenir més d'una traducció a un altre idioma malgrat estar formada completament per mots que no són ni homògrafs ni polisèmics en la llengua original?

(a) No: ho prohibeix el principi de composicionalitat semàntica.

(b) Sí, encara que no continga pronoms o altres expressions anafòriques susceptibles de tenir més d'un antecedent possible.

(c) Sí, però només si conté pronoms o altres expressions anafòriques susceptibles de tenir més d'un antecedent possible.

14. (*) En absència d'informació lèxica o semàntica, l'ambigüitat estructural...

(a) ... és impossible de resoldre.

(b) ... es pot resoldre usant regles derivades d'un estudi de les preferències sintàctiques observades en experiments psicolingüístics.

(c) ... no pot afectar mai el resultat de la traducció automàtica.

15. (*) És possible resoldre en alguns casos l'ambigüitat deguda a un mot homògraf utilitzant exclusivament informació morfològica?

(a) No, aquest tipus d'ambigüitat exigeix un tractament semàntic com a mínim.

(b) No, sempre cal utilitzar informació de caràcter sintàctic per resoldre-la.

(c) Sí, usant la informació morfològica dels mots adjacents.

16. Si un mot té només una forma lèxica i una única traducció a una determinada llengua, pot ser encara ambigu?

(a) No.

(b) Sí, pot ser polisèmic encara que la traducció a aquesta llengua de totes les interpretacions siga la mateixa.

(c) Sí; pot ser homògraf i tractar-se d'un passi gratuït.

17. Si traduïm automàticament la frase espanyola *Ayer cantamos las mismas canciones* i obtenim en anglés *Yesterday we sing the same songs* o en francés *Hier nous chantons les mêmes chansons*, quin tipus d'ambigüitat ha estat mal resolta?

(a) Una ambigüitat lèxica per homografia d'un mot.

(b) Una ambigüitat lèxica pura per polisèmia d'un mot.

(c) Una anàfora.

18. Si traduim automàticament la frase catalana *Hilari no coneix bé la Mariona: cada dia troba sorprenent el que fa* i obtenim en anglés *Hilari does not know Mariona well: every day he finds what he does astonishing* o en francés *Hilari ne connaît pas bien Mariona: tous les jours elle trouve ce qu'il fait etonnant*, quin tipus d'ambigüitat ha estat mal resolta?

(a) L'anàfora d'un pronom buit.

(b) L'anàfora del pronom *que*.

(c) Una ambigüitat sintàctica de l'oració subordinada "el que fa".

19. El principi de composicionalitat diu que la interpretació d'una oració està determinada per les interpretacions dels mots i per la sintaxi. Quan es produeix una ambigüitat perquè no queda clara l'adscripció d'un sintagma preposicional, com en *porta la clau de l'armari gran*, quina n'és la raó?

    (a) L'existència de més d'una estructura sintàctica possible.

    (b) L'ambigüitat categorial de la preposició.

    (c) La polisèmia de la preposició.

20. L'ambigüitat lèxica categorial d'un mot...

    (a) ...no es pot resoldre mai si no s'usa informació semàntica sobre el text o sobre els mots contigus.

    (b) ...no es pot resoldre si no es fa l'anàlisi sintàctica completa de la frase, ja aquesta és l'única manera d'elegir l'anàlisi morfològica correcta.

    (c) ... s'intenta resoldre normalment amb regles basades en les categories lèxiques dels mots que l'acompanyen en la frase.

21. Quin tipus d'ambigüitat es produeix en el pronom feble *li* de la frase *Vaig veure Mario i la seua mare; li vaig dir que m'agradava molt el seu fill*?

    (a) Homografia, perquè el pronom pot, en principi, estar substituint un nom o un altre.

    (b) Una anàfora.

    (c) Polisèmia.

22. Una d'aquestes tres no és un tipus d'ambigüitat lèxica:

    (a) L'homografia.

    (b) L'ambigüitat d'adjunció.

    (c) La polisèmia.

23. De quina classe és l'ambigüitat que presenta l'oració *Expulsaran el portaveu del partit*?

    (a) Lèxica, deguda al fet que el mot *expulsar* és polisèmic.

    (b) Estructural d'origen coordinatiu: no sabem si el sintagma preposicional *del partit* modifica només al segon element *el portaveu* o a tot el sintagma *Expulsaran el portaveu*.

(c) Estructural d'adjunció: el sintagma preposicional *del partit* pot ser un adjunt del sintagma verbal *Expulsaran el portaveu* o del sintagma nominal *el portaveu*.

24. Només una d'aquestes tres afirmacions és certa. Quina?

    (a) Una oració pot ser ambigua sense que cap dels seus mots siga ambigu per ell mateix.

    (b) Una oració només pot ser ambigua si almenys un dels seus mots és ambigu.

    (c) El fet que una oració siga ambigua implica necessàriament que les traduccions de les diverses interpretacions a una altra llengua han de ser diferents.

25. Quin tipus d'ambigüitat es dóna en l'oració *La mata l'enveja* (1: "l'enveja l'està matant"; 2: "la planta li té enveja a ella")?

    (a) Ambigüitat lèxica per polisèmia.

    (b) Ambigüitat lèxica deguda a l'anàfora.

    (c) Ambigüitat estructural deguda a l'ambigüitat lèxica categorial.

26. Quan un adjectiu presenta la mateixa ambigüitat (per exemple, pot tenir més d'una traducció), independentment de com es trobe flexionat en gènere i nombre, direm que l'adjectiu és...

    (a) ... anafòric.

    (b) ... homògraf.

    (c) ... polisèmic.

27. Segons Arnold (2003) els problemes als quals s'enfronta la traducció automàtica són quatre. Indiqueu quina de les afirmacions següents és falsa:

    (a) El problema de l'anàlisi es refereix a la dificultat per resoldre l'ambigüitat d'un enunciat.

    (b) El problema de la síntesi es refereix a l'ambigüitat dels textos traduïts automàticament.

    (c) El problema de la descripció consisteix en el fet que és impracticable descriure de forma suficient i computacionalment eficient tot el coneixement necessari per traduir.

28. Un traductor automàtic per transferència morfològica avançada ...

    (a) ... resol la polisèmia mitjançant l'ús d'un analitzador morfològic.

(b) ...resol la polisèmia mitjançant l'ús d'un desambiguador lèxic categorial.

(c) ...no pot resoldre la polisèmia amb cap dels programes esmentats en les altres dues opcions.

29. Quin tipus d'ambigüitat es dóna en l'oració "*Aston Family Man era el baix de The Wailers*" (Aston era el més baix del grup; Aston tocava el baix en el grup)?

(a) Ambigüitat lèxica per polisèmia.

(b) Ambigüitat lèxica categorial dins de la mateixa categoria lèxica.

(c) Ambigüitat lèxica categorial entre categories lèxiques diferents.

30. Indiqueu quina de les afirmacions següents és falsa:

(a) Hi ha casos en els quals no fa falta resoldre l'ambigüitat per produir una traducció adequada en la llengua meta.

(b) L'ambigüitat sempre representa un problema a l'hora de traduir entre dues llengües.

(c) L'ambigüitat és un dels problemes als quals ha d'enfrontar-se un traductor automàtic.

7.4 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: pot modificar *alcalde* o *expulsaran [l'alcalde]*.

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

(c) Ambigüitat lèxica 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 sintagmes nominals coordinats (*galetes i pa*) o només al segon (*pa*).

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

(f) Ambigüitat per el·lipsi: el subjecte de *vaja* pot ser *jo*, *ell*, *ella*, etc. En el cas dels pronoms de tercera persona, la interpretació estarà determinada pels antecedents que se'ls assignen (ambigüitat lèxica per anàfora).

(g) 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.

(h) L'oració és ambigua perquè dos dels seus mots presenten ambigüitat lèxica pura (polisèmia): *auto* pot ser un automòbil o un tipus de resolució judicial; *suspensión* pot ser l'acció de suspendre (la resolució) o el sistema d'amortidors de l'automòbil. De les quatre combinacions possibles, dues tenen un cert sentit.

(i) L'oració és ambigua per l'ambigüitat lèxica (homografia) de *soldado*. En la primera interpretació és un participi en la forma passiva *haber sido soldado*; en la segona és un substantiu masculí singular. També hi intervé l'ambigüitat lèxica pura (polisèmia) de *despedir* (en la primera *llançar*; en la segona, *deixar sense treball*). Dues de les quatre combinacions tenen sentit.

(j) Ambigüitat estructural pura d'adjunció. El sintagma preposicional *per un policia* pot modificar el sintagma verbal *atrapats a una fàbrica incendiada* o només el sintagma verbal *incendiada*.

(k) Ambigüitat mixta. D'una banda, lèxica: el mot *privado* pot ser un adjectiu (interpretació 1) o un participi (interpretació 2). D'altra banda, estructural: en la primera interpretació, el sintagma preposicional *de caza* modifica el sintagma nominal *coto privado* (*[[coto privado] [de caza]]*); en el segon, només el participi *privado* (*[[coto] [[privado] [de caza]]]*).

(l) Ambigüitat estructural d'origen coordinatiu. El sintagma nominal *cançons de bressol* pot modificar només el segon sintagma verbal *cantar* o el sintagma verbal complet *ballar i cantar* (és a dir, *cançons de bressol* pot ser objecte directe només del segon verb o dels dos).

(m) Ambigüitat estructural pura d'adjunció. El sintagma preposicional *a l'hospital* pot modificar el sintagma verbal *li van donar* o el sintagma verbal *es refà de la pallisa que li van donar*.

(n) Ambigüitat estructural d'adjunció: el sintagma preposicional *a la gallega* pot modificar el nom *polp* per a formar sintagma nominal *polp a la gallega* o modificar el sintagma verbal *serviran polp* (amb nucli *serviran*) i formar el sintagma verbal *serviran polp a la gallega*.

(o) L'oració és ambigua per polisèmia (ambigüitat lèxica) del substantiu *finestra* (de la paret/ del sistema operatiu)

(p) Ambigüitat estructural d'origen categorial. En la primera interpretació *les* és un pronom i *creus* és un verb, i formen junts un sintagma verbal; en la segona *les* és un article i *creus* és un substantiu i formen junts un sintagma nominal.

(q) Ambigüitat estructural d'adjunció. El sintagma (oració subordinada de relatiu) *que s'hagen fet en euros* pot modificar al segon sintagma nominal *les despeses* o al sintagma nominal complet *les entrades i les despeses*.

(r) Ambigüitat de l'oració deguda a l'ambigüitat lèxica per anàfora. El pronom *la* pot tenir tres antecedents: *la venedora*, *la descripció* o *la urbanització projectada* i, per tant, tres interpretacions diferents.

(s) Ambigüitat de l'oració per polisèmia (ambigüitat lèxica) del mot *número* (part d'una actuació / comptes econòmics)

2. (*) Exemples de l'espanyol al català:

   - Ambigüitat que es pot resoldre després de fer una anàlisi morfològica de l'oració: en *mi trabajo*, l'homògraf *trabajo* pot ser substantiu (català *treballo*) o verb (català *treball*), però la presència de *mi* (determinant possessiu) desambigua l'homògraf perfectament.
   - Ambigüitat lèxica que necessita una anàlisi sintàctica per a ser resolta: l'expressió multimot *sesenta y cinco* pot ser un únic numeral (català *seixanta-cinc*) o dos numerals coordinats (català *seixanta i cinc*). L'anàlisi morfològica no és suficient per a detectar que en la frase *En el lugar donde murieron sesenta y cinco quedaron restos* és el primer cas i en la frase *Murieron sesenta y cinco quedaron malheridos* és el segon.
   - Ambigüitat lèxica que necessita una anàlisi semàntica per a resoldrela: el mot polisèmic *destino* en l'oració *El destino estaba escrito en el pasaje arrugado que encontraron* es traduiria pel català *destinació* i en canvi en l'oració *El destino estaba escrito en el libro sagrado que encontraron* es traduiria pel català *destí*; l'elecció exigeix identificar relacions semàntiques entre les interpretacions dels mots.

3. (*) Vegeu el quadre *Per saber més* de la secció 7.2.4. 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.

4. (b)

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

6. (c)

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

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

9. (b)

10. La solució semàntica és més potent i general però exigeix una anàlisi molt profunda de la frase. En alguns casos, la sintaxi podria donar pistes que permetrien una desambiguació molt aproximada. Per exemple, si *fui* va seguit de la preposició *a*, és molt probable que es tracte del verb *ir*; d'altra banda, si va seguit d'un participi passat, és molt probable que es tracte del verb *ser*: es podria fer una categoria gramatical especial per al verb ser i usar tècniques de desambiguació categorial. Si trobem *podem* (*volem*) seguit d'infinitiu, és molt més probable que es tracte del verb *poder* (*voler*) que del verb *podar* (*volar*); de nou, caldria usar una categoria gramatical especial, en aquest cas per als verbs modals.

11. Es poden trobar molts exemples; per exemple, entre $L_1 =$ espanyol i $L_2 =$ català, tenim:

    - Ayer por la mañana vino tarde $\rightarrow$ *Ahir pel demà vi vesprada (Ahir de matí va venir tard).
    - Río porque no llegó a la meta $\rightarrow$ *Riu perquè no va arribar a la fique (Ric perquè no va arribar a la meta).

12. (c)

13. (b). Pot tenir més d'un arbre d'anàlisi sintàctica (ambigüitat estructural).

14. (b)

15. (c)

16. (b)

17. (a). *Cantamos* pot ser present o passat.

18. (a). El pronom buit que fa de subjecte de *fa*.

19. (a)

20. (c)

21. (b). El pronom *li* pot tenir els antecedents *Mario* i *la seua mare*

22. (b)

23. (c)

24. (a)
25. (c)
26. (c)
27. (b)
28. (c)
29. (c)
30. (b)

Capítol 8

Tècniques de traducció automàtica

Aquest capítol descriu les tècniques o, dit d'una altra manera, les estratègies bàsiques usades pels programes de traducció automàtica.

Hi ha dos grans grups de sistemes de traducció automàtica:

- D'una banda, els sistemes de traducció automàtica **basats en regles** (en anglés, *rule-based machine translation*) o **basats en coneixement** (en anglés *knowledge-based machine translation*). En aquests sistemes, la informació necessària per a realitzar la traducció automàtica (diccionaris, regles) l'han escrita persones expertes de manera *deductiva*: és a dir, han pensat en com automatitzar el procés de traducció automàtica i n'han deduït la informació necessària per a realitzar-la. Entre aquests sistemes, podem distingir:
  - Els sistemes de traducció automàtica *indirecta per transferència* (apartat 8.3), entre els quals, podem distingir, d'acord amb el nivell d'abstracció lingüística:
    - els sistemes de transferència morfològica avançada (de vegades anomenats de "traducció directa", tot i que no ho són; apartat 8.3.1);
    - els sistemes de transferència sintàctica (apartat 8.3.3), i
    - els sistemes de transferència semàntica (apartat 8.3.5).
  - Els sistemes de traducció automàtica *per interlingua* (apartat 8.4).
- D'altra banda, els sistemes de traducció automàtica **basats en corpus** (en anglés *corpus-based machine translation*). En aquests sistemes (vegeu l'apartat 8.5), la informació necessària per a realitzar la traducció automàtica s'*aprén* automàticament de manera *inductiva* a partir d'un *corpus* paral·lel, és a dir, de grans quantitats de textos i les seues traduccions, prèviament *segmentats* i *alineats* per posar cada oració d'un

text en correspondència amb la seua traducció en l'altre.[1] Els sistemes basats en corpus més comuns són els de *traducció automàtica estadística*, on el que s'aprenen són models probabilístics de traducció. Durant el decenni de 2010 s'està investigant en sistemes que usen l'anomenat *aprenentatge profund* (en anglés *deep learning*) basat en *xarxes neurals artificials*, les quals es basen vagament en com funciona el cervell humà.

8.1 Funcionament de la traducció automàtica

Els sistemes de traducció automàtica *reals*, és a dir, els que s'usen en la realitat, són el resultat de fer una series d'aproximacions sobre la traducció automàtica *ideal* per fer el problema de la traducció computacionalment abordable.

La majoria dels sistemes de traducció automàtica, independentment de si són basats en regles o en corpus, adopten la que podrien anomenar **aproximació oracional**, segons la qual *traduir texts és traduir oracions*. Aquesta aproximació exclou el tractament d'alguns aspectes de l'estructura del discurs.

Una vegada feta aquesta aproximació general, la resta d'aproximacions depenen del tipus de sistema de traducció automàtica. La majoria dels sistemes de traducció automàtica basats en coneixement aborden la traducció com l'aplicació del *principi de composicionalitat semàntica* (PCS, capítol 7), el qual afirma que la interpretació (el significat) d'una oració es construeix composicionalment a partir de les interpretacions dels mots, seguint els agrupaments dictats pel seu arbre d'anàlisi sintàctica, i també al revés, que les oracions es poden construir composicionalment a partir de les interpretacions (Tellier 2000). Traduir una oració, en aquest esquema, comporta:

- fer-ne l'anàlisi sintàctica completa,
- assignar interpretació a cada mot,
- construir composicionalment una interpretació de l'oració,
- analitzar-la per a obtenir mots i un arbre d'anàlisi sintàctica per a la llengua meta (LM), i
- generar una oració en LM a partir dels mots i l'arbre.

Aquest és bàsicament el *modus operandi* dels sistemes d'*interlingua* (que es discutiran en la secció 8.4) i constitueix l'**aproximació composicional**. No oblidem que aquesta descripció assumeix que l'ambigüitat lèxica (múltiples

[1] Els textos del corpus d'aprenentatge poden a més ser anotats o processats amb algun tipus de processador lingüístic com ara un analitzador morfològic o sintàctic.

interpretacions dels mots) i estructural (més d'un arbre d'anàlisi sintàctica) ha estat idealment resolta.

D'una altra part, els sistemes de traducció automàtica indirecta per transferència (que es discutiran en l'apartat 8.3) són el resultat d'una sèrie d'aproximacions (moltes d'elles inevitables) sobre un model ideal i teòricament motivat basat en el *principi de composicionalitat semàntica*. Aquests sistemes també es poden veure com el resultat d'una sèrie de refinaments inevitables sobre un sistema de traducció *mot per mot* (vegeu l'apartat 8.2), és a dir, com una sèrie d'operacions addicionals que s'han de fer a més d'anar substituint cada mot per un equivalent constant. Per exemple, per a produir traduccions acceptables, ràpides i intel·ligibles, fins i tot entre llengües molt semblants, s'ha d'afegir un processament lèxic robust (per exemple, per a tractar expressions multimot o per a elegir equivalents adequats per a mots lèxicament ambigus) i un processament estructural local o global que es basa en regles simples i ben formulades per a algunes transformacions estructurals (reordenaments, concordança, etc.).

Com en el cas dels traductors professionals, els sistemes de traducció automàtica no sempre necessiten *comprendre* les frases en llengua origen (LO) , és a dir, construir-ne una interpretació explícita. Aquesta noció, que pot semblar polèmica, no ho és tant: qui tradueix com a professional un manual de mecànica de l'automòbil o un text de física teòrica ho pot fer sense haver d'entendre completament les disciplines corresponents. Els sistemes de *transferència* sintàctica (vegeu l'apartat 8.3.3) prenen una drecera i van directament de l'arbre d'anàlisi sintàctica i els mots en LO a l'arbre i els mots en LM. Ho fan aplicant transformacions a l'arbre d'anàlisi sintàctica (*transferència estructural*) i substituint els mots (*transferència lèxica*), sense construir una representació explícita de la interpretació; aquesta és l'**aproximació de transferència**. Depenent del tipus concret de sistema de traducció automàtica per transferència, encara són possibles més aproximacions, com veurem més avall.

Per últim, el sistemes de traducció automàtica estadística (que es discutiran en l'apartat 8.5) fan **assumpcions d'independència estadística** per poder modelar estadísticament el procés de traducció. Per exemple, el *model de traducció*, que s'usa per estimar la probabilitat de què un segment de text en LM siga la traducció d'un segment de text en LO, assumeix que la traducció d'un segment és independent de la traducció de la resta de segments de l'oració; és a dir, assumeix que no cal tenir en compte el context per a traduir cadascun dels segments de l'oració en LO.

8.2 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 tra-

ducció d'una frase es produeix directament, sense que es genere una representació intermèdia de la frase; de vegades també se sol anomenar vagament traducció *mot per mot*. L'estratègia *indirecta* produeix, a partir de la frase en 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.

Una formalització possible de la traducció automàtica directa més senzilla possible és la traducció *mot a mot*: el sistema llig el text original mot a mot d'esquerra a dreta,[2] substitueix cada mot original per un equivalent fix (d'un mot, de més mots, o fins i tot de zero mots) en LM sense tenir en compte el context i escriu els mots un a un i en el mateix ordre en el text meta. Per exemple, si la frase té N mots,

$$m_1\ m_2\ m_3\ \cdots\ m_N$$

la traducció *mot per mot* és

$$T(m_1)\ T(m_2)\ T(m_3)\ \cdots\ T(m_N)$$

on $T(m)$ representa l'equivalent fix del mot m en la LM, que pot tenir zero, un o més mots. Per exemple, la traducció *mot per mot* a l'anglés de la frase catalana *Aquest exercici pràctic és molt senzill*, amb $m_1 =$ *Aquest*, $m_2 =$ *exercici*, etc., podria ser *This exercise practical it is very simple* (incorrecta), on, per exemple, $T(m_4) = T($*és*$) =$ *it is*.

Cap sistema real de traducció automàtica usa aquest model tan rudimentari de traducció, ja que és incapaç de produir traduccions automàtiques útils, ni tan sols per a llengües molt similars: tots els sistemes van més enllà i realitzen operacions addicionals. Per això mateix, el model *mot a mot* es pot usar com a model de referència o *model zero* a l'hora d'estudiar què més fan els sistemes existents, o què més cal fer per a produir traduccions automàtiques útils per a un parell de llengües.

8.3 Traducció indirecta per transferència

Molts dels sistemes indirectes basats en regles són sistemes de *transferència*. Un *sistema de traducció automàtica indirecta per transferència*, o, abreviadament, *sistema de transferència* és el que fa les traduccions en tres fases ben diferenciades anomenades *anàlisi*, *transferència* i *generació*; cada una d'aquestes fases és realitzada per un mòdul (un subprograma) del sistema:

- El mòdul d'*anàlisi* és un mòdul monolingüe que produeix, a partir de la frase en LO, una *representació abstracta del text origen* (RATO). En

[2] S'assumeix que el text està ja segmentat en mots, operació que pot no ser trivial en alguns idiomes com ara el japonés o el xinès, que no usen espais en blanc per a separar els mots.

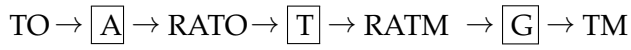


Figura 8.1: Fases d'anàlisi (A), transferència (T) i generació (G) en un sistema de traducció indirecta per transferència (TO = text origen; RATO = representació abstracta del text origen; RATM = representació abstracta del text meta; TM = text meta).

la RATO s'eliminen tots els detalls de la frase en LO que no es consideren 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 RATO.

- El mòdul de *transferència* és un mòdul bilingüe que llig la RATO i genera una altra representació abstracta similar, però per a la LM, la *representació abstracta del text meta* (RATM).

- El mòdul de *generació* (o, menys comunament, *síntesi*) genera a partir de la RATM un text *concret*: la traducció en brut.

Aquestes tres fases s'esquematitzen en la figura 8.1.

Les representacions abstractes *apropen* les dues llengües eliminant-ne alguns detalls específics i destacant característiques generals que poden així ser tractades més fàcilment pel mòdul de transferència. Per exemple, els adjectius catalans van normalment darrere dels substantius mentre que els anglesos van normalment davant; traduir comporta, per tant, canviar els adjectius de posició. Però per a això, cal identificar quines paraules són adjectius i substantius: la fase d'anàlisi es pot encarregar d'aquesta tasca perquè la fase de transferència puga aplicar regles generals sense preocupar-se de quins són els adjectius o els substantius concrets.

L'arquitectura de transferència és el model estàndard per a la traducció automàtica basada en regles o coneixement 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.

L'arquitectura de transferència té tres característiques interessants que mereixen ser esmentades:

Funcionament com a cadena de muntatge. El sistema de transferència funciona com una *cadena de muntatge*: com que els tres mòduls treballen d'esquerra a dreta i en una única passada, no cal que un mòdul espere que l'anterior acabe amb el text: poden treballar paral·lelament; això fa que els sistemes d'aquesta naturalesa siguen molt ràpids.

Modularitat. La divisió en *mòduls* o etapes ben diferenciades (anàlisi, transferència i generació) en permet la reutilització. Per exemple, si hem construït un sistema de transferència que tradueix de l'anglés a l'espanyol:

$$\text{en} \rightarrow \boxed{\text{A}_{\text{en}}} \rightarrow \boxed{\text{T}_{\text{en}\rightarrow\text{es}}} \rightarrow \boxed{\text{G}_{\text{es}}} \rightarrow \text{es}$$

podem aprofitar el mòdul d'anàlisi de l'anglés (A_{en}) per a construir un sistema de l'anglés al català:

$$\text{en} \rightarrow \boxed{\text{A}_{\text{en}}} \rightarrow \boxed{\text{T}_{\text{en}\rightarrow\text{ca}}} \rightarrow \boxed{\text{G}_{\text{ca}}} \rightarrow \text{ca}$$

o usar el mòdul de generació de l'espanyol G_{es} per a construir un sistema del neerlandés a l'espanyol:

$$\text{en} \rightarrow \boxed{\text{A}_{\text{nl}}} \rightarrow \boxed{\text{T}_{\text{nl}\rightarrow\text{es}}} \rightarrow \boxed{\text{G}_{\text{es}}} \rightarrow \text{es}$$

Reversibilitat parcial. Si se separen les dades lingüístiques usades per cada un dels tres mòduls, és possible una *reversibilitat parcial*. Si en el sistema anglés–espanyol de dalt separem les *dades lingüístiques* de cada un dels tres mòduls del *programari* que processa aquestes dades (dA_{en}, $\text{dT}_{\text{en}\rightarrow\text{es}}$, dG_{es}) podem definir un *motor* genèric de traducció (A, T, G) que val per a qualsevol parell de llengües:

$$\text{en} \rightarrow \boxed{\begin{array}{l}\text{dA}_{\text{en}}\\ \text{A}\end{array}} \rightarrow \boxed{\begin{array}{l}\text{dT}_{\text{en}\rightarrow\text{es}}\\ \text{T}\end{array}} \rightarrow \boxed{\begin{array}{l}\text{dG}_{\text{es}}\\ \text{G}\end{array}} \rightarrow \text{es}$$

Si ara volem escriure el sistema de traducció invers, espanyol–anglés, és a dir,

$$\text{es} \rightarrow \boxed{\begin{array}{l}\text{dA}_{\text{es}}\\ \text{A}\end{array}} \rightarrow \boxed{\begin{array}{l}\text{dT}_{\text{es}\rightarrow\text{en}}\\ \text{T}\end{array}} \rightarrow \boxed{\begin{array}{l}\text{dG}_{\text{en}}\\ \text{G}\end{array}} \rightarrow \text{en}$$

podríem traure avantatge del fet que hi ha grans semblances entre les dades lingüístiques de aquest sistema i les del sistema anterior:

- les dades que necessitem per a l'anàlisi de l'espanyol dA_{es} són molt similars a les dades de generació de l'espanyol dG_{es} del sistema existent: per a analitzar l'espanyol podem reciclar una bona part de les dades que s'usaven per a generar-lo en el sistema anterior;

- les dades que necessitem per a la generació de l'anglés dG_{en} són molt similars a les dades d'anàlisi de l'anglés dA_{en} del sistema existent: per a generar l'anglés podem reciclar una bona part de les dades que s'usaven per a analitzar-lo en el sistema anterior;
- les dades de transferència de l'espanyol a l'anglés $dT_{es \to en}$ són molt similars a les dades de transferència de l'anglés a l'espanyol $dT_{en \to es}$ del sistema existent: per a transferir d'espanyol a anglés podem usar una bona part de les dades que s'usaven per a transferir des de l'anglés a l'espanyol en el sistema anterior (per exemple, podem "pegar la volta" als diccionaris bilingües i podríem aprofitar-ne moltes entrades).

8.3.1 Sistemes de transferència morfològica avançada

En els sistemes de *transferència morfològica avançada* —també anomenats sistemes de *transferència sintàctica parcial* o *transformers* (Arnold et al. 1994, 4.2)— la fase d'*anàlisi* analitza morfològicament els mots de la frase i els desambigua en cas d'ambigüitat lèxica categorial però només identifica les relacions (sintàctiques) entre ells usant patrons molt senzills.[3] La secció 8.3.2 dóna més detalls sobre els processos i els mètodes d'anàlisi i generació morfològiques.

De fet, els sistemes de transferència morfològica avançada es poden veure com el resultat de fer una tercera aproximació, afegida a les dues (*aproximació oracional* i *aproximació de transferència*) discutides en l'apartat 8.1:

> **Aproximació de transferència parcial:** Quan les llengües involucrades no són massa diferents sintàcticament (per exemple, quan estan emparentades), no cal fer l'anàlisi sintàctica completa: la transferència lèxica és completa però la transferència estructural és parcial i local i només es fa on és necessària.

La fase de *transferència* pot consistir en un reordenament local (*transferència estructural*) 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 (*transferència lèxica*).

La fase de *generació* podria efectuar la substitució de les formes lèxiques de la LM per les corresponents formes superficials.

Com que els sistemes de transferència morfològica avançada no identifiquen realment les relacions sintàctiques entre els mots de la frase en la

[3] Alguns sistemes de transferència morfològica avançada disponibles en Internet són: SDL Transcend (`http://www.freetranslation.com`) Reverso (`http://www.reverso.net`), i Apertium (`http://www.apertium.org`).

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 (en) **adj subst** $\rightarrow$ (ca) **subst adj**

R_2 (en) **subst**$_1$ **subst**$_2$ $\rightarrow$ (ca) **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 (en) **adj subst**$_1$ **subst**$_2$ **subst**$_3$ $\rightarrow$ (ca) **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,[4] 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; en una oració llarga i amb moltes regles disponibles, açò pot ser un problema greu. 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

[4] En la gramàtica de la llengua, si una regla que estén un sintagma es pot aplicar repetidament a un determinat tipus de sintagma, aquest sintagma es pot allargar indefinidament. Un exemple clàssic d'això 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).

només es visita una vegada cada mot de la frase. Si no es pot aplicar un reordenament al primer mot pendent de processar, es tradueix aïlladament i es continua amb el següent mot.

Per a poder traduir *tall truck driver* seguint aquest esquema, caldria una regla que combinara R_1 i R_2:

R_4 (en) **adj subst**$_1$ **subst**$_2$ $\rightarrow$ (ca) **subst**$_2$ **adj prep_de subst**$_1$

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 substantiu–adjectiu per a identificar una classe de sintagmes nominals senzills, podem fer que la traducció correcta al català del sintagma nominal espanyol *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 espanyol *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):

(es) **subst adj** $\rightarrow$ (ca) **subst adj**
assigna gènere meta: **subst** $\rightarrow$ **adj**
assigna nombre meta: **subst** $\rightarrow$ **adj**

El reordenament i la concordança es poden combinar en la mateixa regla; per exemple, quan es tradueix de l'anglés al català la seqüència adjectiu–substantiu, la regla podria tenir aquesta forma:

(en) **adj subst** $\rightarrow$ (ca) **subst adj**
assigna gènere meta: **subst** $\rightarrow$ **adj**
assigna nombre meta: **subst** $\rightarrow$ **adj**

Les figures 8.2, 8.3 i 8.4 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.

L'estratègia de *transferència morfològica avançada* es pot veure com una formalització de l'estratègia mal anomenada *directa* que s'usava en els programes de traducció automàtica de primera generació (anys cinquanta i seixanta del segle passat): anàlisi morfològica rudimentària, consulta del diccionari bilingüe, i ajustos locals com ara reordenaments. 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, secció 4.2), de vegades erròniament anomenat *traducció directa* en vista de la seua simplicitat, 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".

8.3.2 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 7.2.1). 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, present d'indicatiu, 2a. persona del singular i *viu*, masculí, plural, 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.

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. Fixeu-vos en l'imperatiu espanyol *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 7.2.1): 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.

Per saber més sobre l'anàlisi morfològica

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

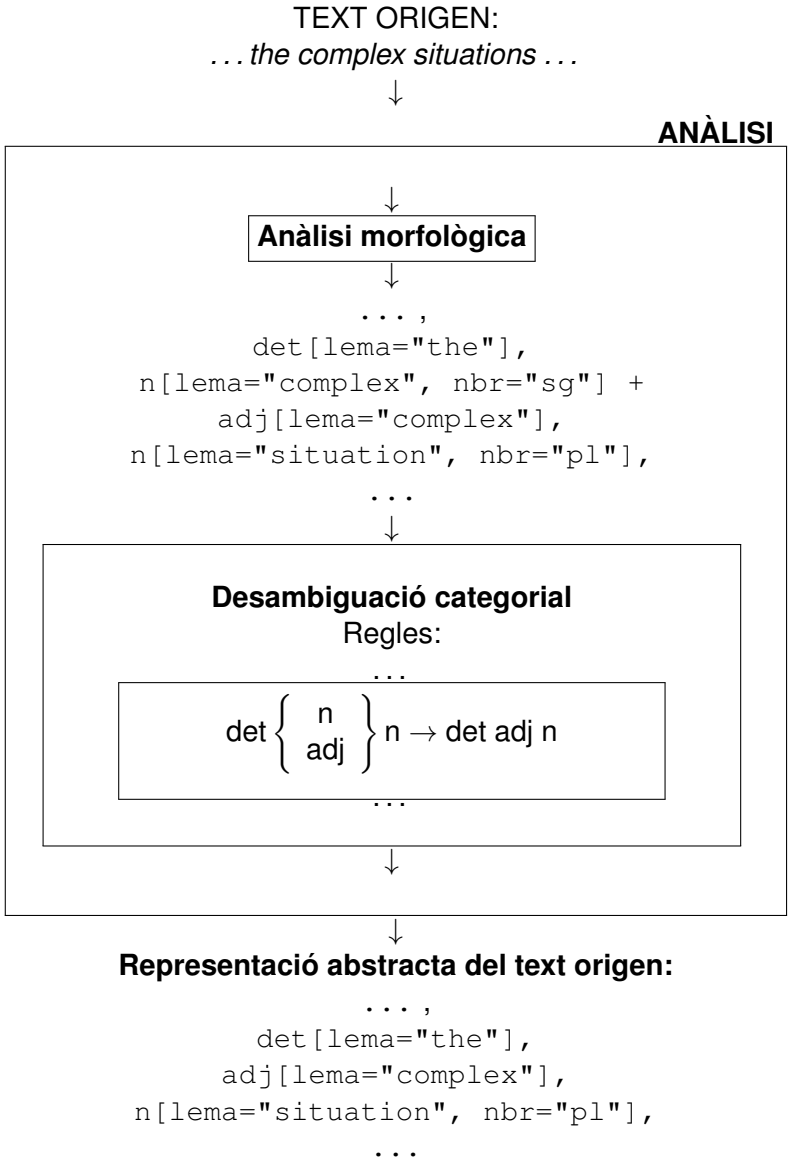


Figura 8.2: Fase d'anàlisi d'un sistema senzill de transferència morfològica avançada. El segment de text *the complex situations* conté el mot *complex* que pot ser un adjectiu (adj) o un substantiu (n); entre les regles del desambiguador categorial hi ha una regla que en aquest cas assigna la categoria d'adjectiu quan va entre un determinant i un substantiu.

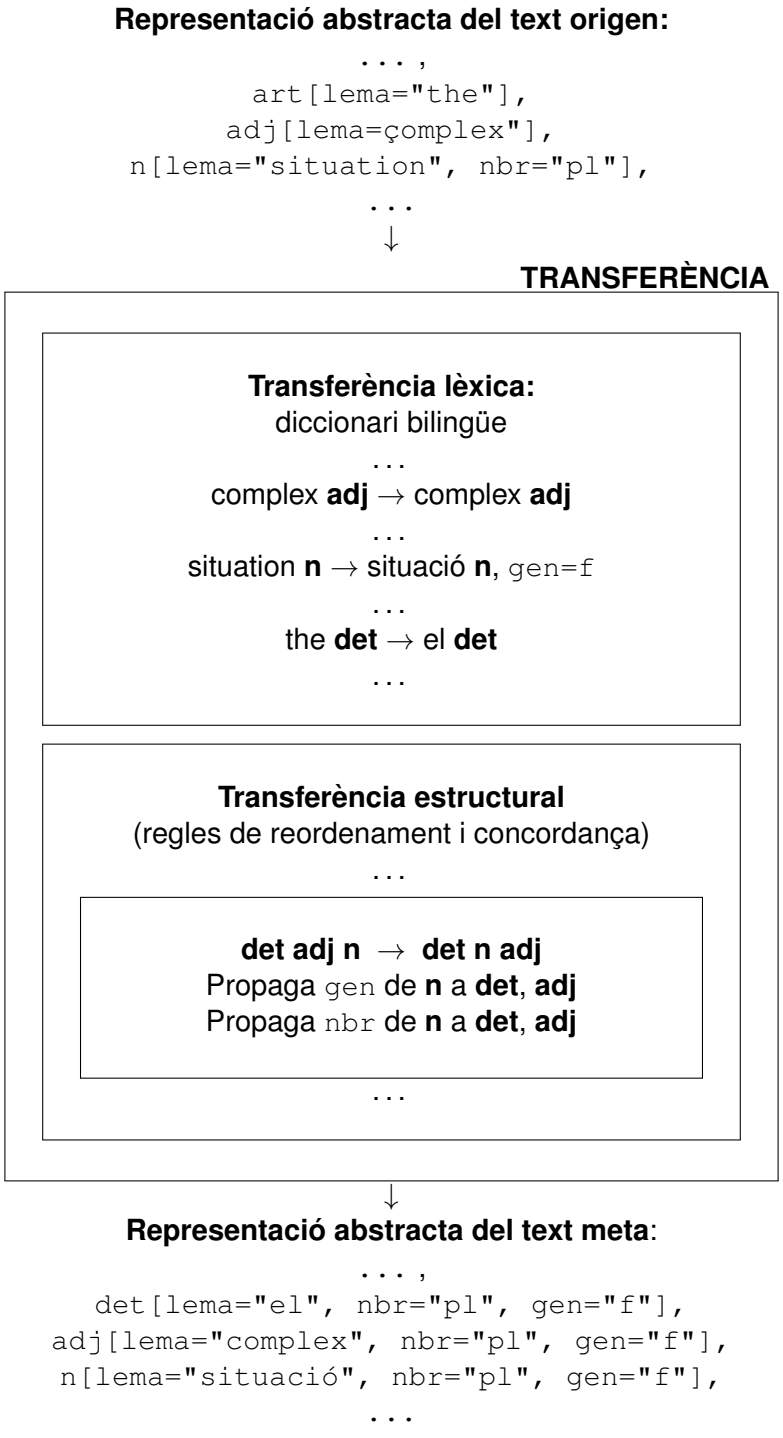


Figura 8.3: Fase de transferència d'un sistema senzill de transferència morfològica avançada, amb regles de transferència estructural que fan el reordenament i la concordança, entre les quals hi ha la regla per reordenar la seqüència determinant–adjectiu–substantiu i assegurar la concordança de nombre i gènere en la llengua meta.

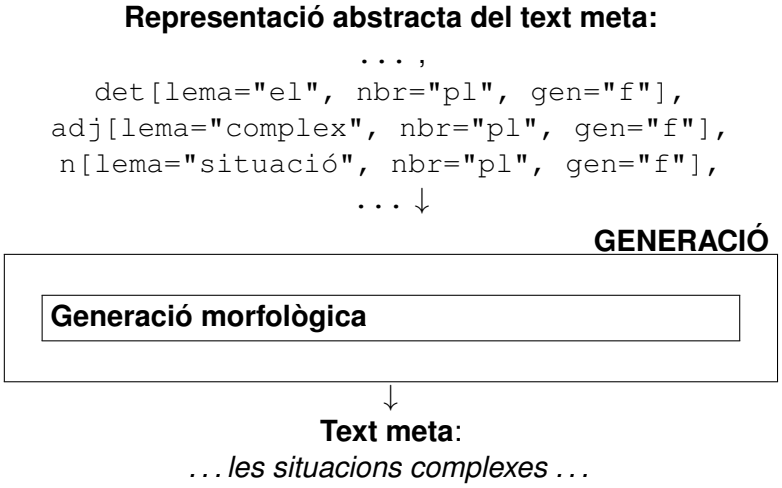


Figura 8.4: Fase de generació d'un sistema senzill de transferència morfològica avançada

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><imp>`.
- 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><imp><3p><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><imp> - <3p><pl>
```

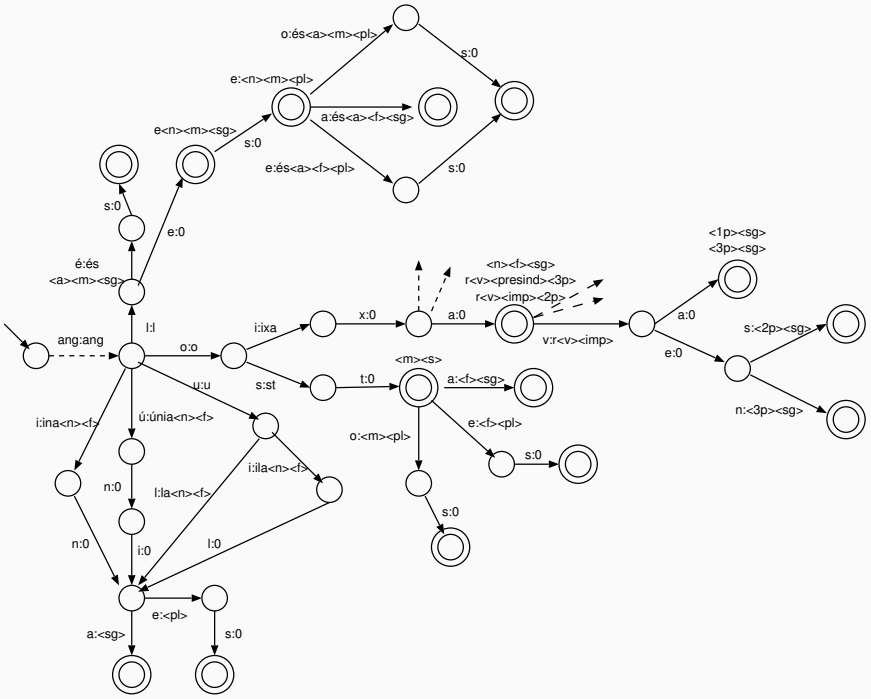
Si la forma superficial haguera estat *angostes*, tot el procés fins a *ango-* hauria estat idèntic. En el cas d'*angoixem*, el procés hauria estat idèntic fins a *angoixe-*:

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

```
r<verb><pressubj><1p><pl>
r<verb><imp><1p><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 (indicat amb una fletxa que no ve de cap lloc i 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 anterior 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.

8.3.3 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 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 8.3.4. En la fase de transferència, s'apliquen regles de *transferència estructural* que transformen la representació sintàctica de la frase d'entrada (la RALO) en una representació sintàctica de la traducció (la RALM) usant regles de transformació d'estructures i de *transferència lèxica* que tradueixen els mots en LO a mots en LM usant un diccionari bilingüe. 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.[5]

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:[6]

$\mathrm{SN} \rightarrow \mathrm{SG\ SN}$
$\mathrm{SN} \rightarrow \mathrm{A\ SN}$
$\mathrm{SN} \rightarrow \mathrm{N}$
$\mathrm{SG} \rightarrow \mathrm{SN\ GS}$

[5] 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.

[6] La sintaxi generativa actual (vegeu Chomsky (1996), Ramos (1992)) 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.

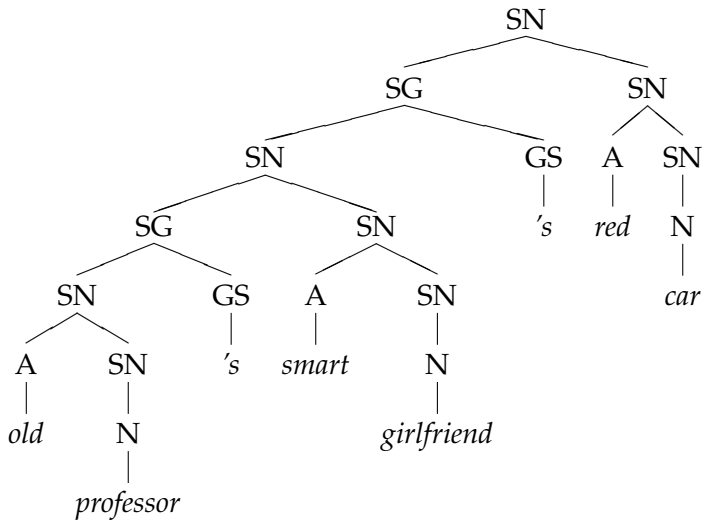
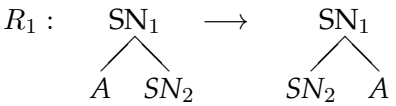
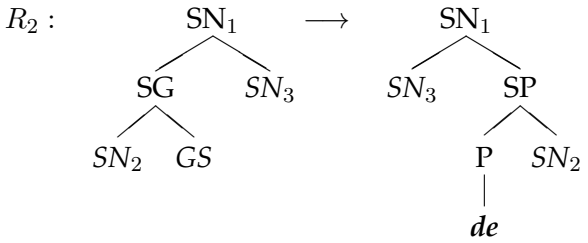


Figura 8.5: Arbre d'anàlisi sintàctica del sintagma nominal *The old professor's smart girlfriend's red car*.

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 figura 8.5. 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:[7]



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



[7]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.

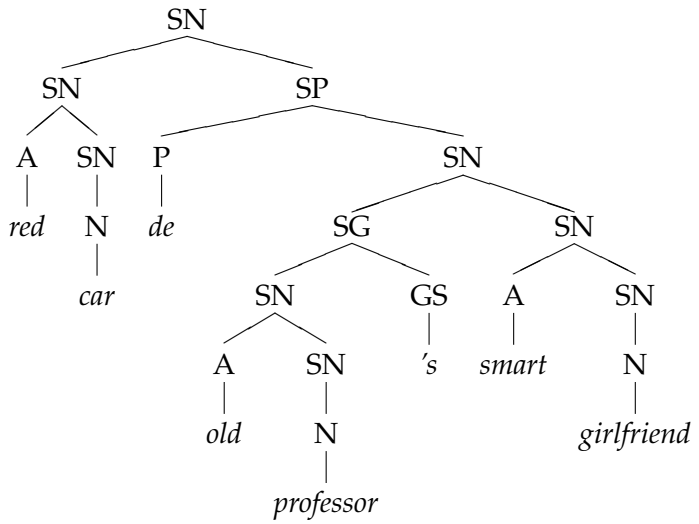


Figura 8.6: Arbre d'anàlisi sintàctica del sintagma *The old professor's smart girlfriend's red car* després d'aplicar-hi la regla R_2 al SN principal o arrel (vegeu el text i la figura 8.5).

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 8.6. 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 figura 8.7 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 entre llengües emparentades, com per exemple de l'espanyol al català, poques vegades es donen situacions com aquesta, ja que, en general, l'ordre dels mots no sol canviar tan radicalment; una construcció espanyola 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* 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[8] per a marcar els sintagmes nominals amb claudàtors:

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

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

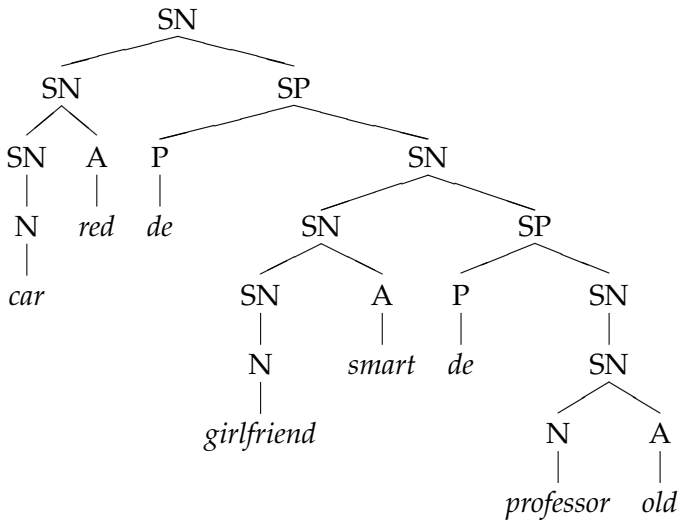


Figura 8.7: Arbre d'anàlisi sintàctica del sintagma *The old professor's smart girlfriend's red car* després d'haver-hi fet tots els reordenaments possibles amb les regles R_1 i R_2 (vegeu el text i les figures 8.5 i 8.6).

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

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 natural en un sistema de transferència sintàctica.

8.3.4 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, les oracions vàlides de la LO. S'ha de tenir en compte que 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; per això, els analitzadors han de ser *robustos* i ser capaços d'entregar anàlisis parcials o incompletes per a oracions que no estaven previstes. 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. Les regles solen correspondre's normalment amb 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.

Per saber més sobre l'anàlisi sintàctica

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 a l'oració completa—, o *descendents* (anglés *top-down*), en cas contrari (recordeu que els arbres d'anàlisi sintàctica estan "cap per avall": l'arrel és a dalt i les fulles, a baix). Si la frase és estructuralment ambigua (vegeu l'apartat 7.2.2), 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 7.2.4).

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

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 *empilen*) i quan el cim de la pila conté elements que segons la gramàtica i el context immediat posterior es poden agrupar en un subarbre, els desempila, 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 l'analitzador consulta en cada pas de l'anàlisi. L'ús de la taula permet construir el programa analitzador independentment de la gramàtica concreta: si canvia la gramàtica, només canvia la taula.

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:

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

Aquesta taula indica què cal fer en cada pas de l'anàlisi. En la taula, el símbol • indica tant el principi com el final de l'oració (l'anàlisi d'una oració comença empilant • 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 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.

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

(8.1) L'home porta la clau.

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

(8.2) **det n v det n** •

L'anàlisi, pas a pas, és la següent:

Pila	Entrada restant...	Acció
•	**det n v det n** •	empilar **det**
• **det**	**n v det n** •	empilar **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** •	empilar **v**
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v**	**det n** •	empilar **det**
• $[_{SN}\mathbf{det}[_{\bar{N}}\mathbf{n}]]$ **v det**	**n** •	empilar **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

8.3.5 Sistemes de transferència semàntica

Els sistemes de traducció automàtica basats en la transferència sintàctica (és a dir, en l'aproximació que diu que es poden transformar d'una banda les estructures sintàctiques —*transferència estructural*— i d'altra banda substituir el lèxic —*transferència lèxica*— fent ambdues operacions independentment) 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 directe*, *objecte indirecte*, etc.) i les relacions semàntiques (*agent*, *pacient*, *destinatari* etc.) dels constituents d'una frase poden variar d'una llengua a una altra. Heus ací alguns exemples:

- En la frase catalana *m'agraden els llimons*, qui produeix el plaer, és a dir, l'agent, (*els llimons*) fa de subjecte en l'oració, mentre que en l'equivalent anglesa (*I like lemons*) hi fa d'objecte directe o en l'equivalent portugués (*Eu gosto de limões*) apareix com a complement preposicional; qui experimenta el plaer fa d'objecte en català i de subjecte en anglés i en portugués. La semàntica d'aquestes frases es podria resumir en l'estructura abstracta

  ```
  donar_plaer(agent=llimons,dest=jo)
  ```

- Però encara pot haver-hi més maneres d'expressar sintàcticament el que a algú li produeix plaer fer una acció:

  – la catalana *m'agrada nadar* o l'espanyola *me gusta nadar*, on l'acció de nadar és el subjecte i el receptor de plaer l'objecte indirecte;

  – la de l'anglés *I like swimming* o el francés *j'aime nager*, on el receptor de plaer és subjecte i l'acció de nadar l'objecte;

  – la del portugués, que introdueix l'acció amb preposició: *eu gosto de nadar*, o

  - la neerlandesa *ik swemme graag* o l'alemanya *ich schwimme gern* on l'acció de nadar passa a ser el verb principal i l'acció d'agradar es converteix en un adverbi.

  La semàntica de totes aquestes frases es podria resumir en l'estructura abstracta

```
donar_plaer(agent=nadar(agent=jo),experimentador=jo).
```

- La noció de *portar un nom* també sol expressar-se de manera sintàcticament divergent. Per exemple, en català diem *Em dic Joan* o *Em diuen Joan*; en anglés *My name is Joan* o *I am called Joan*; en alemany *Ich heiße Joan*, etc.[9]

- En algunes llengües, trobem verbs que de vegades se solen anomenar *ergatius*, com ara en anglés el verb *sink* (*enfonsar*) quan porta subjecte i objecte, no és especial: el subjecte és l'agent i l'objecte el pacient: *we sank the ship*:

  ```
  enfonsar(agent=nosaltres,pacient=vaixell),
  ```

  però quan porta només un subjecte, aquest correspon al *pacient* de l'acció: *the ship sinks*

  ```
  enfonsar(pacient=vaixell),
  ```

  i l'agent sense especificar. En català també hi ha verbs ergatius com ara *bullir* (*Jo bull la llet / La llet bull*).

És cert que aquests casos es podrien tractar de manera particular en un sistema de transferència sintàctica, fent la transformació corresponent en l'arbre d'anàlisi sintàctica per a cada verb específic (és a dir, fent una transferència estructural dependent del lèxic) 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 l'el·lipsi (vegeu les pàgines 122 i 123). 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.

[9] En català de Mallorca s'usa una estructura similar amb el verb defectiu *nòmer*: *Jo nom Joan*.

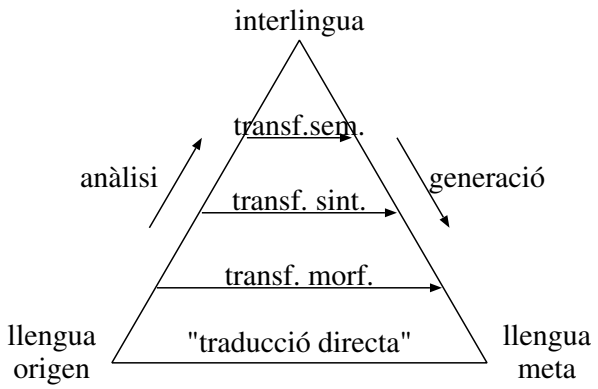


Figura 8.8: Com més profunda i complexa és l'anàlisi del text origen, més senzilla és (menys esforç comporta) la transferència a la representació corresponent de la llengua meta i més complexa la generació. L'anàlisi del text origen és tan profunda en els sistemes d'interlingua clàssics que no és necessària la transferència.

8.4 Sistemes basats en *interlingua*

Els sistemes anomenats d'*interlingua*[10] 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. 8.8). 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. Els sistemes clàssics usen representacions estructurals més o menys complexes per a representar les relacions semàntiques entre els elements de la frase. Però les interlingües no han de ser necessàriament el resultat d'una anàlisi profunda: el que han de ser necessàriament és *neutrals*; per exemple, alguns sistemes històrics com DLT (Hutchins i Somers 1992, cap. 17) usen com a *interlingua* una llengua *pivot* "natural" com l'esperanto, amb anotacions que resolen algunes ambigüitats típiques.[11]

En l'intent de representar els significats de totes les frases de totes les llengües, les interlingües clàssiques acabarien per ser "models del món".

[10] S'ha de tenir en compte que el terme *interlingua* es pot referir també a una llengua internacional planificada —no tan famosa com l'*esperanto*— molt basada en el llatí i amb vocabulari europeu, i que no té res a veure amb la traducció automàtica.

[11] Aquesta aproximació pot ser particularment útil quan les llengües entre les quals ha de traduir el sistema tenen una gran similitud sintàctica i semàntica, com en el cas de les llengües romàniques, amb l'excepció, potser, del romanés.

Això fa que, actualment, només s'hagen desenvolupat sistemes d'interlingua clàssics 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}, 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 8.9 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 cap a estructures lingüísticament neutrals), 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 8.10 il·lustra el cost d'afegir L_4 al sistema.

8.5 Sistemes de traducció automàtica basats en corpus

Totes les tècniques de traducció automàtica descrites fins ara són de naturalesa *deductiva*, és a dir, estan basades en teories i coneixements lingüístics sobre la traducció. Però recentment (sobretot en els primers anys del tercer mil·lenni) s'està produint un creixement espectacular de tècniques *inductives* de traducció automàtica, en les quals el sistema *aprén* automàticament a traduir entre dues llengües a partir d'un corpus paral·lel suficientment gran d'oracions en LO acompanyades de la seua traducció a la LM (vegeu

[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.

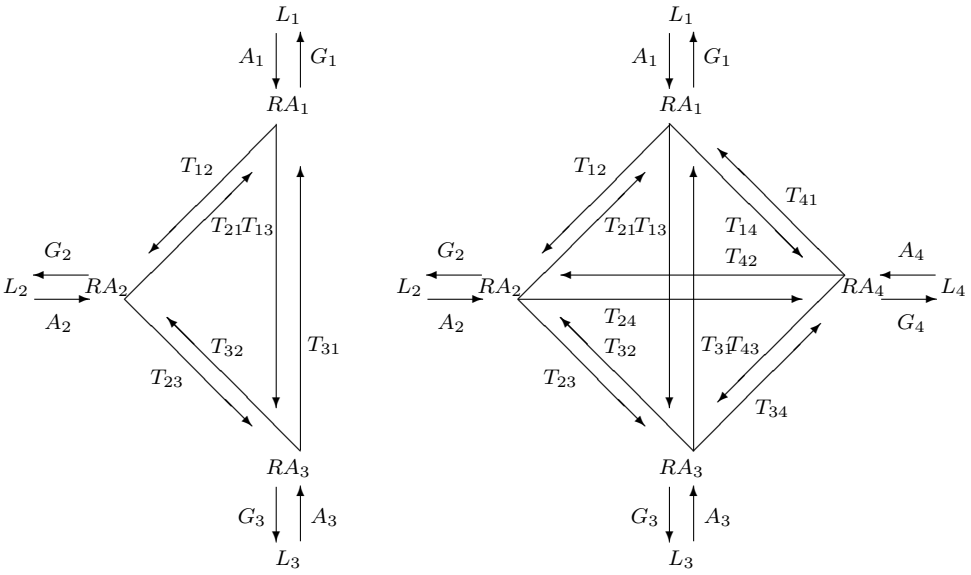


Figura 8.9: Cost d'afegir una quarta llengua L_4 a un sistema de transferència. Les entitats RA_1 a RA_4 són les representacions abstractes (tant RALO com RALM) que usen els mòduls de transferència.

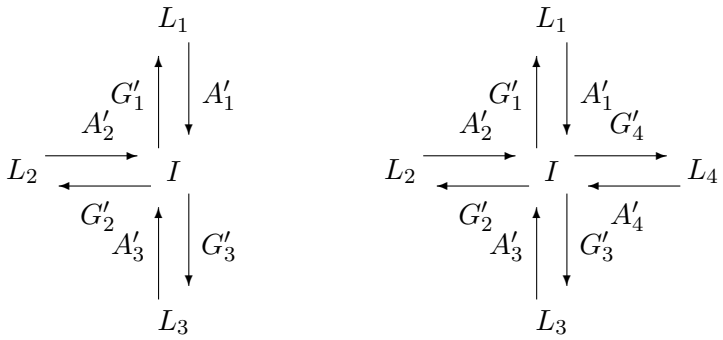


Figura 8.10: Cost d'afegir una quarta llengua L_4 a un sistema d'interlingua.

Anglés	Espanyol
It has been exciting in many ways .	Ha sido un trabajo apasionante en varios sentidos .
As the shadow rapporteurs know , this has been my first report during my time in Parliament and it has been a good learning experience .	Como bien saben los ponentes alternativos , éste ha sido el primer informe en el que he trabajado durante mi mandato parlamentario , y me ha venido muy bien como experiencia formativa .
It has also been very challenging to work on three reports and therefore also with other rapporteurs .	También ha sido un gran desafío trabajar en tres informes , y por lo tanto con otros ponentes .
It has been exciting .	Ha sido emocionante .

Figura 8.11: Oracions paral·leles anglés-espanyol extretes del corpus paral·lel Europarl (`http://www.statmt.org/europarl/`) amb les actes del Parlament Europeu de període 1996–2011.

l'apartat 10.2). Aquestes aproximacions inductives també reben el nom de *traducció automàtica basada en corpus*.

8.5.1 Sistemes de traducció automàtica estadística

La principal tècnica de traducció automàtica basada en corpus es la *traducció automàtica estadística* (en anglés *statistical machine translation*; SMT), la qual va ser inventada cap a finals dels huitanta per un grup d'investigadors d'IBM (Brown et al. 1990); els sistemes actuals són una evolució d'aquests.

A l'hora de traduir hi ha una diferència fonamental entre els sistemes basats en regles o coneixement i els sistemes estadístics: mentres que els sistemes basat en regles produeixen únicament una traducció, els sistemes estadístics generen una gran quantitat d'*hipòtesis de traducció* (idealment totes les possibles) i utilitzen models estadístics per *puntuar* les hipòtesis generades i escollir la millor de totes. Els principals models estadístics que s'usen per puntuar les hipòtesis de traducció són el *model de traducció* i el *model de llengua*, els quals s'expliquen més avall. La combinació d'aquests models fa que la hipòtesi de traducció que rep la puntuació *global* més alta no siga necessàriament la hipòtesi de traducció millor segons cada model per separat.

El **model de traducció** s'aprén a partir d'un corpus paral·lel amb les oracions ja alineades com el que es mostra en la figura 8.11. Primerament, s'han d'obtenir els *alineaments entre els mots* (vegeu-ne un exemple en la

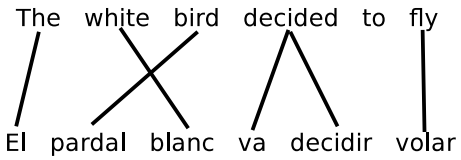


Figura 8.12: Alineament entre les paraules de l'oració en anglés *The white bird decided to fly.* and les paraules de l'oració en català *El pardal blanc va decidir volar.*

figura 8.12) per a després estimar el model de traducció a partir d'aquests alineaments.

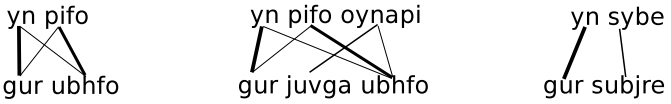
Tot i que sembla una tasca difícil per a un ordinador, els alineaments entre els mots es poden obtenir automàticament sense usar cap coneixement sobre les llengües dels textos a alinear mitjançat un procés iteratiu. La figura 8.13 il·lustra aquest procés amb un corpus menut de tres oracions paral·leles; si us fixeu, sense tenir cap coneixement de les llengües (perquè han estat inventades) les persones també som capaces d'obtenir aquests alineaments. El procés comença assumint que, per a cada oració paral·lela, tots els mots de l'oració en LM poden ser traducció de cadascun dels mots de l'oració en LO, i per tant, els assigna la mateixa probabilitat. En cada iteració el programa alineador visita totes les oracions paral·leles del corpus i va refinant aquestes probabilitats fins que l'estructura d'alineament queda definida. Aquest refinament es produeix perquè en cada iteració les probabilitats de la iteració anterior s'usen per acumular evidència en tot el corpus sobre la probabilitat de les correspondències entre mots i, a més, perquè els mots que són traducció mútua solen aparéixer junts en les mateixes oracions paral·leles, la qual cosa no succeeix si dos mots no són traducció un de l'altre.

Una vegada obtinguts els alineaments entre els mots, podem aprendre models probabilístics que indiquen, per exemple, la probabilitat que la traducció d'un determinat mot en una llengua siga la traducció d'un determinat mot en l'altra (un model de traducció de paraules o diccionari bilingüe probabilístic), o la probabilitat que la traducció d'una seqüència (segment) de mots en una llengua siga la traducció d'una seqüència (segment) de mots en l'altra (un model de traducció de segments). Aquest últim model l'usen els sistemes de traducció automàtica estadística basats en segments bilingües (en anglés *phrase-based statistical machine translation*; Koehn (2010)), els quals són els més usats en l'actualitat.

Però per produir bones traduccions no podem usar el model de traducció únicament perquè les traduccions serien poc naturals, gramaticals i fluides. El motiu és que el model de traducció no té en compte l'ordre en què apareixen els segments traduïts en la LM, ni el context en què aparei-

(a) Inicialització: Tots els alineament són igualment probables.

(b) Primera iteració: El alineament entre *yn* i *gur* guanya força.



(c) Segona iteració: El alineament entre *pifo* i *ubhfo* guanya força.

(d) Tercera i última iteració: L'estructura d'alineament que estava "oculta" queda al descobert.

Figura 8.13: Exemple que il·lustra el procés iteratiu que permet obtenir l'alineament entre les paraules de les oracions d'un corpus paral·lel. En aquest exemple el corpus consta de tres oracions paral·leles en dos llengües inventades. Aquestes oracions paral·leles són: *yn pifo–gur ubhfo*, *yn pifo oynapi–gur juvga ubhfo* i *yn sybe–gur subjre*. El gruix de les línies que connecten les paraules representa la probabilitat de l'alineament.

xen els segments en LO a l'hora de puntuar les seues possibles traduccions. Aquestes deficiències es mitiguen parcialment amb l'ús d'un model de la LM.

Un **model de llengua** es un model probabilístic que serveix per a mesurar la versemblança d'una oració o text en LM; es a dir, la seua fluïdesa o gramaticalitat.[14] Aquests models s'aprenen de forma automàtica a partir d'un corpus de text en LM i es basen en comptar la freqüència de segments de longitud fixa, normalment segments de fins a cinc paraules, per evitar assignar una versemblança nul·la a oracions que, tot i que són correctes, no apareixen en els corpus d'entrenament.[15] El model de llengua té en compte

[14]S'assumeix que el model de llengua s'aprén de textos naturals i gramaticalment correctes en LM.

[15]Per evitar assignar versemblances nul·les a una oració, a més d'utilitzar segments de poques paraules, aquests models també usen tècniques de suavitzat (en anglés *smoothing*)

l'ordre de les paraules i per tant assigna una versemblança major a l'oració *M'agrada menjar pernil del bo* que a l'oració *del bo menjar pernil M'agrada*, tot i que contenen els mateixos segments de text (*del bo*, *menjar pernil* i *M'agrada*). A més té en compte, tot i que indirectament, el context en què apareixen les paraules en LM, de manera que assigna una versemblança major a l'oració en espanyol (LM) *No piensa con la cabeza* que a l'oració *No piensa con el cabo*, on els segments *la cabeza* i *el cabo* són dues possibles traduccions del segment en català *el cap* que apareix en l'oració en LO *No pensa amb el cap*.

Per saber més sobre sistemes de traducció automàtica estadística

A més dels models de traducció i de la LM els sistemes de traducció automàtica estadística basats en segments bilingües combinen altres models per a establir la puntuació global d'una hipòtesi de traducció. A continuació es descriuen molt breument aquests models i per a què s'usen:

Model de reordenament lèxic: La seua funció és modelar diferents operacions de reordenament que es poden fer a l'hora de disposar les traduccions dels segments en LO. Les probabilitats d'aquestes operacions depenen dels segments concrets que s'estan reordenant i són tres: traducció monòtona (quan no hi ha reordenament), reordenament (quan la posició de la traducció del segment en qüestió i la de l'anterior s'intercanvien) i traducció discontinua (qual la traducció del segment es mou a una altra posició en l'oració en LM; es a dir, quan no és cap de les altres dues operacions).

Ponderació lèxica: Els segments usats per traduir poden ser molt llargs (normalment fins a 7 paraules), la qual cosa fa molt difícil estimar bé la seua probabilitat de traducció perquè els segments llargs solen aparéixer poques vegades en els corpus d'entrenament; això fa necessari l'ús d'un altre model per estimar la qualitat dels segments bilingües. Aquest model usa les probabilitats de traducció entre els mots (un diccionari bilingüe probabilístic) per obtenir un indicador de la qualitat d'els segments bilingües. Per exemple, la qualitat del segment bilingüe (*la comissió de balanços de finançament*, *the funding balance comission*), on l'alineament entre els mots és *la–the*, *comissió–comission*, *balanços–balance* i *finançament–funding*, depén de les probabilitats de traducció dels mots que han estat alineats.

Nombre total de paraules de l'oració: Quan es puntuen les hipòtesis de traducció es multipliquen moltes probabilitats, és a dir valors entre 0 i 1, de manera que com més llarga siga una traducció més probabilitats es multipliquen i més fàcil es arribar a tenir una puntuació molt prop de zero. Això fa que els sistemes preferisquen les traduccions curtes. Per a evitar això s'introdueix un model que compta el nombre de paraules en la hipòtesi de traducció i que fa que tinga relació amb el nombre de paraules de l'oració origen.

Nombre de segments: Aquest model es similar a l'anterior, però comptant el nombre de segments bilingües que s'han usat per a produir una hipòtesi de traducció. Com més llargs siguen els segments, menys segments s'usaran i més context tindran; i a l'inrevés, com més curts siguen els segments més segments faran falta per produir la hipòtesi de traducció.

---

de les probabilitat.

Tots aquests models (i els anteriors) es combinen per a obtenir una puntuació global per a cada hipòtesi de traducció i poder escollir així la millor. Aquesta combinació es fa assignant un pes (importància) a cada model que s'obté mitjançat un procés automàtic (*tuning*) que intenta maximitzar la *qualitat* de les traduccions proporcionades pel sistema en traduir un corpus de *desenvolupament*.

Consulteu el llibre de Koehn (2010) per saber més sobre els models que s'usen per a traduir, el procés de *tuning* i les mesures automàtiques de la qualitat que usen.

Per saber més sobre sistemes basats en corpus

Hi ha hagut altres aproximacions inductives a la traducció automàtica, com ara els sistemes de *traducció automàtica basada en exemples*, tot i que a hores d'ara ja no s'usen. La *traducció automàtica basada en exemples* intenta construir *plantilles* de traducció a partir dels exemples observats en el corpus d'oracions paral·leles i *generalitzar-les* perquè servisquen en noves situacions. Per exemple, si sabem que el substantiu anglés *ski* es tradueix per *esquí* i que la locució substantiva *ski station* es tradueix per *estació d'esquí* podem generalitzar aquesta última locució substituint *ski* per qualsevol altres substantiu N, de manera que la traducció de "*N station*" és "estació de N"; així, si la traducció de *train* és *tren*, la traducció de *train station* és *estació de tren*, etc. (exemple pres de Carl et al. 2001). Fixeu-vos que la traducció automàtica basada en exemples pot necessitar que la mostra de frases i traduccions estiga, a més, anotada lingüísticament (en l'exemple, indicant quins mots o estructures funcionen com un nom).

8.6 Qüestions i exercicis

Els exercicis marcats amb (*) són més difícils.

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 de l'espanyol al català que usa regles que reordenen seqüències de formes lèxiques segons les categories lèxiques. 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 espanyoles 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 suahili a les quinze llengües i de les quinze llengües cap al suahili. En una oferta de treball, WorldTrans demana experts en suahili però no demana cap expert en traducció entre suahili i cap de les quinze llengües. Quina classe de sistema de traducció automàtica és LetTrans? Justifiqueu la resposta.

4. (*) Imagineu que teniu un sistema de traducció automàtica que treballa amb dues llengües, diguem-ne A i B, en els dos sentits de traducció: $A \rightarrow B$ i $B \rightarrow A$, que traduïm un text origen T en llengua A a la llengua B mitjançant aquest traductor automàtic, generant un text T', i que després usem aquest mateix sistema per a traduir T' de nou a la llengua A; anomenarem T'' el nou text en llengua A.

$$T \overset{A \rightarrow B}{\longrightarrow} T' \overset{B \rightarrow A}{\longrightarrow} T'' \tag{8.3}$$

El text T'' serà previsiblement diferent del text T. Trieu dues llengües A i B i indiqueu quins canvis són previsibles, classificant-los segons la naturalesa lingüística dels fenòmens que han causat els canvis, explicant la raó del resultat si cal amb un exemple. Heu d'indicar *tres tipus diferents* de canvi.

5. (*) Imagineu que sou part d'un equip de desenvolupament d'un sistema de traducció automàtica de l'anglés al català basat en l'estratègia de transferència morfològica avançada (apartat 8.3.1). Els informàtics del projecte us demanen consell sobre les regles de reordenament del sistema, ja que, per motius tècnics, només poden afegir-n'hi tres.

   Indiqueu quines serien les 3 regles que proposaríeu, tenint en compte que han de produir, com a mínim, tres oracions ben traduïdes en el corpus d'oracions següents (la traducció ideal s'indica entre parèntesis, tot i que no sempre podrà ser aconseguida):

   (a) *A dark autumn night* (*Una nit fosca de tardor*)
   (b) *A high tide* (*Una marea alta*)
   (c) *A magic dark silhouette* (*Una silueta fosca màgica*)
   (d) *An autumn tide* (*Una marea de tardor*)
   (e) *A dark magic silhouette* (*Una silueta màgica fosca*)
   (f) *A dark autumn high tide* (*Una marea alta de tardor fosca*)
   (g) *A dark night* (*Una nit fosca*)

   Deixeu de banda la concordança i centreu-vos només en els reordenaments. Assenyaleu quina seria la traducció del sistema per a totes les oracions anteriors usant el conjunt de regles que heu proposat.

6. 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.

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

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

8. (*) Dues traduccions possibles del mot català *cap* a l'espanyol 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.

9. 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.

10. 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 tripletes lema–categoria–informació morfològica possibles.

11. 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.

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

    (a) Els sistemes de transferència morfològica avançada.

    (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 avançada.

    (b) Els sistemes de transferència semàntica superficial.

    (c) Els sistemes d'*interlingua*.

14. Primerament, elegiu un idioma meta (francés, anglés o alemany) i un idioma origen (català o espanyol). Després, per als idiomes elegits, doneu un exemple de traducció mot a mot inacceptable en *tres* d'aquests cinc casos:

    (a) homografia mal resolta d'un mot

(b) polisèmia mal resolta d'un mot

(c) problemes de concordança

(d) ambigüitat estructural mal resolta

(e) problemes amb l'ordre dels mots

15. *Interlingua*, a més de ser el nom de la representació intermèdia dels sistemes indirectes sense transferència, és el d'una llengua artificial bàsicament d'arrel llatina, amb una flexió simplificada, i amb un vocabulari dissenyat per a ser comprensible a molts europeus. Una característica important d'interlingua és que els determinants (*un*, *le*, *alcun*, *iste*, *mi*, *tu*, etc.) i els adjectius són invariables. Els plurals dels noms es fan amb *-s* o *-es*. Imagineu que tenim un sistema de transferència morfològica avançada que tradueix d'interlingua al català (o a l'espanyol) usant aquestes quatre regles:

R_1 detecta **det**–**n** i escriu trad(**det**)–trad(**n**), fent concordar trad(**det**) en gènere i en nombre amb trad(**n**)

R_2 detecta **det**–**n**–**adj** i escriu trad(**det**)–trad(**n**)-trad(**adj**), fent concordar trad(**det**) i trad(**adj**)en gènere i en nombre amb trad(**n**)

R_3 detecta **n**–**adj** i escriu trad(**n**)–trad(**adj**), fent concordar trad(**adj**) en gènere i en nombre amb trad(**n**)

R_4 detecta **adj**–**n** i escriu trad(**n**)–trad(**adj**), fent concordar trad(**adj**) en gènere i en nombre amb trad(**n**)

Si no es pot usar informació de concordança, la traducció dels determinants i els adjectius es fa en masculí singular. Indica quines traduccions al català (o a l'espanyol) produirà aquest sistema per a les frases següents i per què:

(a) *Un longe viage*

(b) *Un longe viages*

(c) *Un viages longe*

(d) *Longe viages*

(e) *Un governamento non democratic*

(f) *Un governamentos non democratic*

(g) *Tu melior ideales*

(h) *Un bon solution*

16. (*) La traducció d'una oració es pot veure com una interpretació d'aquesta (és a dir, com l'expressió en la llengua meta del seu significat). El *principi de composicionalitat semàntica* postula que la interpretació

d'una oració es construeix combinant les interpretacions dels mots seguint precisament les agrupacions successives (constituents) que indica l'arbre d'anàlisi sintàctica de l'oració, partint dels mots i anant cap a l'arrel de l'arbre. Indiqueu en quin (o quins) tipus de sistema de traducció automàtica trobem un disseny que aplica, exactament o aproximadament, el principi de composicionalitat. Raoneu breument la resposta.

17. El programari que porten instal·lat les naus de la confederació galàctica inclou un programa que tradueix una de les llengües majoritàries del planeta Zkanagg, el tazkannwat, al català. El sistema és un sistema de transferència morfològica avançada estàndard, que llegeix els textos d'esquerra a dreta, mot a mot, busca en l'entrada els patrons de categories lèxiques que conté en el seu catàleg, selecciona el més llarg, reordena i concorda els mots del patró, els escriu, i continua després de la zona reordenada. Algunes traduccions són errònies perquè el sistema no té un catàleg massa complet de regles. Fixeu-vos en els exemples i digueu quins són els patrons que detecta i quines les regles de reordenament associades.

(8.4) *Thlong u knaar uw phlagyw.*
Va adquirir el navegant el-OBJ control-OBJ
TA: El navegant va adquirir el control (correcta).

(8.5) *Thlong u knaar qimratt uw phlagyw.*
Va adquirir el navegant estelar el-OBJ control-OBJ
TA: El navegant estelar va adquirir el control (correcta).

(8.6) *Thlong u knaar na Zkannag uw phlagyw.*
Va adquirir el navegant de Zkannag el-OBJ control-OBJ
TA: El navegant de Zkannag va adquirir el control (correcta).

(8.7) *Thlong u knaar qimratt na Zkannag uw phlagyw.*
Va adquirir el navegant estelar de Zkannag el-OBJ control-OBJ
TA: *El navegant estelar va adquirir de Zkannag el control.
Correcta: El navegant estelar de Zkannag va adquirir el control.

18. Els mots no són tots igualment freqüents en els textos. De fet, si ordenem els mots d'un gran corpus de text real (de qualsevol tipus i de qualsevol idioma) pel nombre de vegades que hi apareixen, començant pel més freqüent, el nombre d'aparicions es redueix dramàticament segons que anem baixant per la llista. Típicament, el mot més freqüent

pot arribar a constituir el 10% de tot el text, però el segon només cobreix al voltant del 5%, el tercer al voltant del 3%, etc.; quan arribem al 100é mot més freqüent ja hem de parlar del 0,1% (una vegada cada 1.000 mots), i si arribem a la posició 1000, del 0,01% (una vegada cada 10.000 mots). En resum, la distribució no és gens homogènia: uns pocs mots són els més freqüents i la majoria són moltíssim menys freqüents. De fet, és típic que la majoria dels mots siguen *hapax legomena*, és a dir, mots que han aparegut només una vegada en tot el corpus. Si haguéreu de supervisar la construcció dels diccionaris d'un sistema de traducció automàtica, per a què us podrien servir aquestes constatacions estadístiques?

19. (*) Els sistemes de traducció automàtica entre dues llengües amb sintaxi similar no necessiten fer massa reordenaments perquè l'ordre dels mots no varia massa d'una llengua a altra. A pesar d'això, la traducció mot per mot no és practicable perquè el gènere i el nombre gramatical d'alguns substantius varia i els adjectius, articles, etc., que l'acompanyen no concordarien correctament: cast. *una señal muy clara* → cat. **una senyal molt clara* (correcte: *un senyal molt clar*); cast. *me gusta la leche fría* → ital. **mi piace la latte fredda* (correcte: *mi piace il latte freddo*). Una manera d'identificar zones on s'ha d'establir la concordança és detectar seqüències de mots, de manera similar a com es fa en els sistemes de transferència morfològica, però sense reordenar-les. Per exemple, detectar la seqüència **art**–**subst** pot servir per propagar el gènere i el nombre del substantiu a l'article. Fixeu-vos en les frases espanyoles següents i les traduccions al català fetes per un sistema que usa aquesta estratègia i deduïu quines són les seqüències que detecta i quines no. Justifiqueu la vostra resposta.

    (a) *Nos ofreció un postre → Ens va oferir unes postres*

    (b) *Nos ofreció un postre buenísimo → Ens va oferir unes postres boníssimes*

    (c) *Nos ofreció un buen postre → *Ens va oferir un bon postres*

    (d) *Nos ofreció un postre típico buenísimo → *Ens va oferir unes postres típiques boníssim*

    (e) *Nos ofreció un postre muy bueno → *Ens va oferir unes postres molt bo*

20. Indiqueu quina d'aquestes afirmacions és falsa.

    (a) Els sistemes de transferència sintàctica fan anàlisi sintàctica sense fer anàlisi morfològica.

    (b) Els sistemes de transferència sintàctica només usen informació bilingüe en una de les tres fases.

(c) La fase de transferència d'un sistema de transferència sintàctica realitza transformacions d'arbres d'anàlisi sintàctica d'acord amb regles determinades.

21. Elegeix la seqüència que està en l'ordre temporal correcte:

    (a) Preedició, postedició, traducció per transferència, disseminació.

    (b) Preedició, traducció per transferència, disseminació, postedició.

    (c) Preedició, traducció per transferència, postedició, disseminació.

22. En quin tipus de sistema de traducció automàtica tindrien bàsicament la mateixa representació les frases *David és vist per Lluc* i *Lluc veu David*?

    (a) En un sistema de transferència morfològica.

    (b) En un sistema de transferència semàntica o d'interlingua clàssic.

    (c) En un sistema de transferència sintàctica.

23. Quantes llengües naturals ha de conéixer l'equip d'experts que ha d'incorporar una nova llengua a un sistema de traducció automàtica basat en interlingua que ja en té 7?

    (a) Set.

    (b) Una.

    (c) Vuit.

24. Com més profunda és l'anàlisi en un sistema de traducció automàtica...

    (a) ...més complexa és la transferència.

    (b) ...més senzilla és la generació.

    (c) ...més senzilla és la transferència.

25. Si una oració té només una ambigüitat lèxica pura, té només un arbre únic d'anàlisi sintàctica. Per tant, si es tradueix aquesta oració amb un sistema de traducció automàtica indirecta per transferència sintàctica...

    (a) ...el sistema es bloquejarà perquè només opera a nivell sintàctic

    (b) ...l'ambigüitat lèxica no afecta el resultat perquè no afecta la sintaxi

    (c) ...pot encara produir-se un error en la traducció per causa de l'ambigüitat lèxica de transferència

26. Un sistema de traducció automàtica per transferència tradueix en qualsevol sentit entre quatre llengües. Si volem afegir-hi una cinquena llengua perquè traduïsca en qualsevol sentit entre cinc llengües, quants mòduls nous cal escriure?

    (a) 4 de transferència, un d'anàlisi i un de generació

    (b) 5 de transferència, un d'anàlisi i un de generació

    (c) 8 de transferència, un d'anàlisi i un de generació

27. En quina de les tres fases d'un sistema de transferència s'usen els diccionaris bilingües?

    (a) En la d'anàlisi.

    (b) En la de generació

    (c) En la de transferència.

28. Un amic meu ha dissenyat un sistema de traducció automàtica entre l'espanyol i el portugués, però tot i que m'assegura que no ha programat cap tractament de l'ambigüitat estructural, el seu sistema tradueix perfectament un munt d'oracions amb aquest tipus d'ambigüitat. És açò possible?

    (a) No. Probablement ha dissenyat també un mòdul de preedició i el sistema elimina automàticament qualsevol causa d'ambigüitat.

    (b) Sí, açò pot ocòrrer quan es donen els anomenats *passis gratuïts*; de segur que, si insistim, trobarem alguna oració que hi serà traduïda malament.

    (c) Sí, si es tracta d'oracions en què aquesta ambigüitat es deu a mots polisèmics i el programa té un diccionari prou complet.

29. Els informàtics que participen en el disseny d'un sistema de traducció per interlingua t'informen que cada una de les fases del sistema s'ha d'executar en un ordinador diferent. Quants ordinadors hem de comprar?

    (a) Dos, un per a la fase d'anàlisi i un altre per a la de generació.

    (b) Dos, un per a la fase d'anàlisi i un altre per a la de transferència.

    (c) Tres, un per a la fase d'anàlisi, un altre per a la de transferència i un tercer per a la de generació.

30. Quants mòduls d'anàlisi i de generació hem d'afegir en total a un sistema basat en transferència que ara mateix permet traduir entre 4 llengües, si volem incorporar-hi una llengua més de manera que el sistema puga traduir (tant en un sentit com en l'altre) entre totes les llengües existents i la nova?

Primer mot		Segon mot		Expressió	
fondos	(410)	*estructurales*	(203)	*fondos estructurales*	(63)
precio	(415)	*máximo*	(202)	*precio máximo*	(2)
algunos	(403)	*sectores*	(211)	*algunos sectores*	(1)
hacia	(409)	*ellos*	(204)	*hacia ellos*	(0)
otra	(411)	*crisis*	(203)	*otra crisis*	(0)

Taula 8.1: Freqüències d'aparició de parells de mots sobre economia.

(a) 2

(b) 4

(c) 6

31. Si una forma superficial té només una forma lèxica, però dues possibles traduccions a una altra llengua...

    (a) ... es tracta d'un mot homòfon.

    (b) ... es tracta d'un mot homògraf.

    (c) ... probablement un sistema autòmatic haurà de recórrer a informació estadística o regles sobre el context per triar una de les solucions.

32. (*) En un corpus de textos en espanyol sobre economia de 925.461 mots estudiem quan apareixen mots conjuntament. En concret, i per posar un exemple, estudiem parells de mots gramaticalment vàlids on el primer mot apareix unes 400 vegades en total en el corpus i el segon mot hi apareix unes 200. Fixeu-vos en la taula 8.1 de freqüències d'aparició d'alguns parells. A pesar que tant el primer mot com el segon mot de cada parell tenen freqüències similars, en algun cas les freqüències d'aparició conjunta són molt elevades i en uns altres casos són molt més reduïdes. Podríeu explicar la causa d'aquesta variació? Per a quina aplicació de la informàtica a la traducció podrien servir els resultats d'un estudi numèric com aquest?

33. Elegeix una llengua origen (català, espanyol, anglés, francés o alemany) i una llengua meta (català, espanyol, anglés, francés o alemany) i dóna *tres* exemples de frases que es poden traduir acceptablement *mot per mot* però tals que si canviem *un mot* de les frases per un altre de la mateixa categoria, la traducció *mot per mot* resulta incorrecta. En cada una de les frases, la raó lingüística per la qual la segona traducció és incorrecta ha de ser diferent.

34. (*) Estudieu els següents sintagmes nominals en maori (una llengua polinèsia parlada en Nova Zelanda):

(8.8) *Te* *whare* .
Art. def. sg. casa .
La casa.

(8.9) *Ngā* *whare* .
Art. def. pl. casa .
Les cases.

(8.10) *Te* *whare* *nui* .
Art. def. sg. casa gran .
La casa gran.

(8.11) *Te* *whare* *nui* *o* *te* *aroha* .
Art. def. sg. casa gran de Art. def. sg. amor .
La casa gran de l'amor.

(8.12) *Ngā* *whare* *nui* .
Art. def. pl. casa gran .
Les cases grans.

Com en els exemples, en maori la majoria dels noms i adjectius són invariables. Imagineu que sou part d'un equip de desenvolupament d'un sistema de traducció automàtica del maori al català (o a l'espanyol) basat en l'estratègia que hem anomenat en l'apartat 8.3.1 *transferència morfològica avançada*.[16] Especifica completament *dues* regles (indicant possibles reordenaments i operacions per a assegurar la concordança) que permeten donar la traducció correcta de les oracions de dalt i de les següents. No us preocupeu de la contracció preposició–article.

(8.13) *Ngā whare nui o te aroha* (Les cases grans de l'amor)

(8.14) *Te hau o te aroha* (El vent de l'amor)

(8.15) *Ngā pukapuka o te whare* (El llibre de la casa)

(8.16) *Ngā ingoa o te pukapuka nui* (Els noms del llibre gran)

[16] És a dir, llig les oracions mot a mot d'esquerra a dreta i fa l'anàlisi morfològica de cada mot, prova de detectar la seqüència més llarga de mots que concorda amb alguna seqüència de categories lèxiques que té en el seu catàleg, processa la seqüència, i continua immediatament després de la seqüència processada.

(8.17) *Te ingoa o ngā whare* (El nom de les cases)

35. Es vol construir un sistema de traducció automàtica que traduïsca entre qualssevol dues llengües del grup format pel portugués, el gallec, el català, l'espanyol i l'italià. A més, es requereix que es puguen afegir fàcilment altres llengües com l'occità, el sard o l'asturià. No es busca la perfecció sinó més aviat traduccions en brut ràpides i fàcils d'entendre o de corregir (és a dir, amb pocs errors). Tenint en compte les llengües implicades, argumenteu a favor i en contra d'usar un sistema d'interlingua clàssic (amb anàlisi semàntica profunda) o un sistema de transferència, indicant en cada cas com haurien de ser les representacions intermèdies usades.

36. Tenim un sistema de traducció automàtica multilingüe que tradueix en qualsevol direcció entre les llengües que considera. Per a afegir-hi una nova llengua hem escrit 6 mòduls. Com era el sistema abans de l'addició de la nova llengua?

    (a) D'interlingua amb 4 llengües (hem afegit la quinta).

    (b) De transferència amb 2 llengües (hem afegit la tercera)

    (c) De transferència amb 4 llengües (hem afegit la quinta)

37. Quin dels tres mòduls d'un sistema de traducció automàtica de transferència espanyol–anglés conté les regles que indiquen que el passat de *bring* és *brought* i que el plural de *foot* és *feet*?

    (a) El de transferència.

    (b) El d'anàlisi.

    (c) El de generació.

38. Quin tipus de sistema de traducció automàtica per transferència analitza els textos originals fins arribar a categories com ara *agent*, *pacient*, *destinatari*, *instrument*, *experimentador*, etc.?

    (a) Els de transferència morfològica avançada.

    (b) Els de transferència sintàctica.

    (c) Els de transferència semàntica.

39. Tenim un sistema basat en interlingua que tradueix entre 6 idiomes (L_1, L_2, ..., L_6) i volem incorporar l'idioma L_7. Els experts que hi treballaran...

    (a) ...han de saber traduir entre la llengua L_7 i les altres sis.

    (b) ...no necessiten saber res de les llengües L_1 a L_6.

(c) ...han d'escriure 12 mòduls de transferència més, 6 des de la llengua L_7 i 6 cap a la llengua L_7.

40. Quin dels tres mòduls d'un sistema de traducció automàtica indirecta per transferència és monolingüe i tracta amb la llengua meta?

(a) El de generació.

(b) Tots els mòduls són bilingües, no n'hi ha cap de monolingüe.

(c) El de transferència.

41. En quin dels tres mòduls d'un sistema de traducció automàtica indirecta per transferència es fan els reordenaments dels mots de la llengua original perquè l'ordre siga l'adequat en la llengua meta?

(a) En el d'anàlisi.

(b) En el de transferència.

(c) En el de generació.

42. Un traductor automàtic per transferència morfològica avançada ...

(a) ...resol la polisèmia mitjançant l'ús d'un analitzador morfològic.

(b) ...resol la polisèmia mitjançant l'ús d'un desambiguador lèxic categorial.

(c) ...no pot resoldre la polisèmia amb cap dels programes esmentats en les altres dues opcions.

43. Indiqueu quina de les afirmacions següents és certa. Per norma general, els sistemes de traducció automàtica ...

(a) ...tradueixen cadascuna de les oracions una per una sense tenir en compte la resta d'oracions del text a traduir.

(b) ...tradueixen directament (mot per mot) de a llengua origen a la llengua meta.

(c) ...necessiten construir una interpretació completa del text abans de traduir-ho.

44. Els sistemes de traducció automàtica estadística ...

(a) ...aprenen a traduir a partir de diccionaris bilingües fets a mà i de textos monolingües en la llengua meta.

(b) ...aprenen a traduir a partir de textos *comparables* en ambdues llengües (textos que parlen del mateix però no són traducció mutua) i de textos monolingües en la llengua meta.

(c) Cap de les altres respostes es correcta.

45. Per a què usen els sistemes de traducció automàtica estadística el *model de llengua*?

    (a) Per a mesurar la versemblança (fluïdesa) de les traduccions.

    (b) Per a emmagatzemar les diferents alternatives de traducció d'un segment de text.

    (c) Els sistemes de traducció automàtica estadística no fan servir cap *model de llengua*.

8.7 Solucions

1. Per exemple, L_1=espanyol 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 abraça el segment *de francés* i trenca el sintagma;

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

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

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

4. Tipus de canvis (per exemple, català→espanyol→català):

    - Canvi d'un mot per un sinònim per causa de l'elecció diferent d'equivalents en un sentit i en un altre, *darrer*→*último*→*últim* o fins i tot per un que no ho és, *direcció*→*dirección*→*adreça*.
    - Canvi d'un mot per un altre per causa d'una homografia en alguna de les dues llengües *com aquest*→*como este*→*menjo aquest*; *riu sec*→*río seco*→*ric sec*.

- Pèrdua de mots: *en tinc dos→tengo dos→tinc dos*; *hi van arribar tard→llegaron tarde→van arribar tard*.
- Canvis de concordança: *La dona cosia el coixí cansada→La dona cosía la almohada cansada→La dona cosia el coixí cansat*. Quan tradueix del català a l'espanyol, *cansada* no concorda amb coixí i es tradueix independentment, però a la tornada *almohada* sí que concorda amb *cansada* i el sistema els tradueix com si formaren un sintagma.

5. Per exemple, amb les regles

   - $R_1 : a\ n\ \rightarrow\ n\ a$,
   - $R_2 : a_1\ a_2\ n\ \rightarrow\ n\ a_2\ a_1$
   - $R_3 : n_1\ n_2\ \rightarrow\ n_2$ "de" n_1

   es tradueixen bé totes excepte la (a) i la (f), que quedarien: "*Una [tardor fosca]$_{R_1}$ nit" i "*Una [tardor fosca]$_{R_1}$ [marea alta]$_{R_1}$" perquè les regles són incapaces de reconéixer els sintagmes complets.

6. (b)
7. (c)
8. (b), vegeu l'apartat 7.2.4.
9. (c)
10. (c)
11. (b)
12. (c)
13. (c)
14. (només a tall d'exemple) Si la llengua origen és el català i la llengua meta és l'anglés, tenim:

    (a) homografia mal resolta d'un mot: *ara rius → now *rivers* en comptes de *now you laugh*.

    (b) polisèmia mal resolta d'un mot: *rebrem el president a l'estació → we will welcome the president at the *season* en comptes de *at the station*.

    (c) problemes de concordança: *aquella gent estava feliç → *that people *was happy* en comptes de *those people were happy*.

(d) ambigüitat estructural mal resolta: *Dona'm la clau d'aquell sistema → give me the key *from that system* en comptes de *give me the key to that system*.

(e) problemes amb l'ordre dels mots: *Jo he estat sempre un professional responsable → *I have been always a professional responsible* en comptes de *I have always been a responsible professional*

15. S'hi indiquen les traduccions i, entre claudàtors, la regla aplicada en cada cas:

(a) *Un* $[_{R_4}$ *longe viage*] → *Un viatge llarg*

(b) *Un* $[_{R_4}$ *longe viages*] → **Un viatges llargs*

(c) $[_{R_2}$ *Un viages longe*] → *Uns viatges llargs*

(d) $[_{R_4}$ *Longe viages*] → *Viatges llargs*

(e) $[_{R_1}$ *Un governamento*] *non democratic* → *Un govern no democràtic*

(f) $[_{R_1}$ *Un governamentos*] *non democratic* → **Uns governs no democràtic*

(g) *Tu* $[_{R_4}$ *melior ideales*] → **El teu millors ideals*

(h) *Un* $[_{R_4}$ *bon solution*] → **Un bona solució*

16. Entre els tipus de sistemes de traducció automàtica indirectes, el primer que comença a aplicar, almenys parcialment, el principi de composicionalitat és el de transferència sintàctica, ja que construeix la traducció usant com a pas intermedi un arbre d'anàlisi sintàctica de l'oració original. Per tant, els sistemes amb anàlisis més avançades (transferència semàntica, interlingua) també l'apliquen.

Però els sistemes de transferència sintàctica no apliquen exactament el principi de composicionalitat semàntica, ja que es basen en l'aproximació que es poden traduir separadament: d'una banda, els mots (transferència lèxica) substituint-los pels seus equivalents i, d'altra banda, els arbres, transformant-ne l'estructura. Aquesta aproximació pot no funcionar perquè de vegades les transformacions dels arbres depenen de la interpretació de mots concrets i de parts de l'oració. En aquest sentit, els sistemes de transferència semàntica i d'interlingua proven de construir una representació semàntica a partir de l'arbre i de la semàntica dels mots, de manera que fan una interpretació més general del principi.

17. (8.18)

Thlong	*u*	*knaar*	*uw*	*phlagyw.*
Va adquirir	el	navegant	el-OBJ	control-OBJ

TA: El navegant va adquirir el control (correcta).

R_1: **verb art nom → art nom verb**

Resultat correcte.

(8.19) *Thlong u knaar qimratt uw phlagyw.*
Va adquirir el navegant estelar el-OBJ control-OBJ
TA: El navegant estelar va adquirir el control (correcta).

R_2: **verb art nom adj** $\rightarrow$ **art nom adj verb**

(8.20) Resultat correcte.

Thlong u knaar na Zkannag uw phlagyw.
Va adquirir el navegant de Zkannag el-OBJ control-OBJ
TA: El navegant de Zkannag va adquirir el control (correcta).

R_3: **verb art nom prep nompropi** $\rightarrow$ **art nom prep nompropi verb**

Resultat correcte.

(8.21) *Thlong u knaar qimratt na Zkannag uw*
Va adquirir el navegant estelar de Zkannag el-OBJ
phlagyw.
control-OBJ
TA: *El navegant estelar va adquirir de Zkannag el control.
Correcta: El navegant estelar de Zkannag va adquirir el control.

No ha estat capaç de detectar el patró **verb art nom adj prep nompropi** i aplica la regla R_2 que és la més llarga que concorda. El resultat és que el sintagma preposicional *de Zkannag* queda darrere del verb.

18. Com que l'objectiu de l'equip que dissenya els diccionaris és que tinguen la cobertura més alta possible (és a dir, que deixen el mínim possible de mots sense traduir), l'única estratègia raonable és la d'ordenar els mots de la llengua original per freqüències d'aparició i anar introduint-los en el diccionari en aquest ordre, de manera que en cada moment sempre estem augmentant la cobertura del diccionari tan ràpidament com és possible.

19. Vegem que passa amb cada una de les oracions:

    (a) Nos ofreció un postre $\rightarrow$ *Ens va oferir unes postres*: La traducció és correcta. Sembla que reconeix la seqüència (1) **art–subst** i propaga el nombre i el gènere del substantiu a l'adjectiu.

    (b) Nos ofreció un postre buenísimo $\rightarrow$ *Ens va oferir unes postres boníssimes*: La traducció és correcta. Sembla que reconeix la seqüència (2) **art–subst–adj** i propaga el nombre i el gènere del substantiu tant a l'article com a l'adjectiu,

(c) Nos ofreció un buen postre → **Ens va oferir un bon postres*: No funciona. No reconeix la seqüència **art–adj–subst**, i tradueix mot per mot.

(d) Nos ofreció un postre típico buenísimo → **Ens va oferir unes postres típiques bonissim*: funciona incorrectament perquè no reconeix la seqüència completa **art–subst–adj–adj**; en canvi, sí reconeix la seqüència més curta (2) **art–subst–adj** i propaga el gènere del substantiu només a l'article i al primer adjectiu. Després, el sistema continua traduint mot per mot.

(e) Nos ofreció un postre muy bueno → **Ens va oferir unes postres molt bo*: funciona incorrectament perquè no reconeix la seqüència completa **art–subst–adv–adj**; en canvi, sí reconeix la seqüència més curta (1) **art–subst** i propaga el gènere del substantiu només a l'article. Després, el sistema continua traduint mot per mot.

El sistema només ha usat dues seqüències (1: **art–subst** i 2: **art–subst–adj**) per a intentar fer la concordança.

20. (a)

21. (c), vegeu l'apartat 6.5.

22. (b)

23. (b)

24. (c)

25. (c)

26. (c)

27. (c)

28. (b)

29. (a)

30. (a)

31. (c)

32. Si la distribució dels mots fóra al atzar, la freqüència de tots els parells de mots seria la mateixa i molt baixa. Però hi ha mots que tendeixen a estar junts (col·locacions, unitats lèxiques multimot, unitats terminològiques) més que l'atzar.

    Per exemple, el mot "fondos" apareix davant del mot "estructurales" 63 vegades de les 203 vegades que apareix "estructurales", és a dir,

unes 3 de cada 10 vegades, quan a l'atzar apareixeria 410 vegades per cada 925.461, és a dir, unes 4 vegades cada 10.000. Per tant, apareix quasi mil vegades més freqüentment que l'atzar.

Es pot demostrar que, a pesar de ser menys freqüents, "precio máximo" o "algunos sectores" també tendeixen a estar junts per damunt de l'atzar, pot ser per ser col·locacions pròpies del tema econòmic.

Un estudi de bigrames (parelles) com aquests pot servir:

- primàriament, per a identificar unitats terminològiques ("fondos estructurales", "Real Decreto", "política monetaria"), col·locacions ("hacer frente", "tomar posiciones"), o noms d'entitat ("Nueva York", "Rodrigo Rato", "Unión Europea") pròpies del text en qüestió.
- secundàriament, per a decidir automàticament, per a un mot que té diverses traduccions, quina és la traducció que "sona més natural" davant o darrere de la traducció d'una altra.

33. Els següents exemples estan presos per al parell espanyol–català; en cada cas, la primera frase és un exemple de traducció correcta i el segon d'incorrecta:

    Homografia: Le traje un [sombrero] → Li vaig portar un [barret]; Le traje un [traje] → Li vaig portar un [vaig portar]* (correcte: vestit).

    Polisèmia: El [canto] de la sirena → El [cant] de la sirena; El [canto] de la moneda → El [cant] de la moneda*. (correcte: cantell, viu)

    Concordança de gènere o nombre : [La] indicación era [inequívoca] → [La] indicació era [inequívoca]; [La] señal era [inequívoca] → [La] senyal era [inequívoca]* (correcte: el, inequívoc)

    Anàfora: Cualquier [indicación] es importante para quien la comprenda → Qualsevol [indicació] és important per a qui [la] comprenga; Cualquier [señal] es importante para quien la comprenda → Qualsevol [senyal] és important per a qui [la] comprenga.*

34. Dues regles són suficients (la resta va bé mot per mot):

    R_1: • detectar "determinant nom";
    • propagar el nombre (sing./pl.) del determinant maori (te/ngā) al nom català;
    • propagar el gènere (masc./fem.) del nom català al determinant català.

    R_2: • detectar "determinant nom adjectiu";

- propagar el nombre (sing./pl.) del determinant maori (te/ngā) al nom i a l'adjectiu catalans;
- propagar el gènere (masc./fem.) del nom català al determinant i a l'adjectiu catalans.

35.
- Avantatges d'interlingua:
  - són necessaris menys mòduls nous quan s'afig una llengua nova al sistema (només un d'anàlisi i un de generació).
  - no calen experts bilingües per a construir mòduls de transferència (és poc versemblant que existisquen experts asturià-català o asturià-sard).
- Desavantatges d'interlingua:
  - Vista la semblança sintàctica entre les llengües involucrades, sembla excessivament costós fer l'esforç de dissenyar una representació d'interlingua, fer l'anàlisi i la generació completa dels textos (lèxica, sintàctica, semàntica) quan una transferència morfològica completa i sintàctica parcial seria suficient.
- Avantatges de transferència:
  - Les llengües són prou similars perquè un sistema de transferència morfològica completa i sintàctica parcial amb poques regles done resultats acceptables.
- Desavantatges de transferència:
  - Per descomptat, cada vegada que s'afig una llengua a un sistema amb N llengües s'han d'escriure $2N$ mòduls de transferència i calen experts bilingües per a construir-los tots.

Els desavantatges d'interlingua s'atenuarien si en comptes d'una representació interlingual semàntica (basada en nocions com ara *agent*, *pacient*, *destinatari*, *temps*, etc.) fóra més aviat de naturalesa lèxica. Fins i tot, podria ser similar a una llengua humana. El llatí clàssic no, perquè a pesar de ser l'origen de totes les llengües del sistema té una sintaxi —verb final— i morfologia –declinació— molt diferents; la llengua artificial anomenada interlingua —"le lingua international facile e de aspecto natural elaborate per linguistas professional como denominator commun del linguas le plus diffundite in le mundo"[17]— amb anotacions sintàctiques i marques de desambiguació podria ser una millor opció.

36. (b)

37. (c)

[17]URI: `http://www.interlingua.com`.

38. (c)
39. (b)
40. (a)
41. (b)
42. (c)
43. (a)
44. (c)
45. (a)

Capítol 9

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.

9.1 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 mètode d'avaluació o de mesurament estàndard és un bon indicador de la magnitud del problema".

Un concepte clau és el d'*utilitat*. La traducció automàtica serà *millor* o *de més qualitat* com més *útil* siga per a un propòsit previst. La *utilitat* depén de l'aplicació. Si la traducció automàtica s'usa per a la disseminació, és a dir, com a base per a produir un text adequat per a ser publicat, serà més útil com menys esforç siga necessari per a convertir-la en adequada (posteditar-la). Però si la traducció automàtica s'usa tal com és per a una aplicació d'assimilació, és a dir, per a comprendre un text escrit en una altra llengua, la utilitat augmenta amb la seua intel·ligibilitat.

9.2 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:
   - l'**avaluació d'adequació**, que serveix per a "determinar la idoneïtat [utilitat] dels sistemes de TA en un context operacional especificat" —per exemple, per a decidir si el sistema de TA és útil per a traduir el correu comercial d'una empresa alimentària—;
   - 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
   - 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.
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) o professionals externs (*consultors*) contractats a l'efecte;
   (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. Tradicionalment, la qualitat s'ha considerat de manera relativament desconnectada de les aplicacions concretes, i s'ha vist com una combinació (en proporcions difícils de determinar[1]) de diversos factors, com ara: la *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); la *naturalitat* o *gramaticalitat* del text; l'adequació de l'estil o del registre dels documents traduïts, etc.

[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".

Aquesta avaluació se sol 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 (en anglés *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]

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); com s'ha discutit en el capítol 6, la noció central és la de *propòsit* de la traducció:

- L'**avaluació** d'un sistema que s'usa **per a la *disseminació*** de textos s'ha de fer estimant d'alguna manera l'esforç de postedició, ja que el sistema serà més útil com més reduït siga aquest esforç.
  Una possible mesura quantitativa de la qualitat que aproxima (Sager 1993, p. 264) l'esforç de postedició per part de professionals de la traducció és la *taxa d'error per mot* (o taxa de mots corregits). Aquesta mesura s'ha de calcular sobre un conjunt suficientment gran de textos *representatius* de la tasca de traducció i es calcula com el percentatge d'insercions, esborraments i substitucions de mots estrictament necessaris per a transformar la traducció automàtic en brut en una traducció adequada al propòsit. Aquesta mesura té l'inconvenient que dóna la mateixa importància a totes les operacions de correcció, independentment del mot i això pot no ser adequat perquè l'esforç necessari per a corregir tots els mots no és el mateix.[4]
  Una altra mesura de l'esforç de postedició és el temps que es tarda en posteditar una traducció automàtica per fer-la adequada al propòsit previst. Mesurar el temps de postedició té l'inconvenient que no tots els posteditors són igualment eficients ni tenen la mateixa experiència posteditant.
- L'**avaluació** d'un sistema usat **per a l'*assimilació*** d'informació s'ha de fer de manera diferent: ací la utilitat està relacionada més amb la intel·ligibilitat, i es podria determinar directament a través de qüestionaris de comprensió (Jones et al. 2007) o similars (O'Regan i Forcada 2013; Ageeva et al.

[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 a l'espanyol (Mira i Gimènez i Forcada 1998; Forcada 2000).

[4]Per exemple, no és el mateix corregir un article que ha estat mal concordat amb el substantiu a què acompanya, que un terme d'especialitat que per a poder corregir-lo potser ens em de documentar abans.

2015) o indirectament, per exemple, estudiant l'èxit a l'hora d'executar una tasca amb les instruccions traduïdes automàticament (Doherty et al. 2012).

(b) *La facilitat d'ús 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).

Per saber més sobre els problemes de posteditar per determinar la qualitat

La determinació de la qualitat d'una traducció en brut per còmput del nombre de correccions necessàries no està exempta de problemes:

- Si suposem que existeix una única traducció acceptable del text origen (el que és suposar molt) 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 amb el nombre mínim d'operacions d'inserció, esborrament i substitució de mots; aquest nombre mínim es pot considerar una *distància*, i, de fet, matemàticament, ho és: s'anomena *distància d'edició* (en anglés *edit distance*). 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. Avaluar (corregir) la traducció en brut comporta 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. A més, aquest tipus d'avaluació és bastant costós, ja que per a obtenir una medició fiable de la qualitat és necessari corregir textos de milers de paraules.

9.2.1 Anàlisi de costos i beneficis

En el cas concret d'una aplicació de disseminació, finalment, des d'un punt de vista econòmic, el que és rellevant a l'hora de decidir si s'adopta o no un sistema de traducció automàtica és *una comparació dels costos i dels beneficis* d'usar aquest sistema de TA en comptes d'usar exclusivament els serveis de professionals de la traducció: per exemple, si costa més (en des-

peses de personal) la postedició (revisió) dels textos meta produïts pel sistema (afegint-hi el cost d'usar el sistema de TA) que la traducció completa dels textos origen per part de professionals, l'adopció del sistema de TA no convé a una empresa. Per ampliar la primera aproximació esmentada en la pàgina 105,

$$\textbf{cost}\left(\begin{array}{c}\text{traducció automàtica}\\ + \\ \text{postedició}\end{array}\right) < \textbf{cost}(\text{traducció professional}),$$

hauríem de considerar altres factors, és a dir, totes les despeses en què s'incorre quan s'adopta la traducció automàtica seguida de postedició:

- **Costos de *funcionament*** (cost efectiu per mot), que ha de tenir en compte:
  - l'amortització del traductor automàtic (en cas d'adquisició)
  - el servei tècnic i el manteniment del sistema
  - la migració (adaptació dels programes que s'usen, l'adquisició de sistemes informàtics)
- **Costos de *preedició* i de preparació**: cal preparar i potser preeditar (vegeu l'apartat 6.5) els textos que s'han de traduir
- **Costos de postedició**: depén de la *qualitat* del text en brut i de la formació dels posteditors, als quals se'ls pot pagar per hores de treball, per quantitat de text corregit, etc.
- **Costos de formació**, ja que els professionals han d'aprendre a usar una nova tecnologia:
  - **Formació en l'ús del programa de traducció automàtica**: els professionals han d'aprendre a usar, configurar i potser mantenir el nou programari associat
  - **Formació en postedició,** la qual ha de permetre que els professionals
    - coneguen el comportament del programa de traducció automàtica (per exemple, quins són els errors típics que comet);
    - aprenguen tècniques de correcció, com ara l'ús avançat del processador de textos (macroinstruccions, substitució de patrons, etc.)

9.3 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, p. 262) 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" (Sager 1993, p. 261) 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" (Sager 1993, p. 261). 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.

Per saber més sobre avaluació: avaluació predictiva

Hi ha un tipus d'avaluació que es pot considerar com a cas particular de l'avaluació diagnòstica definida en l'apartat 9.2, 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*, i s'aplica principalment als sistemes de TA basats en regles.

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, és a dir, de transferència morfològica, sintàctica, etc., vegeu el cap. 8), i, en segon lloc, un conjunt de textos o frases d'avaluació (en anglés, *test suite*) que els permeta obtenir detalls concrets sobre les dades lingüístiques (p.ex., les 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 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 (normalment els exemples on el sistema no tradueix adequadament donen molta més informació que els exemples que es tradueixen adequadament). 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 8.3).

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 comportament del sistema en aquestes situacions concretes.

Hi ha una relació molt estreta entre les tècniques descrites i l'anomenada *enginyeria inversa*, o determinació detallada de l'estratègia usada per un programa (en aquest cas, de traducció automàtica) per a reprogramar-la en un altre.

9.4 Qüestions i exercicis

1. Quina característica d'un sistema de traducció automàtica s'ha de considerar com a especialment important quan s'avalua l'aplicació del sistema a l'*assimilació*?

   (a) els seus útils de postedició assistida.

   (b) els seus útils de preedició assistida.

   (c) la velocitat de resposta.

2. Per què és difícil avaluar la qualitat d'una traducció automàtica comptant la quantitat mínima de postedició necessària per fer-lo adequat quan no hi ha una traducció de referència?

   (a) No és que siga difícil; sense traducció de referència és absolutament impossible.

   (b) Perquè aquesta tasca no es pot fer sense conéixer profundament l'estratègia usada pel sistema de traducció automàtica.

   (c) És relativament senzill corregir el text perquè siga adequat però és molt difícil fer-ho fent-hi el mínim nombre de canvis necessaris.

3. Quan es vol usar un sistema de traducció automàtica per a l'*assimilació* d'informació, a què donaríeu *menys* pes en l'avaluació?

   (a) Facilitat de postedició de la traducció en brut.

   (b) Intel·ligibilitat de la traducció en brut.

   (c) Velocitat.

4. Trieu la resposta errònia. A l'hora de decidir l'adopció d'un sistema de traducció automàtica per a la postedició ...

   (a) ... haurem de fer una avaluació amb textos semblants als que cal traduir.

   (b) ... els textos a usar en l'avaluació hauran de tenir un nombre de mots suficient que ens permeta extrapolar els resultats a la resta de texts.

   (c) ... haurem d'avaluar, mitjançant qüestionaris o un mètode equivalent, la intel·ligibilitat dels textos traduïts en brut.

5. Esteu avaluant un sistema de traducció automàtica. Indiqueu quina d'aquestes tres magnituds no usaríeu com a indicador directe de l'esforç de postedició.

   (a) La intel·ligibilitat del text meta en brut.

   (b) El nombre mínim de mots que cal canviar en el text meta en brut per a fer-lo adequat al propòsit previst, expressat com a percentatge respecte al nombre total de mots.

   (c) El temps necessari que cal invertir per a convertir el text meta en brut en un text adequat al propòsit previst, expressat com a minuts per cada 1.000 paraules.

Quant al concepte d'*avaluació predictiva*, mireu a més els exercicis 2, 17 i 19 del capítol 8.

9.5 Solucions

1. (c)
2. (c)
3. (a)
4. (c)
5. (a)

Capítol 10

Memòries de traducció

> *Existing translations contain more solutions to more translation problems than any other existing resource* (Isabelle et al. 1993)

10.1 Introducció

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.

L'**objectiu** és, per tant, aprofitar traduccions anteriors per a no repetir l'esforç quan s'han de fer traduccions noves. Ha de quedar clar que, per a poder fer-ho, els textos originals i les traduccions han d'estar en format informatitzat (fitxers de text).

[1] Com veurem més avall, en comptes de traduir mot a mot fent una simple substitució de cada mot origen pel(s) mot(s) meta corresponents, 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.

10.2 Bitextos

Suposarem que, com a resultat del treball anterior de traducció, tenim parells de textos (E, D) on E és el *text esquerre* (en la *llengua esquerra*) i D és el *text dret* (en la *llengua dreta*), i volem traduir de la llengua esquerra a la llengua dreta.[2] De fet, quan dos textos E i D són equivalents (és a dir, traducció un de l'altre) direm que el parell (E, D) és un *bitext* o *text paral·lel*.

Vegeu aquest exemple on la llengua esquerra és el català i la llengua dreta, l'espanyol:

> E: "Tenim textos equivalents i els volem aprofitar per a fer noves traduccions. Quan dos textos són equivalents diem que formen un bitext."
>
> D: "Tenemos textos equivalentes y los queremos aprovechar para hacer nuevas traducciones. Cuando dos textos son equivalentes decimos que forman un bitexto."

Però els bitextos complets no es poden aprofitar tal com estan perquè és molt improbable que ens encarreguen de nou exactament la mateixa tasca de traducció, és a dir, que ens encarreguen traduir de nou un text E' quan tenim ja el bitext (E', D'). El que sí que és probable és que algunes parts del text nou E' apareguen també en la part esquerra d'alguns dels bitextos que ja tenim. Per això, necessitem obtenir a partir d'ells bitextos més menuts, amb parts esquerres que tinguen possibilitat d'aparéixer de nou en el futur.

10.2.1 Segmentació de bitextos

La primera operació consisteix a *segmentar* o *dividir* automàticament cada un dels dos textos en unitats més menudes, que anomenarem *segments*, usant algun criteri programable (vegeu més avall). La figura 10.1 mostra el resultat de la segmentació dels textos E i D. Com s'hi veu, pot ser que inicialment el nombre de segments del text esquerre E siga diferent del nombre de segments del text dret D.

10.2.2 Alineament de bitextos. Unitats de traducció

El resultat de la segmentació encara no és útil: no queda clara la correspondència entre els segments esquerres i els segments drets. Necessitem *revisar la segmentació*, és a dir, ajuntar segments o partir segments, en un costat o en altre, fins que tinguem el mateix nombre de segments i siguen traducció mútua. D'aquesta operació se'n diu *alinear* el bitext.

Direm que un bitext (E, D) està *alineat* si les dues parts tenen el mateix nombre N de segments, $E = e_1e_2e_3 \dots e_N$ i $D = d_1d_2d_3 \dots d_N$, i els seus

[2] Parlem d'*esquerra* i *dreta* perquè pot ser que no siga important (o que no se sàpia) quin dels dos textos és l'original i quin és la traducció.

E	D
e_1	d_1
e_2	d_2
e_3	d_3
...	...
...	d_M
...	
e_L	

Figura 10.1: Els bitextos E i D, segmentats, respectivament, en L i M segments: $E = e_1 e_2 \dots e_L$ i $D = d_1 d_2 \dots d_M$. En l'exemple, el text esquerre té dos segments més (és a dir, $L = M + 2$).

E	D
e_1	d_1
e_2	d_2
e_3	d_3
...	...
e_N	d_N

Figura 10.2: Els bitextos E i D, segmentats i alineats, en N unitats de traducció $(e_1, d_1), (e_2, d_2), \dots (e_N, d_N)$

segments són traducció mútua: e_1 és traducció de d_1 (o viceversa), e_2 és traducció de d_2, etc. És a dir, el text alineat ens proporciona N bitextos (segments paral·lels) (e_1, d_1), (e_2, d_2), (e_3, d_3), etc., més menuts (per això els escrivim amb minúscula); aquests bitextos s'anomenen normalment *unitats de traducció* (UT): vegeu la figura 10.2.

Així és més probable que quan ens donen un text nou E' tinguem traduccions per a alguns dels seus fragments. El bitext de més amunt es podria alinear per a formar les unitats de traducció

(*Tenemos*, *Tenim*)
(*textos equivalents*, *textos equivalents*)
(*i els volem aprofitar*, *y los queremos aprovechar*)
...
(*formen un bitext*, *forman un bitexto*)

Com ja s'ha dit més amunt, una operació molt important per a la reutilització o el *reciclatge* de traduccions antigues és la d'*alinear* els textos i les traduccions existents per a identificar fragments o *unitats de traducció* que

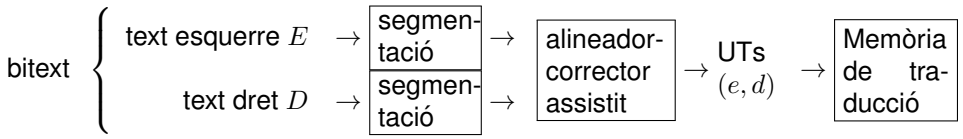


Figura 10.3: Esquema del procés d'*alineament* d'un bitext existent per a alimentar una memòria de traducció.

es puguen reutilitzar posteriorment. Una *memòria de traducció* és una base de dades en la qual cada fitxa (registre) conté una unitat de traducció, i té, per tant, com a mínim dos camps: el text esquerre i el text dret.

L'operació d'alineament de bitextos existents és una de les tasques que es pot realitzar amb l'ajuda d'un programa de memòries de traducció. La figura 10.3 mostra un esquema del procés.

Alineament automàtic

L'alineament automàtic de textos traduïts no és una tasca senzilla; són necessaris coneixements previs sobre les llengües involucrades (per exemple, correspondències entre mots o alineaments prèviament validats per una persona experta[3]), per això, la majoria dels sistemes de memòria de traducció usen mecanismes molt senzills per a segmentar els textos, tant per a alinear bitextos com per a dividir un text esquerre nou en segments per a cercar-los en la base de dades.[4] Per tant, els segments obtinguts no són en general els *ideals* (vegeu més avall). Els mecanismes de segmentació usuals intenten identificar unitats similars a l'oració usant la puntuació i el format com a indicadors, amb la idea que el nombre d'oracions serà bàsicament el mateix en el text dret i l'esquerre (això pot fallar). Però és que, al contrari del que podria semblar, no és gens senzill distingir quan un punt representa el final d'una oració:

```
[...] molt tard.   Després van anar [...]
[...] per CC.OO.   Els d'UGT no van [...]
[...] per CC.OO.   i per UGT [...]
[...] per al Sr.   Martínez [...]
```

En els dos primers casos el punt és el final de l'oració però en els dos últims no. Cal definir clarament les regles de segmentació; els programes

[3] O obtingut mitjançant mètodes estadístics com els descrits en l'apartat 8.5.

[4] És important usar el mateix mètode per a segmentar el bitext abans d'alinear-lo i per a segmentar el nou text esquerre a traduir.

de memòries de traducció solen permetre que la persona usuària modifique o refine aquestes regles. De fet, hi ha un format XML estàndard per a especificar i intercanviar regles de segmentació, anomenat SRX[5] (*segmentation rules exchange*).

Convé, a més, esmentar un altre aspecte addicional important. Els documents de text, a més de caràcters i paraules, contenen informació de *format* que cal considerar en el procés de segmentació i alineament. Quan es tracta de traduir documents de text, la persona usuària vol alinear *text* i en molts casos probablement no necessita veure els codis de format per a validar o corregir un alineament.[6] A més, les unitats de traducció que se n'obtinguen de l'alineament haurien de ser independents del format del document. Els programes més recents resolen això amb *filtres* o *conversors* per als formats de text més freqüents, filtres que tracten d'ocultar al màxim possible a la persona usuària les característiques de format dels documents perquè puguen concentrar-se en les textuals.[7] Els professionals valoren molt la capacitat de gestionar el format de manera senzilla i eficient, ja que la preservació del format de les traduccions és una de les tasques que els fa perdre més temps.

Alineament assistit

És possible que els dos textos no tinguen el mateix nombre d'"oracions" o que l'estratègia per a segmentar-los falle per alguna causa i l'alineament no siga perfecte. La majoria dels programes de memòria de traducció ofereixen a la persona usuària la possibilitat de validar o modificar (unint o dividint segments en el text esquerre o en el text dret) l'alineament automàtic inicial usant una interfície senzilla i intuïtiva abans d'incorporar els segments resultants a la memòria de traducció.

Per saber més sobre l'alineament ideal

Hem descrit un tipus d'alineament possible, basat en unitats fonamentalment equivalent a les oracions, però, quin seria l'alineament òptim d'un bitext? És a dir, quina seria la millor manera de dividir-lo en unitats de traducció? La longitud de les unitats de traducció pot anar des dels mots fins a les oracions senceres. La probabilitat que un fragment esquerre (procedent dels bitextos ja existents) e aparega en un nou text E' és

[5] `http://www.unicode.org/uli/pas/srx/srx20.html`

[6] La situació és diferent quan s'estan traduint els missatges o textos inclosos en *programes* d'ordinador, com a part de les tasques anomenades genèricament de *localització* (adaptació de programes d'ordinador als usuaris d'una regió i idioma concrets); en aquest cas, pot ser molt útil veure tot el programa a més dels textos.

[7] A voltes, qui tradueix ha de gestionar explícitament les etiquetes de format per assegurar una bona traducció: tots els programes ofereixen la possibilitat d'*editar etiquetes* manualment: les etiquetes s'agrupen i simplifiquen per fer més fàcil la faena.

tant més gran com més menut és el fragment. Però si el fragment és massa menut és més probable que la traducció present en la memòria de traducció siga més imprecisa per ambigua (poden aparéixer correspondències múltiples entre les quals s'hauria d'elegir: per exemple a una part esquerra e li poden correspondre dues parts dretes d i d' diferents en unitats de traducció diferents). D'altra banda, si els fragments són massa llargs, és més improbable que siguen ambigus, però és molt menys probable que es repetisquen exactament en textos futurs. Per exemple, el fragment espanyol *las decepciones* es pot correspondre en català amb *les decebes* o *les decepcions* però el fragment *las decepciones sufridas* només pot aparéixer alineat amb *les decepcions patides*. El *fragment ideal* seria el que és suficientment menut per a aparéixer sovint però suficientment complet com per a tenir una traducció constant. És a dir, per una banda, es dóna 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). La grandària ideal és, per tant, un compromís entre la cobertura dels fragments menuts i la precisió dels fragments més grans.

10.2.3 La memòria de traducció com a base de dades

Una vegada alineats els bitextos les unitats de traducció s'organitzen perquè tant el programa com la persona usuària hi puguen accedir eficientment; per exemple, com una base de dades. 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 d'unitats de traducció. Molts programes marquen les unitats de traducció amb un codi que indica la temàtica o la naturalesa o el nom del bitext del qual s'ha extret les unitats de traducció, de manera que la temàtica del nou document servisca per a localitzar les unitats de traducció més adequades en cada cas.

10.3 Traducció amb memòries de traducció

L'organització de les UT en una base de dades permet, a més, *recuperar* de la memòria de traducció, quan s'està traduint un text nou E', els segments esquerres, $e'_1, e'_2, \ldots$ i *construir*, partint dels segments drets corresponents, la traducció desitjada D'.

La memòria de traducció pot contenir unitats de traducció amb segments esquerres idèntics o similars. En cas de trobar un segment idèntic (*concordança exacta*, en anglés *exact match*) per al qual només hi haja una traducció disponible, només cal inserir-ne la traducció directament.

Però, com que això succeeïx poques vegades, ja que els segments són normalment oracions i és difícil que es repetisquen *exactament*, la major part dels sistemes comercials usen estratègies per a no desaprofitar unitats de traducció que continguen parts esquerres *similars* a la nova (les anomenades *concordances parcials*, o, en anglés, *fuzzy matches*).

Alguns programes, si no troben una UT que té la part esquerra idèntica a l'observada en el nou text (e'), però troben una, (e, d), la part esquerra e de la qual es diferencia en un mot o en un cert percentatge dels mots de e', presenten com a *traducció aproximada* la part dreta (d) corresponent a e. Normalment, els sistemes, quan troben un segment similar però no idèntic, destaquen gràficament les diferències (per exemple, amb colors); així, la persona usuària pot fer-hi les modificacions necessàries perquè la traducció resultant siga correcta. Normalment, se sol establir un *llindar* (en anglés *fuzzy match threshold*) de manera que no es presenten propostes per davall d'una certa *puntuació de concordança parcial* mínima (en anglés *fuzzy match score*). Les puntuacions de concordança parcial se solen expressar com a percentatges, on el 0% indica falta total de concordança i el 100% concordança exacta, i el llindar se sol establir per damunt del 60%.

Alguns sistemes, fins i tot, són capaços d'usar les bases de dades lèxiques o terminològiques de la persona usuària per a proposar traduccions per als mots discordants. Per exemple, si a la memòria hi ha la UT (*Connecteu l'ordinador a la impressora.*, *Conecte el ordenador a la impresora*) però el nou text conté la frase *Connecteu l'ordinador a la xarxa*, el programa pot trobar les correspondències (*xarxa*, *red*) i (*impressora*, *impresora*) en una base de dades lèxica i usar-les per a proposar la traducció correcta (*Conecte el ordenador a la red*). D'altres programes usen estratègies pròpies (no descrites) per a construir traduccions usant fragments de parts dretes de més d'una UT. En general, la utilitat d'una memòria de traducció 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*).

Hi ha dues modalitats d'ús de les memòries de traducció:

Interactiva: Qui està traduint rep diverses propostes per a cada nou segment e', entre les quals elegeix la més adequada per a produir la traducció corresponent d'. Aquesta modalitat comporta l'accés per part de qui tradueix a la memòria de traducció.

Pretraducció: Qui està traduint rep com a molt una única proposta per a cada nou segment e', elegida automàticament. En aquesta modalitat, qui tradueix no té accés a la memòria de traducció.

Tant en un cas com en l'altre, poden passar tres coses, tal com es veu en l'exemple de la figura 10.4:

- Que es done una concordança exacta, com en el cas dels segments e'_1 i e'_5 i per tant, només calga comprovar que la proposta de la memòria es pot usar com la seua traducció.
- Que es done una concordança parcial, com en el cas dels segments e'_2 (similar a e_{112}) i e'_3 (similar a e_{47}), i calga posteditar, respectivament,

E'	D' en construcció
$e'_1 = e_{234}$ (100%)	$d'_1 = d_{234}$
$e'_2 \simeq e_{112}$ (87%)	posteditar $d_{112} \rightarrow d'_2$
$e'_3 \simeq e_{47}$ (93%)	posteditar $d_{47} \rightarrow d'_3$
e'_4	generar d_4 des de zero
$e'_5 = e_{51}$ (100%)	$d'_5 = d_{51}$
...	...

Figura 10.4: Traducció d'un nou bitext E' i els tipus de situacions de concordança que s'hi poden donar: concordança exacta per a e'_1 i e'_5, concordança parcial per a e'_2 i e'_3, i absència de concordança raonable per a e'_4.

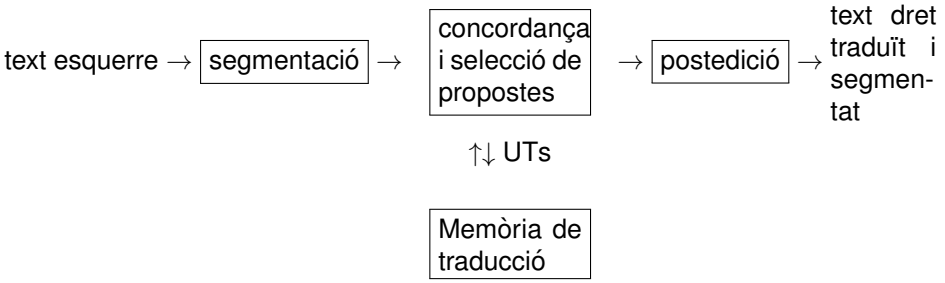


Figura 10.5: Esquema del procés de traducció d'un nou text esquerre usant una memòria de traducció.

les propostes d_{112} i d_{47} de la memòria de traducció, per a produir d'_2 i d'_3.

- Que no es done cap concordança raonable, com en el cas del segment e'_4, i calga generar la traducció d'_4 des de zero.

Tant en el cas interactiu com en el de la pretraducció, qui tradueix ha d'acabar la traducció. El procés es mostra en la figura 10.5.

10.3.1 Ampliació de la memòria

Una vegada feta la pretraducció o la selecció interactiva de propostes de traducció para cada un dels segments del text esquerre E', el programa les mostra alineades amb els segments del text esquerre perquè la persona usuària les puga posteditar i produir el text dret correcte D'. Les noves unitats de traducció (e', d') que s'hagen produït en el procés es poden enviar

a la memòria de traducció. La repetició del cicle pretraducció–correcció–enviament de noves UT a la memòria enriqueix la memòria de traducció i permet traduir cada vegada amb menys esforç, ja que la cobertura de la memòria augmenta.

10.4 Productes

En l'actualitat, hi ha molts programes de traducció assistida per ordinador basats en memòries de traducció disponibles comercialment (*SDL Trados* de SDL International, *Déjà Vu* d'Atril, *IBM Translation Manager*, *Transit* de Star, etc., per nomenar algunes de les més conegudes). També hi ha una alternativa interessant de codi obert i distribució gratuïta, anomenada OmegaT.[8] Aquests programes de traducció assistida són molt populars entre les persones que es dediquen professionalment a la traducció, molt més populars que els programes de traducció automàtica. Això pot ser, d'una banda, perquè els professionals tenen així la impressió que el programa no tradueix i els relega a un mer paper de correctors, sinó que *organitza* i fa més eficient el treball de traducció professional i, d'altra banda, perquè les memòries de traducció *conserven l'estil* i *les decisions terminològiques* de traduccions anteriors, que poden variar d'un equip a altre, mentre que els sistemes de traducció automàtica solen basar-se en seleccions terminològiques i d'estil de propòsit general, encara que siga dins d'una temàtica concreta.

S'ha de tenir en compte que la traducció automàtica pot constituir una alternativa acceptable en aquells segments on la memòria de traducció no pot fer una proposta raonable, i, de fet, molts programes de traducció assistida per ordinador combinen l'accés a memòries de traducció amb l'accés a la traducció automàtica. La puntuació de concordança parcial per davall de la qual propostes de la memòria de traducció comencen a ser menys útils que la traducció automàtica depén de la cobertura de la memòria de traducció per al tipus de text, de la llengua origen i la llengua meta, i, per descomptat, del sistema concret de traducció automàtica: per a llengües molt similars com ara l'espanyol i el català, podria passar que només les concordances millors que, diguem-ne, el 95% foren acceptables, mentre que si les llengües són més distants, com l'espanyol i el turc, podria passar que la traducció automàtica només fóra útil per a puntuacions de concordança molt baixes.

Hi ha tecnologies de traducció automàtica que s'assemblen molt al funcionament de les memòries de traducció, com ara la traducció automàtica estadística (vegeu l'apartat 8.5). Un exemple clàssic és l'edició bilingüe (espanyol–català) d'*El Periódico de Catalunya* (vegeu l'epígraf A.2.3), que es prepara diàriament amb un mètode completament automàtic que funciona en molts aspectes de manera similar a una memòria de traducció.

[8] `http://www.omegat.org`

10.5 Intercanvi de memòries de traducció

10.5.1 El format d'intercanvi TMX

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. El format TMX segueix les especificacions XML (vegeu l'apartat 4.4); és a dir, les memòries TMX són un tipus de document XML, definit, per tant, per una DTD concreta.[9]

La figura 10.6 mostra part d'una memòria de traducció en el format TMX. S'hi mostren només dues unitats de traducció (`tu`) dins del cos (`body`), cada una amb el seu identificador únic (`tuid`). Cada unitat de traducció conté dues variants (`tuv`), cada una en una llengua (`xml:lang="..."`), a més d'un element `prop` que conté la clau (`key`) per a cercar la unitat de traducció en la base de dades. Abans del cos, l'element arrel `tmx` conté una capçalera (`header`) amb informació sobre la creació i les característiques de la memòria.

10.5.2 Altres problemes

Però, fins i tot quan ja s'ha resolt aquest problema tècnic, 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 l'explotació de grans memòries de traducció distribuïdes (en les diverses màquines d'una xarxa) està lluny de ser trivial. Per exemple, en el cas de l'espanyol i el català, una gran memòria de traducció alimentada amb les traduccions fetes només en l'àmbit de les administracions autonòmiques 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.

[9] `http://www.ttt.org/oscarstandards/tmx/tmx14b.html`

```
<?xml version='1.0' encoding='ISO-8859-1' ?>
<!DOCTYPE tmx SYSTEM 'tmx13.dtd'>
<tmx version='1.3'>
 <header creationtool='Waikoloa'
                  creationtoolversion='1.00'
                  datatype='plaintext'
                  segtype='paragraph'
                  adminlang='EN-US'
                  srclang='EN-US'
                  o-tmf='okLiteTM'>
 </header>
 <body>
 <!-- ... -->
   <tu tuid='511'>
   <prop type='tmkey'>a thesaurus error occurred word
   is ending the current session</prop>
   <tuv xml:lang='EN-US'>
    <seg>A thesaurus error occurred. Word is ending
    the current session.</seg>
   </tuv>
   <tuv xml:lang='FR-FR'>
    <seg>Une erreur s'est produite pendant l'exécution
    du dictionnaire des synonymes. Word met fin à la
    session en cours.</seg>
   </tuv>
  </tu>
  <tu tuid='512'>
   <prop type='tmkey'>a thumbnail preview is not
    available for this file</prop>
   <tuv xml:lang='EN-US'>
    <seg>A thumbnail preview is not available for
         this file.</seg>
   </tuv>
   <tuv xml:lang='FR-FR'>
    <seg>Il n'y a pas d'aperçu disponible pour
         cette image.</seg>
   </tuv>
  </tu>
  <!-- ... -->
</body>
</tmx>
```

Figura 10.6: Exemple de memòria de traducció en TMX; s'hi mostren només dues unitats de traducció.

10.6 Qüestions i exercicis

1. Indica quina d'aquestes afirmacions és certa:

   (a) Les memòries de traducció són bàsicament sistemes de traducció directa i, per tant, la unitat bàsica de traducció que usen és el mot.

   (b) Les memòries de traducció usen informació sobre les categories lèxiques dels mots per a decidir els alineaments.

   (c) Per a no haver de traduir un text nou des de zero amb un sistema de ajuda a la traducció basat en memòries de traducció, és necessari que hi haja textos originals i traduïts que hagen estat alineats.

2. Una memòria de traducció es pot veure com una base de dades ...

   (a) ... on cada registre és una llengua i cada camp una unitat de traducció.

   (b) ... on cada registre és una oració i cada camp un mot.

   (c) ... on cada registre és una unitat de traducció i la variant en cada llengua es guarda en un camp diferent.

3. Quina característica dels textos que s'han de traduir fa que l'ús de memòries de traducció siga la solució adequada?

   (a) El fet que els textos estiguen escrits amb un lèxic monosèmic, és a dir, precís i no gens ambigu.

   (b) La repetitivitat.

   (c) La similitud entre les llengües origen i meta.

4. Quant al funcionament, a quins sistemes de traducció automàtica s'assemblen més les memòries de traducció?

   (a) Als sistemes de traducció automàtica directes o mot per mot.

   (b) Als sistemes de traducció automàtica per interlingua.

   (c) Als sistemes de traducció automàtica per transferència.

5. Les memòries de traducció comercials actuals segmenten i alineen els textos ...

   (a) ... fent un balanç òptim entre cobertura i precisió.

   (b) ... usant la informació sintàctica com a pista.

   (c) ... usant regles que analitzen la puntuació i el format.

6. Indica quina d'aquestes afirmacions és falsa:

(a) Els resultats produïts per una memòria de traducció no necessiten revisió, ja que es basen en traduccions correctes realitzades anteriorment per professionals.

(b) Les memòries de traducció poden organitzar els segments i les traduccions corresponents en bases de dades similars a les bases de dades lèxiques o terminològiques.

(c) Una memòria de traducció és un sistema de traducció directa amb equivalències entre segments de text observades en textos anteriorment traduïts.

7. Les memòries de traducció usen els signes de puntuació per a segmentar les oracions abans d'alinear els textos. En particular, l'aparició d'un punt ("`.`") és moltes vegades un bon indicador del final d'una oració, però no sempre. Quan no? Doneu almenys *quatre* excepcions diferents a aquesta regla i descrigueu, en cada cas, una regla senzilla que permeta decidir amb seguretat raonable quan ens trobem en cada una d'aquestes excepcions, usant el mínim possible d'anàlisi lingüística del text.

8. La majoria de les memòries de traducció comercials divideixen els bitextos en unitats de traducció ...

   (a) ... aproximadament equivalents a una oració, usant regles senzilles i relativament independents de la llengua per a segmentar (dividir) cada text en oracions.

   (b) ... en mots i petites unitats multimot (entre dos i quatre mots) de gran repetitivitat.

   (c) ... equivalents a una oració, usant una anàlisi lingüística detallada del text per a determinar l'extensió de cada oració.

9. Que hem de fer amb els bitextos existents per a poder reutilitzar la informació que contenen per a fer noves traduccions amb una memòria de traducció?

   (a) Passar-los a XML.

   (b) Segmentar cada un dels dos textos en oracions.

   (c) Segmentar els dos textos i alinear-los.

10. Estem traduint un segment en un programa de traducció assistida (com ara OmegaT) i el programa ens dóna una sèrie de propostes que vénen de la memòria de traducció. Quin d'aquests indicadors indica millor l'esforç que haurem de fer per acabar de traduir el segment?

    (a) El percentatge de concordança parcial de la millor proposta.

(b) El nombre de propostes.

(c) La longitud en paraules de la millor proposta (com més llarga, millor).

11. La majoria dels programes de traducció assistida basats en memòries de traducció no donen una de les tres informacions següents:

(a) El percentatge de coincidència entre el nou segment a traduir i el segment origen de la unitat de traducció proposada.

(b) Els mots que cal canviar en el segment meta de la unitat de traducció proposada.

(c) Els mots del segment origen de la unitat de traducció proposada que es diferencien de les del nou segment a traduir.

12. Quina d'aquestes característiques d'un treball de traducció el fa més tractable amb memòries de traducció?

(a) La repetitivitat dels textos.

(b) La proximitat entre les llengües origen i meta.

(c) Que els textos origen i meta estiguen codificats en ISO-8859-1 (*Latin-1*).

13. Per poder explotar les traduccions presents en un bitext (o text paral·lel) en un programa d'ajuda a la traducció basat en memòries de traducció ...

(a) ...hi ha prou amb incloure els bitexts com a font d'unitats de traducció per al seu ús.

(b) ...és necessari alinear els segments prèviament; aquesta operació es pot fer automàticament, encara que pot requerir supervisió.

(c) ...primer cal convertir el bitext al format binari que use el programa d'ajuda a la traducció que estiguem utilitzant.

14. En general, l'ús d'un programa gestor de memòries de traducció fa que el procés de traducció siga més eficient quan ...

(a) ...els textos a traduir són curts.

(b) ...els textos a traduir són molt repetitius.

(c) ...la llengua origen i meta de la traducció pertanyen a la mateixa família de llengües.

15. Com podem generar una memòria de traducció a partir d'un text i de la seua traducció?

(a) Cal segmentar i alinear els segments.

(b) Cal segmentar el text, però l'alineament no és necessari perquè ja sabem que un text és traducció de l'altre.

(c) No es pot, les memòries de traducció es creen en anar traduint els segments.

16. Podem usar una memòria de traducció amb segments en anglés i espanyol per traduir de l'alemany a l'espanyol?

(a) Sí, si tracten del mateix tema.

(b) Sí, si la longitud dels segments és la mateixa.

(c) No.

10.7 Solucions

1. (c)
2. (c)
3. (b)
4. (a)
5. (c)
6. (a)
7. El punt apareix en moltes construccions que no indiquen el final de una oració. Heus ací uns exemples i com detectar-los per a no segmentar.

(a) Punts de milers (1.259), milions (1.032.200), decimals saxons (3.4), números de telèfon francesos (01.10.23.87.49), dates (20.01.1999), etc. *Solució:* Si detectem [xifra] "." [xifra] (sense blancs), no segmentem.

(b) Punts en mig de sigles: CC(.)OO., EE(.)UU., etc. *Solució:* si detectem [majúscula] "." [majúscula] (sense blancs), no segmentem.

(c) Punts al final d'abreviatura de cortesia: Sr(.), Dra(.), Excm(.), etc. o d'altres que precedeixen noms propis (Avda., Pça.) o números (Tel.). *Estratègia de solució:* si detectem "." [blanc] [minúscula], no segmentem; si detectem "." [blanc] [número], no segmentem; si detectem "." [blanc] [majúscula], què fem?: depén de l'abreviatura (la decisió és impracticable sense vocabularis de noms propis).

(d) Punts al final de sigles: CC.OO(.), O.N.U(.). *Solució:* si detectem [majúscules] "." [majúscules] ".", no segmentar.

(e) Punts en URIs i adreces de correu electrònic. *Solució:* la millor seria una regla general (patró o expressió regular) que detectara aquestes entitats i evitara segmentar-les. Possibles escapatòries: si detectem [minúscula] "." [minúscula] (sense blancs), no segmentem.

(f) Punts suspensius ("..."). *Solució:* no segmentar mai "..." si es troba.

8. (a)
9. (c)
10. (a)
11. (b)
12. (a)
13. (b)
14. (b)
15. (a)
16. (c)

Apèndix A

Traducció automàtica espanyol–català

La inclusió d'aquest apèndix en el llibre té tres objectius:

1. Estudiar amb una miqueta més de detall els problemes que planteja la traducció automàtica entre dues llengües emparentades. Podríeu pensar que, sent l'espanyol i el català tan similars sintàcticament, els problemes serien poc importants: el capítol intenta convéncer-vos que les coses no són tan senzilles com podrien semblar a primera vista.

2. Il·lustrar, amb un parell de llengües concret, alguns dels conceptes tractats en els capítols anteriors.

3. Donar una breu notícia de les experiències de traducció automàtica espanyol–català existents.

A.1 Problemàtica de la traducció automàtica espanyol–català

A.1.1 Introducció

Les aplicacions potencialment més interessants de la TA espanyol–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 territoris de parla catalana, on la presència del català és encara minoritària. Quan la llengua original dels documents és l'espanyol, es podria usar un sistema de TA per a generar esborranys de documents catalans (o, fins i tot, documents pràcticament

correctes si els documents espanyols estan escrits en un llenguatge controlat).

En el cas concret de l'espanyol 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 (mireu la p. 6.5.1).

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 de l'espanyol 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 espanyols pels seus equivalents catalans. De fet, el model de traducció automàtica *mot per mot* (definit en la pàg. 150, i que no s'ha de confondre amb el que se sol anomenar *traducció literal*) és el model de referència que usarem en aquest capítol: els tres grups de *problemes* que es presenten en aquest 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.

A.1.2 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: per exemple, l'espanyol 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*, *pels*, i *del*, sempre queda alguna indicació d'aquesta unió, com ara un apòstrof o un guionet. Per exemple, en espanyol, 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 8.3.1). Per sort, aquests problemes es poden resoldre de manera senzilla usant analitzadors morfològics com els que es descriuen en l'epígraf 8.3.2.

A.1.3 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 7.2.1). L'espanyol —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 invo-

lucra 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 (algunes particularment freqüents, com ara *para*, preposició i verb; *una*, determinant i verb, i *como*, radverbi relatiu, preposició i verb).

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

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. femṡing.).

   (b) En *-as*:
      - 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*:

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

- 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)

A.1.4 Divergències de traducció

Imaginem que hem pogut segmentar el text espanyol 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 espanyols. Vegem quins són alguns dels problemes:

Concordança de gènere i nombre: De vegades el gènere i el nombre d'un mot varien de l'espanyol 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 (opcional) de la concordança del participi, inexistent en espanyol en situacions com ara *todavía no la hemos estudiado con profundidad* → *encara no l'hem estudiada amb profunditat*.

L'article neutre: L'espanyol posseeix l'anomenat *article neutre*, que no té correspondència en català estàndard (*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 l'espanyol 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 8.3.3).

Els pronoms febles: Els principals problemes es troben en la traducció de *lo*, ja que pot correspondre en català a alguna forma del pronom 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 espanyoles *se la*, *se lo*, etc. pot correspondre de vegades a alguna forma de *li* o *els*, i en el fet que l'espanyol 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 espanyol i català: les preposicions espanyoles 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.

A.2 Experiències de TA espanyol–català

En aquesta secció es descriuen breument quatre experiències de traducció automàtica de l'espanyol al català (SALT, el traductor espanyol–català de Lucy Software, els traductors Automatictrans i el de *El Periòdico de Catalunya*, interNOSTRUM), i una d'elles (Apertium) amb més detall.

A.2.1 SALT, de la Generalitat Valenciana

El programa SALT (la versió actual és la 4.0) porta el nom de l'antic*Servei d'Assessorament Lingüístic i Traducció* de la Conselleria de Cultura, Educació i Ciència (ara Conselleria d'Educació, Investigació, Cultura i Esport) de la Generalitat Valenciana; es tracta d'un programa que s'executa en els sistemes operatius Windows, GNU/Linux i MacOS. El desenvolupament del programa el va iniciar a finals dels noranta per part d'un equip de programadors dirigit per Rafael Pinter sota la direcció lingüística de Josep Lacreu, en aquell moment responsable d'aquest servei. Inicialment, la disponibilitat del programa va ser més aviat reduïda i el seu llançament es va retardar per les discussions quant a l'estàndard de valencià que havia de produir; actualment es pot descarregar gratuïtament de diversos servidors d'Internet[2] i també el distribueixen els serveis de normalització lingüística d'algunes universitats. SALT 4.0 s'executa com una extensió dels processadors de textos LibreOffice i OpenOffice.org, tradueix textos en espanyol a la variant valenciana del català i està concebut també com una ajuda a les persones que volen començar a generar documents en valencià (entre altres eines, inclou diccionaris i guies de consulta completíssimes). L'Acadèmia Valenciana de la Llengua va declarar *oficials* "els continguts" del programa SALT 2 (acord de 20 de maig del 2002). [3]

A.2.2 El traductor espanyol–català de Lucy Software

El sistema de traducció automàtica espanyol–català de Lucy Software, originalment desenvolupat per l'empresa Incyta de Cornellà en col·laboració amb la Universitat Autònoma de Barcelona és un sistema de transferència sintàctica estàndard (vegeu l'apartat 8.3.3), hereu del sistema METAL de l'empresa Siemens. El seu desenvolupament va anar passant d'una empresa a altra: en l'actualitat el desenvolupa la multinacional Sail Labs i el distribueix l'empresa Incyta, S.L.. El programa es pot usar en Internet[5] i els resultats són de gran qualitat.

[2] com ara `http://www.ceice.gva.es/polin/val/salt/apolin_salt4.htm` i `https://www.softcatala.org/wiki/Rebost:Salt`

[3] L'any 2000, l'empresa Autotrad de València va llançar el programa Ara, llançat. El gerent de l'empresa era Rafael Pinter, responsable informàtic de SALT. Ara era bàsicament una versió bastant millorada de la primera versió de SALT, amb una aparença molt similar però amb algunes diferències: p.e., produïa textos en català oriental estàndard, podia dialogar amb la persona usuària en espanyol i en català, i permetia programar tasques de traducció perquè s'executaven sense necessitat que la persona usuària les atenguera. El cost (l'any 2004) era de 45 euros per llicència. El lloc web de l'empresa[4] no sembla funcionar correctament, i és possible que el programa ja s'estiga comercialitzant.

[5] `http://www.lucysoftware.com/catala/traduccio-automatica/kwik-translator-/`

A.2.3 El traductor d'*El Periódico de Catalunya* i AutomaticTrans

Una experiència interessant (Fité 2006) de traducció espanyol–català per a la disseminació és l'edició bilingüe del diari *El Periódico de Catalunya*;[6] el text original —en espanyol la major part de les vegades— es tradueix usant un sistema de traducció automàtica basat en corpus combinat en tècniques similars a les *memòries de traducció* (vegeu el capítol 10) i després és revisat pels redactors de tancament del mateix periòdic abans de ser publicat.

Un programa similar (i, segons les nostres notícies, d'origen comú) a l'usat per *El Periódico de Catalunya* s'anomenava abans AutomaticTrans i ara probablement el comercialitza l'empresaAT Language Solutions[7]

A.2.4 interNOSTRUM

Un equip d'investigadors de la Universitat d'Alacant, finançat per l'extinta Caja de Ahorros del Mediterráneo i per la mateixa Universitat, va desenvolupar entre 1998 i 2006 sota la direcció d'un dels autors d'aquest llibre un sistema de traducció automàtica espanyol–català anomenat interNOSTRUM (Canals-Marote et al. 2001a,b). L'objectiu del projecte era desenvolupar un sistema de traducció automàtica de l'espanyol a les variants estàndards del català i el sistema invers corresponent. Durant l'últim decenni, ha estat un dels sistemes de traducció automàtica més usats en Internet.

La versió actual d'interNOSTRUM (que va estar accessible de forma gratuïta a través de l'URL `http://www.internostrum.com` i que el gener de 2016 encara estava accessible a través de l'adreça `http://torsimany.ua.es/index.php`) genera, quasi instantàniament, esborranys de traduccions al català llestes per a ser corregides (posteditades).

interNOSTRUM tradueix textos en formats ANSI, HTML i RTF de l'espanyol al català oriental i al revés 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).

El traductor estava escrit per executar-se sobre el sistema operatiu GNU/-Linux i és encara accessible, com ja s'ha dit, a través d'un servidor d'Internet;[8] i és un sistema de transferència morfològica avançada com el descrit en l'apartat 8.3.1; el disseny del sistema és molt similar al del sistema Apertium que es descriu més avall en la secció A.2.5, del qual és precursor.

[6] Disponible per Internet: `http://www.elperiodico.es`.

[7] `https://www.at-languagesolutions.com/`

[8] Tot i que ja no es manté: també hi ha disponible una versió per a servidors basats en el sistema operatiu Windows.

A.2.5 Apertium

Apertium[9] (Forcada et al. 2011), iniciat a la Universitat d'Alacant el 2005, és una plataforma de programari lliure o de codi font obert que permet construir sistemes de traducció automàtica de transferència morfològica avançada (com els de l'apartat 8.3.1); això vol dir que es pot descarregar i copiar lliurement però també que programadors i lingüistes poden modificar el programari, els diccionaris, les regles, etc. i distribuir versions modificades, ja que a més de l'executable del programari que necessitem per usar-lo, se'n distribueix el codi font, és a dir, la forma del programari que permet als experts modificar-lo.

El primer sistema que es va construir sobre la plataforma Apertium va ser el sistema espanyol–català (en l'actualitat hi ha disponibles en Apertium més de 40 sistemes de traducció automàtica diferents).

Apertium és pot usar gratuïtament en línia a través de moltes webs[10] però també és pot instal·lar en localment. La versió completa —per exemple per a muntar un servidor per a un entorn de producció— s'instal·la sobre ordinadors amb sistema operatiu GNU/Linux, però hi ha moltes altres versions que funcionen sense necessitat de connexió a Internet, com ara:

- Una aplicació per al sistema operatiu Android, *Apertium offline translator*;[11]
- Una aplicació de sobretaula, *apertium-caffeine* [12] per a GNU/Linux, Windows o MacOS (requereix que s'hi haja instal·lat Java);
- Un extensió per al programa de traducció assistida OmegaT, anomenada *apertium-omegat*.[13]

El disseny d'Apertium Com s'ha dit més amunt, els sistemes de traducció automàtica basats en l'arquitectura Apertium són tots sistemes de transferència morfològica avançada. L'arquitectura d'un sistema de traducció concret basat en Apertium és flexible: depenent de la llengua origen i la llengua meta, es poden seleccionar mòduls diferents. La figura A.1 representa la configuració més comuna d'un sistema de traducció automàtica basat en Apertium: la construcció del text d'entrada es va fent etapa per

[9] www.apertium.org

[10] Per exemple: http://www.apertium.org, http://apertium.ua.es, http://apertium.uoc.edu, http://politraductor.upv.es i http://aplica.prompsit.com.

[11] https://play.google.com/store/apps/details?id=org.apertium.android

[12] http://wiki.apertium.org/wiki/Apertium-Caffeine

[13] http://wiki.apertium.org/wiki/Apertium-OmegaT

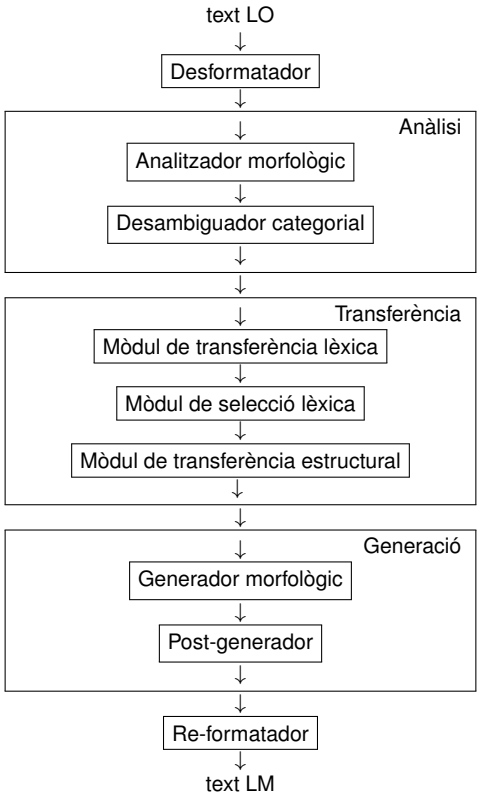


Figura A.1: Arquitectura més comuna dels sistemes de traducció automàtica basats en Apertium.

etapa, com en la cadena de muntatge d'una fàbrica d'automòbils. El sistema espanyol–català segueix l'arquitectura de la figura A.1 però la versió disponible en 2016 no té (encara) mòdul de selecció lèxica.

Apertium separa efectivament el *motor de traducció* (que consisteix en mòduls genèrics, comuns a totes les parelles llengua origen–llengua meta) i les *dades lingüístiques* específiques d'una parella llengua origen–llengua meta, tal com es discuteix en la 152. Això permet que les persones expertes que desenvolupen dades lingüístiques (diccionaris, regles) per a un sistema determinat no hagen de preocupar-se de com està programat el motor de traducció.

Els següents paràgrafs descriuen alguns dels mòduls amb més detall.

Subprogrames basats en tècniques d'estats finits: Els mòduls d'*anàlisi morfològica*, *transferència lèxica*, *generació morfològica* i *postgeneració* estan basats en *transductors d'estats finits*, similars als descrits en el quadre "per saber més" de l'epígraf 8.3.2. Aquesta tecnologia permet velocitats de processament de l'ordre de 10.000 mots per segon en equips estàndards, velocitats que pràcticament no depenen de la grandària dels diccionaris. Els *transductors d'estats finits* usats en Apertium 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.

Analitzador morfològic: L'analitzador morfològic 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.

Transferència lèxica (transferència lèxica): El subprograma de consulta del diccionari bilingüe 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 o formes lèxiques corresponents 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 A, 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 lèxiques 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 espanyol, 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 Apertium 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 mòdul de desambiguació lèxica categorial: Aquest programa s'encarrega de decidir, quan l'analitzador morfològic entrega, per a un mot homògraf, més d'una forma lèxica, quina és la forma lèxica més adequada en el context. Els desambiguadors lèxics categorials combinen (vegeu l'apartat 7.2.4) regles de base lingüística que permeten eliminar algunes formes lèxiques i models estadístics, entrenats sobre un corpus de referència, que assignen 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.

El mòdul de transferència estructural: Malgrat la gran semblança entre l'espanyol i el català, hi ha divergències gramaticals considerables (vegeu el capítol A):

- 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, molt similars a les que es discuteixen en l'apartat 8.3.1: la solució 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 mòdul en formen el *catàleg* de patrons. El funcionament del subprograma es basa en un esquema patró–acció:

- Llegeix el text (analitzat i ja 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'usen variables d'*estat*, una espècie de *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 pocs milers de mots per segon).

A.3 Qüestions i exercicis

Aquests exercicis poden servir per a repassar els conceptes tractats en aquest apèndix.

1. Quina d'aquestes tres tasques és més difícil en un sistema de traducció automàtica espanyol–català?
   (a) Decidir la traducció del pronom espanyol *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) de l'espanyol:
   (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* espanyol. En traducció automàtica de l'espanyol 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 espanyol.

A.4 Solucions

1. (a)
2. (a)
3. (b)

Bibliografia

AECMA (2007). AECMA Simplified English. http://www.simplifiedenglish-aecma.org/Simplified_English.htm.

Ageeva, E., Forcada, M., Tyers, F., Pérez-Ortiz, i J.A. (2015). Evaluating machine translation for assimilation via a gap-filling task. En *Proceedings of EAMT 2015, The Eigtheenth Annual Conference of the European Association for Machine Translation (Antalya, May 11-13, 2015)*, pages 137–144.

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. (2003). En Somers, H., editor, *Computers and Translation: A translator's guide*, chapter Why translation is difficult for computers, pages 119–142. John Benjamins, Amsterdam i Philadelphia.

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.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., i Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Canals-Marote, R., Esteve-Guillen, A., Garrido-Alenda, A., Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Perez-Antón, P., i Forcada, M. (2001a). El sistema de traducción automática castellano-catalán interNOSTRUM. *Procesamiento del Lenguaje Natural*, 27:151–156. XVII Congreso de la Sociedad Española de Procesamiento del Lenguaje Natural, Jaén, España, 12–14.09.2001.

Canals-Marote, R., Esteve-Guillen, A., Garrido-Alenda, A., Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Perez-Antón, P., i Forcada, M. (2001b). The Spanish-Catalan machine translation system interNOSTRUM. En *Proceedings of MT Summit VIII: Machine Translation in the Information Age*, pages 73–76. Santiago de Compostela, Spain, 18–22.09.2001.

Carl, M., Iomdin, L. L., Pease, C., i Streiter, O. (2001). Towards a dynamic linkage of example-based and rule-based machine translation. *Machine Translation*, 15(3):223–257.

Chomsky, N. (1996). *The minimalist program*. MIT Press, Cambridge, Massachusetts.

Doherty, S., Kenny, D., i Way, A. (2012). A user-based usability assessment of raw machine translated technical instructions. En *The 10th Biennial Conference of the Association for Machine Translations in the Americas (AMTA 2012)*, San Diego, California, USA. 28 oct. – 1 nov. 2012.

Don, J., Kerstens, J., Ruys, E., i Zwarts, J. (1996). Lexicon of linguistics. `http://www-uilots.let.ruu.nl/~Hans.Leidekker/lexicon/ll.html`.

Douglas, S. i Hurst, M. (1996). Controlled language support for perkins approved clear english (pace). En *Proceedings of the First International Workshop on Controlled Language Applications*, volume 93, page 105. Citeseer.

Fité, R. (2006). El periódico, una experiencia en traducció automàtica. *Tradumàtica*.

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)*.

Forcada, M. i Pérez-Ortiz, J. (2009). *Informática Aplicada a la Traducció: notes de classe amb exercicis i problemes resolts*. Alacant.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., i Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Hovy, E. (1993). How MT works. *Byte*, (gener):167–176.

Huijsen, W.-O. (1998). Controlled language–an introduction. En *Proceedings of CLAW*, volume 98, pages 1–15.

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.

Isabelle, P., Dymetman, M., Foster, G., Jutras, J., Macklovitch, E., Perrault, F., Ren, X., i Simard., M. (1993). Translation analysis and translation automation. En *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research:, vol. 2, ≪Distributed computing≫*, pages 1133–1147. IBM Press.

Jakobson, R. (1966). On the linguistic aspects of translation. En Brower, R., editor, *On translation*. Oxford Univ. Press, Oxford.

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., i Emonts, M. (2007). ILR-based MT comprehension test with multi-level questions. En *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.

Kohl, J. R. (2008). *The Global English Style Guide: Writing Clear, Translatable Documentation for a Global Market*. SAS Institute.

Krauwer, S. (1993). Evaluation of MT systems: a programmatic view. *Machine Translation*, 8.

Lappin, S. i Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

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.

Masterman, M. (1967). *Machine Translation*, chapter Mechanical pidgin translation: An estimate of the research value of "word-for-word" translation into a pidgin language, rather than into the full normal form of an output language. North Holland.

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.

O'Regan, J. i Forcada, M. L. (2013). Peeking through the language barrier: the development of a free/open-source gisting system for basque to english based on `apertium.org`. *Procesamiento del Lenguaje Natural*, 51:15–22.

O'Brien, S. (2003). Controlling controlled english. an analysis of several controlled language rule sets. *Proceedings of EAMT-CLAW*, 3:105–114.

Radford, A., Atkinson, M., Britain, D., Clahsen, H., i Spencer, A. (1999). *Linguistics: an introduction*. Cambridge Univ. Press, Cambridge.

Radford, A., Atkinson, M., Britain, D., Clahsen, H., i Spencer, A. (2009). *Linguistics: an introduction, 2nd. ed.* Cambridge Univ. Press, Cambridge.

Ramos, J. R. (1992). *Introducció a la sintaxi*. Tàndem, València.

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.

Schwitten, R. (2007). Controlled natural languages. `http://www.ics.mq.edu.au/ rolfs/controlled-natural-languages/`.

Somers, H. i Rutzler, C. (1996). Machine translation. En Owens, R., editor, *The translator's handbook*. Aslib, Londres, 3a. edició.

Tellier, I. (2000). Semantic-driven emergence of syntax: the principle of compositionality upside-down. En *Proc. 3rd Conference on the The Evolution of Language*, pages 220–224, Paris.

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`).