

Curs urgent de traducció automàtica

Mikel L. Forcada

Departament de Llenguatges i Sistemes Informàtics

Universitat d'Alacant

E-03071 Alacant, Spain

Traducció Automàtica: Fonaments i Aplicacions

Universitat d'Alacant, 2004

Índex

1. Què és la traducció automàtica (TA)? Aplicacions
2. Formats de text
3. Com funciona la TA?
4. Per què és difícil la TA?
5. Avaluació de la traducció automàtica
6. TA de pàgines web
7. Memòries de traducció

Què és la traducció automàtica (TA)? /1

La traducció, ...

... mitjançant un sistema informàtic ...

... (ordinador(s) + programes) ...

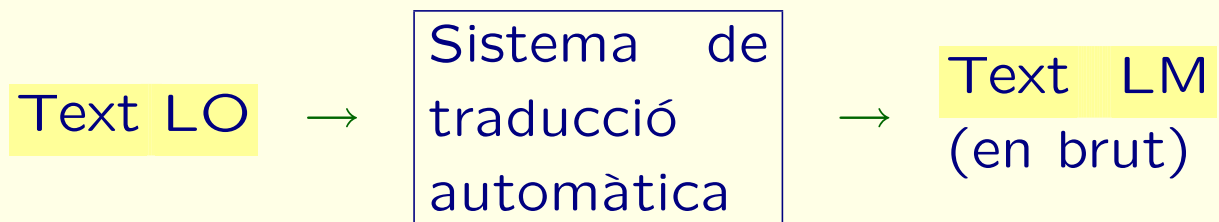
... de textos informatitzats en la llengua origen (LO)...

... a textos informatitzats en la llengua meta (LM).

[Atenció al format dels textos!!]

Què és la traducció automàtica (TA)? /2

Esquemàticament:



Aplicacions de la TA /1

Dos grans grups:

- **Assimilació:** traducció efímera, idealment instantània, per a la revisió o la comprensió de documents en una altra llengua. P.e., navegació per internet, xat (*chat*), etc.
- **Disseminació:** traducció permanent, idealment amb pocs errors, per a la publicació. P.e., producció d'esborranys per a *posteditar*.

Aplicacions de la TA /2

Preedició i postedició: els professionals col·laboren amb el sistema de TA en aplicacions de disseminació:

- **Preedició:** preparació del text per a evitar lèxic o construccions que donen problemes de traducció amb un sistema de traducció automàtica.
- **Postedició:** correcció del text traduït en brut per a fer-lo adequat al propòsit previst.

Conèixer bé com funciona el sistema de TA ajuda molt en ambdues tasques.

Aplicacions de la TA /3

Alternativa a la preedició: llenguatge controlat.

- Els autors escriuen ja pensant en el tractament automatitzat del text.
- S'eviten lèxic i construccions problemàtiques.
- Es minimitza la postedició.
- Consistència d'estil, comprensibilitat, mantenibilitat.
- Però els autors l'han de conèixer i aplicar!
- Se'ls pot ajudar amb eines informàtiques.

Aplicacions de la TA /4

La postedició és convenient quan

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

Perquè siga eficient:

- cal ser competent en la llengua meta → generar un text genuí a partir del text en brut
- cal conèixer el sistema de TA → reconèixer l'origen dels errors, predir-ne el comportament

Formats de text /1

Un text informatitzat és, com qualsevol porció de dades informatitzada, una seqüència de bits, és a dir, d'uns i zeros: 000101010010100111101001010010....

Els bits van normalment en grups de 8 (bytes o octets). Amb 8 bits es poden fer $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^8 = 256$ combinacions: 00000000 (0), 00000001 (1), 00000010 (2), ..., 11111111 (255).

Hi ha moltes maneres d'organitzar els octets per a emmagatzemar textos. Molts problemes provenen de discrepàncies quant a la manera de fer-ho.

Formats de text /2

Dos aspectes importants: **codificació** i **format** propiament dit.

- **Codificació:** Assignació d'un codi (una seqüència d'un o més octets concreta) a cada caràcter possible de la llengua corresponent (per exemple: "a" → 01100001 (97); "?" → "00111111" (63), etc.)

Formats de text /3

- **Format propiament dit:** Els textos, a més de caràcters, contenen informació de format.

És necessària l'assignació de codis per a regular altres característiques del text:

- Per a codificar l'aparença visual o de presentació, per exemple, “inici cursives”, “final negretes”, “lletra de 16 punts”), o
- Per a codificar l'estructura (és a dir, l'organització del contingut, per exemple, “títol de secció”, “llista numerada”, “nota a peu de pàgina”, “fila d'una taula”, etc.).

Formats de text /4

Codificacions d'1 octet (“unibyte”):

- **ASCII:** Assigna codis de 7 bits, del 0000000 (0) al 1111111 (127), (sobra un bit de l'octet) als caràcters anglesos (sense accents, etc.)
- **ANSI o ISO:** família de codificacions que aprofiten els codis del 128 al 255 per a caràcters internacionals.

En Europa occidental: ISO-8859-1 (o Latin-1); més recentment, ISO-8859-15 (o Latin-9; conté el símbol de l'euro)

En Windows s'usa CP-1252 que és molt similar (però no idèntic) a l'ISO-8859-1.

Formats de text /5

Codificacions de més d'1 octet (“multibyte”: japonés, xinés, coreà, devanagari. . . .:)

- **Unicode (ISO-10646)**: Assigna codis de 31 bits (4 octets) i permet codificar $2^{31} = 2\,147\,483\,648$ caràcters.
- **UTF-8**: Versió d'Unicode que només usa més d'un octet quan cal:
 - codis del 0 al 127: 1 octet (compatible amb ASCII):
 - codis del 128 al 2047: 2 octets;
 - codis del 2048 al 65535: 3 octets, etc.

Formats de text /6

Necessitat de format (estructural o presentacional) més enllà de la codificació de caràcters. La informació de format es pot codificar:

- Com a seqüències de caràcters (anomenades **marques**) que es poden llegir amb un editor senzill de text com el Bloc de notes: La família SGML (ara XML): HTML i XHTML (pàgines web), NewsML (notícies), etc.; RTF, TeX (processadors de textos); Postscript (impressores), etc. Un exemple:

```
<p>Un paràgraf curt amb un mot <em>emfatitzat</em>.</p>  
(HTML vist a través d'un editor de text)
```

Formats de text /7

- **Amb codis no interpretables com a caràcters:** (no visibles a través d'un editor de text senzill) Adobe PDF (impressió, presentació), formats semisecrets de processadors de textos comercials com .doc de Microsoft, etc.

Formats de text /8

El problema wysiwyg (“what you see is what you get” : “el que veus és el que obtindràs”):

Les persones usen la presentació visual per a comunicar l'estructura lògica dels documents (a persones vidents!).

Els processadors de textos actuals ens mostren el document tal com quedarà imprés mentre l'estem editant.

Sucumbim a la temptació de treballar directament sobre la presentació (negretes, màrgens, tipus) en comptes de sobre l'estructura lògica (seccions, títols, etc.)

Formats de text /9

El problema wysiwyg:

Si fem això, i més endavant volem canviar el tipus de lletra dels títols de secció o de les paraules estrangeres. . .

. . . ens toca canviar-los un per un!

Resultat: “el que veus és tot el que tens”

Podem evitar-ho? Sí.

Formats de text /10

Com? Usant estils.

Marquem estructuralment les parts (elements): títol de segon nivell, text emfatitzat, exemple, etc.

I després assignem un estil de presentació a cada part (per exemple, els títols de segon nivell pode anar numerats automàticament i en Helvètica de 14 punts, l'èmfasi pot ser en negreta i l'exemple en cursiva)

Canviar la presentació de totes les aparicions d'un element és fàcil: només cal canviar l'estil associat a l'element.

Formats de text /11

En aplicacions d'internet, la separació estructura–presentació es fa així:

- La informació —el contingut— s'estructura usant XML o HTML
- La presentació es genera (en el servidor o en el navegador) usant fulls d'estil escrits en CSS o en XSL

Formats de text /12

document (XML o HTML)
full d'estil (XSL o CSS)

→

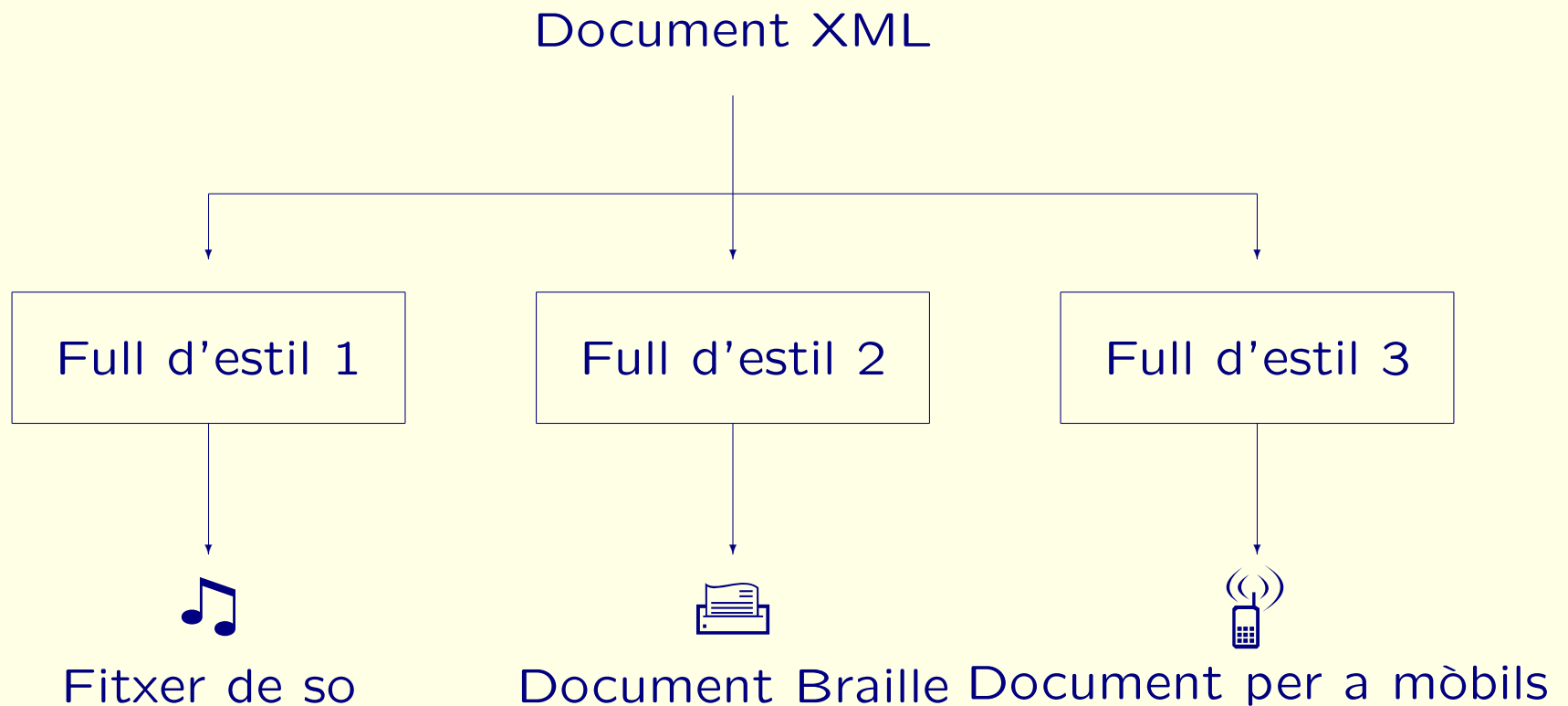
→

Processador
de fulls
d'estil

→ presentació

Formats de text /13

Accessibilitat (no tots els receptors són vidents):



Formats de text /14

Els sistemes de TA han de ser capaços:

- De separar del text a traduir la informació de format;
- de reintegrar adequadament la informació de format al text després de traduir-lo;
- i, idealment, d'usar la informació de format per a decidir quines parts cal traduir.

La preservació del format estalvia temps a la persona traductora/correctora (que es concentra en la part lingüística de la faena).

Com funciona la TA? /1

Primera aproximació [!!]: Traduir textos és traduir oracions.

Traduir oracions suposa:

- Construir una interpretació (un significat) a partir de l'oració en LO.
- Construir una oració en LM a partir de la interpretació.

Com funciona la TA? /2

Principi de composicionalitat [semàntica]:

La interpretació d'una oració es construeix ...

... a partir de les interpretacions dels mots ...

Escriuen cartes ≠ Escriuen articles

... component-les seguint les agrupacions indicades per l'estructura sintàctica de l'oració.

Israel amenaça Palestina ≠ Palestina amenaça Israel

Com funciona la TA? /3

Però alerta! Les oracions poden ser ambigües (és a dir, tenir més d'una interpretació):

- perquè els mots tenen més d'una interpretació (ambigüitat **lèxica**)
- perquè l'oració té més d'una possible anàlisi sintàctica (ambigüitat **sintàctica**)
- per ambdues coses alhora.

(en veurem exemples més endavant)

Elegir la interpretació correcta no és trivial per a un sistema informàtic (normalment només pot usar part del **cotext**).

Com funciona la TA? /4

Esquemàticament:



En alguns sistemes de TA s'intenta representar directament les interpretacions amb una interlingua (un llenguatge estructurat artificial).

Com funciona la TA? /5

Però... els traductors professionals realment necessiten interpretar o comprendre completament una oració per a traduir-la?

“... interacciones independientes del espín en unidades de la sección eficaz del neutrino de Dirac...” →

“... interaccions independents de l'espín en unitats de la secció eficaz del neutrí de Dirac...”

“... tornillos que unen el volante de inercia al árbol de levas →

*“... caragols que uneixen el volant d'inèrcia a l'arbre de lleves
...”*

No: Transformen estructures o patrons i substitueixen el lèxic (parant especial esment al terminològic).

Com funciona la TA? /6

Això permet fer la segona aproximació [!!]:

La majoria dels sistemes de TA no construeixen completament la interpretació, ...

... sinó que transformen l'estructura sintàctica de l'oració en LO en una estructura sintàctica vàlida per a l'oració en LM i...

... substitueixen els mots de l'oració en LO per equivalents adequats en LM...

... fent les dues operacions bastant independentment.

Com funciona la TA? /7

Per a programar un sistema de TA cal formular tots els processos de traducció de forma explícita i mecanitzable (adéu “intuïció lingüística”!).

A més, la mecanització ha de ser eficient (programes ràpids i compactes) i s’ha de dur a terme en un temps raonable:

- Això exigeix una reflexió lingüística (traductològica) sobre els processos de traducció per part dels dissenyadors del sistema.
- A més, pot comportar més aproximacions, simplificacions, compromisos i sacrificis.

Com funciona la TA? /8

Per tant...

Podem esperar que un bon sistema de TA ens allibere de la part més mecànica (mecnitzable) de la tasca de traducció.

Però no podem esperar —per bo que siga— que compregui el text, resolgui les ambigüitats sempre correctament i produïsci textos en una variant genuïna de la llengua meta.

Per què és difícil la TA? /1

Els quatre problemes de la traducció automàtica (Arnold 2003):

1. El problema de l'anàlisi
2. El problema de la síntesi
3. El problema de la transferència
4. El problema de la descripció

Per què és difícil la TA? /2

El problema de l'anàlisi: La forma no determina completament el contingut (la interpretació). També s'anomena ambigüitat:

- Portaven notícies de Grècia (tema o procedència?)
- Ha venut les taronges que ha comprat a Joan (Joan ven o compra?)
- Treballa en l'estudi que li han encarregat (prepara un document o dissenya un taller d'artista?)

Per què és difícil la TA? /3

El problema de la síntesi: El contingut no determina completament la forma (hi ha més d'una manera de dir el mateix en qualsevol llengua):

- Quina hora és?
- Com és de tard? (de: Wie spät ist es?)
- Quines hores són (pt: Que horas são?)

Els expedients s'obrin o s'inicien?

Les sessions es clouen, es tanquen, es rematen o s'alcen?

Per què és difícil la TA? /4

El problema de la transferència: Les llengües divergeixen. És a dir, hi ha diferències irreductibles en la manera en que el mateix contingut s'expressa en llengües diferents:

- ca: M'agrada nadar (M' objecte, agrada, verb, nadar subjecte)
- en: I like swimming (I subjecte, like verb, swimming objecte)
- de: Ich schwimme gern (Ich subjecte, schwimme, verb, gern, adverbi)

Totes volen dir `produir_plaer(agent=nadar(agent=jo),destinatari=jo)`

Per què és difícil la TA? /5

El problema de la descripció (represa): construir un sistema de traducció automàtica comporta la gestió d'una gran quantitat de coneixement, que s'ha d'elicitar, aplegar, descriure, i representar en una forma útil i computable.

Avaluació de la traducció automàtica /1

Volem avaluar l'adopció d'un sistema de traducció automàtica per a la disseminació.

Les traduccions en brut s'hauran de posteditar (corregir): com menys correccions, més qualitat: millor.

D'acord: com avaluem la qualitat?

Avaluació de la traducció automàtica /2

Per avaluar la qualitat, cal:

- elegir una mostra suficient de textos representatius,
- traduir-la automàticament,
- i comptar la quantitat de correcció mínima necessària per a fer que la traducció siga adequada al propòsit previst.

Sembla senzill, però...

Avaluació de la traducció automàtica /3

...no ho és gens!

- és difícil elegir prou text representatiu per endavant;
- la noció d'adequació és de vegades difícil d'especificar:
- és difícil fer el mínim de correccions (cal buscar traduccions adequades que se n'obtinguen amb poques correccions);
- tot el procés és molt costós (temps de correcció).

Avaluació de la traducció automàtica /4

Però la qualitat dels textos traduïts en brut no ho és tot!

Fem un pressupost: si adoptem la traducció automàtica, d'una banda, ens estalviem els costos de traducció humana, però tenim despeses noves:

- despeses de funcionament i
- despeses de formació (s'ha d'aprendre a usar una nova tecnologia)

Avaluació de la traducció automàtica /5

Despeses de funcionament:

- **Cost del sistema de TA** (cost efectiu per mot): amortització (sistema en propietat), cost per mot (sistema llogat), servei tècnic i manteniment, costos de migració (adaptació de programes, adquisició de sistemes), i (no oblidem) el cost d'avaluació!
- **Cost de preedició i preparació:** cal preparar i potser preeditar els textos i això ho ha de fer algú, cobrant.
- **Cost de postedició:** depèn de la qualitat; pot baixar amb la formació; depèn de com paguem als posteditors (per mot, per temps), etc.

Avaluació de la traducció automàtica /6

Despeses de formació:

- **Formació en ús del programa de TA:** ús pròpiament dit, configuració i manteniment; ús de nou programari associat.
- **Formació en postedició:**
 - coneixement del programa de TA (errors típics);
 - tècniques de correcció, ús avançat del processador de textos, macroinstruccions, substitució de patrons, etc.

Avaluació de la traducció automàtica /7

I potser ens hem deixat encara alguna cosa!

Avaluar la traducció automàtica no és fàcil.

La lliçò? Desconfieu de les primeres impressions.

TA de pàgines web/1

La traducció automàtica de pàgines web és com la TA d'altres documents de text, però hi ha algunes diferències:

- les pàgines web són hipertextos: contenen enllaços a d'altres pàgines web
- de vegades són actives: contenen programes que s'executen durant la presentació
- de vegades són dinàmiques: el servidor no les té guardades sinó que les genera automàticament quan se sol·liciten

TA de pàgines web/2

Dos usos bàsics de la TA de pàgines web:

- **Disseminació:** TA per a construir i mantenir servidors d'Internet multilingües
- **Assimilació:** TA durant la navegació ("navegació traduïda"), en el client, en el servidor que conté la informació, o en un altre servidor (p.e., interNOSTRUM)

TA de pàgines web/3

Els requisits són diferents en cada cas:

- **Disseminació:** la TA ha de ser de qualitat, potser seguida de postedició (la web traduïda és percebuda com a definitiva)
- **Assimilació:** la TA ha de ser molt ràpida, “instantània”, com si formara part del procés de presentació de la traducció: la qualitat no és tan crucial (la traducció és percebuda com a provisional)

Un incís: el format de les pàgines web/1

Les pàgines web són documents especials:

- La majoria de les pàgines web estan escrites en (algun dialecte no estàndard de) HTML (HyperText Markup Language, “llenguatge de marques per a hipertextos”).
- HTML conté, a més de text senzill, marques per a controlar la presentació i per a enllaçar altres documents.
- Quan editem amb programes especialitzats (Composer, Frontpage, Dreamweaver, etc.) no veiem les marques sinó l’aparença aproximada del document.

La transparència següent conté un exemple.

Un incís: el format de les pàgines web/2

```
<HTML>
<HEAD>
<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 HTML. Els documents HTML poden contenir
<A HREF="http://www.internostrum.com">enllaços</A>
a altres documents HTML, locals o remots.</P>
</BODY>
</HTML>
```

Un incís: el format de les pàgines web/3

Encapçalament de nivell 1

Encapçalament de nivell 2

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

Un altre encapçalament de nivell 2

Aquest és l'últim paràgraf d'aquest document HTML. Els documents HTML poden contenir [enllaços](#) a d'altres documents HTML, locals o remots.

TA de pàgines web (represa)/4

Traduir un document HTML comporta:

- Identificar les porcions del document que corresponen a text que ha de ser llegit i traduir-les;
- Adaptar els enllaços a la nova situació (potser ja no poden enllaçar el mateix document!).

L'adaptació d'enllaços depén de la situació.

TA de pàgines web /5

Els enllaços contenen URIs (adreces d'altres documents):

Podeu visitar també els nostres

`
productes.`

TA de pàgines web/6

Una miradeta als URIs dels enllaços:

`http://www.servidor.ct/es/prod/ta.html`

L'URI (localitzador) indica:

- L'esquema (`http:` protocol de transferència d'hipertext)
- El nom de la màquina que fa de servidor (`www.servidor.ct`).
- La ruta que identifica el recurs concret (`/es/prod/ta.html`) dins del servidor

TA de pàgines web/7

Traduir suposa adaptar els URIs dels enllaços. Per exemple, en un servidor bilingüe espanyol–català, si un enllaç des d'una pàgina en espanyol apunta a l'URI

```
http://www.servidor.ct/es/prod/ta.html
```

La traducció catalana hauria d'apuntar a l'URI:

```
http://www.servidor.ct/ca/prod/ta.html
```

TA de pàgines web/8

Però el text mateix de les pàgines web conté de vegades material especial que no cal traduir:

- **URIs:** `www.pujol.com` (no és “`www.colina.como`”)
- **Adreces de correu electrònic:** `andreu.fuster@correu.com` (no és “`andrés.carpintero@correo.como`”)

TA de pàgines web/9: pàgines generades al servidor

Moltes vegades els documents HTML no són al disc dur, sinó que són generats per un programa que s'executa en el servidor durant la navegació.

Possiblement es generen pàgines diferents per a cada perfil de visitant.

Els detalls de la traducció d'aquest tipus de documents queden fora de l'abast d'aquest curs, però presenten reptes considerables.

Webs preparades per a la TA: aspectes lingüístics/1

Si preveiem que una web ha de ser traduïda automàticament a una altra llengua, podem preparar el text origen.

Conèixer el sistema concret de TA ens pot ajudar a evitar els mots o les construccions que donen lloc a problemes.

La noció és coneguda de fa temps i s'anomena **llenguatge controlat**.

Webs preparades per a la TA: aspectes lingüístics/2

Alguns consells independents de l'idioma:

- Fer pàgines i paràgrafs curts.
- No usar textos en gràfics (imatges), sinó icones.
- Usar estructures gramaticals senzilles.
- Usar vocabulari bàsic (freqüent, quotidià), però...
- Evitar els mots polisèmics i els homògrafs (homònims).
- Evitar les abreviatures.

Webs preparades per a la TA: aspectes lingüístics/3

Més consells independents de l'idioma:

- No usar el format per a transmetre informació crucial; millor usar text.
- Repassar l'ortografia.
- Evitar les expressions idiomàtiques (no òbviament composicionals).

Memòries de traducció/1

Els traductors (humans) han generat moltíssimes traduccions.

Hi ha a l'abast nombrosos textos electrònics bilingües on la versió en un idioma és una bona traducció de la versió en l'altre i viceversa.

No es podria aprofitar aquest treball per a traduir documents nous (reciclatge automàtic de traduccions?) → Alternativa a la traducció automàtica.

Memòries de traducció /2

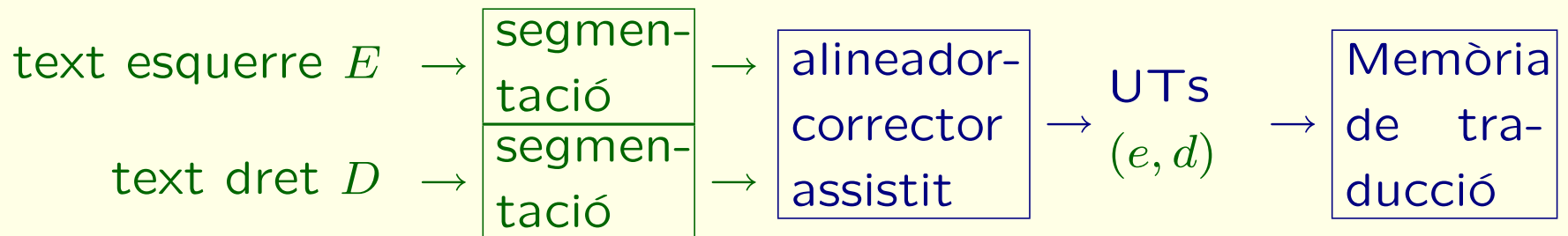
Per a aprofitar aquests bitextos cal:

- Alinear-los (indicar quines parts són traducció de quines);
- Segmentar-los en unitats de traducció (UT);
- Organitzar les UT en una base de dades eficient.

Totes aquestes tasques, tan automàticament com siga possible.

Memòries de traducció/3

Esquema del procés de segmentació i d'alineament d'un parell de textos existent per a alimentar una memòria de traducció.



Memòries de traducció/4

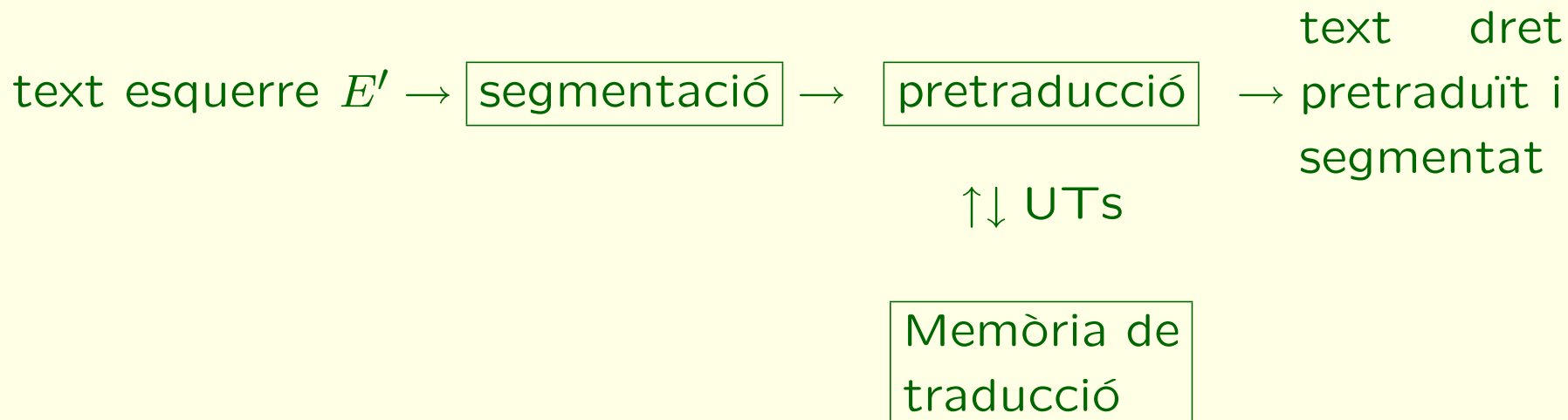
Per a traduir textos nous cal:

- Segmentar-los en unitats que puguem correspondre amb les UT existents
- Substituir els segments trobats per les traduccions corresponents.

Aquest és el fonament de les memòries de traducció.

Memòries de traducció/5

Esquema del procés de pretraducció d'un nou text esquerre E' usant una memòria de traducció.



Memòries de traducció/6

Alguns productes comercials (preus de 600 euros cap amunt):

- Déjà Vu d'Atril (<http://www.atril.com/ca/>)
- Transit de Star (<http://www.star-transit.com/es/>)
- Trados (www.trados.com)
- SDLX (www.sdlintl.com)

Solen contenir, a més de la memòria de traducció, altres útils com ara bases de dades lèxiques (“terminològiques”), etc. Hi ha productes Open Source com ara OmegaT.

Memòries de traducció/7

Quan funcionen bé les memòries de traducció?

- Quan tenim moltes traduccions alineades en la memòria
- Quan els tipus de textos a traduir són molt repetitius
- Quan la terminologia i la fraseologia són estables en la memòria

Però:

- sempre cal revisar la pretraducció
- A canvi: la pretraducció revisada es pot afegir ja a la memòria de traducció per usar-la en el futur.

Memòries de traducció/8

Sobre la segmentació:

- Els programes de MT segmenten els textos en “oracions” usant la puntuació i el format.
- A canvi, troben en la memòria segments aproximats a més dels idèntics (i produeixen traduccions aproximades).
- Hi ha (des de 1998) un format estàndard internacional de MT independent del programa: TMX (Translation Memory eXchange), que permet l’intercanvi de memòries entre equips de traducció.